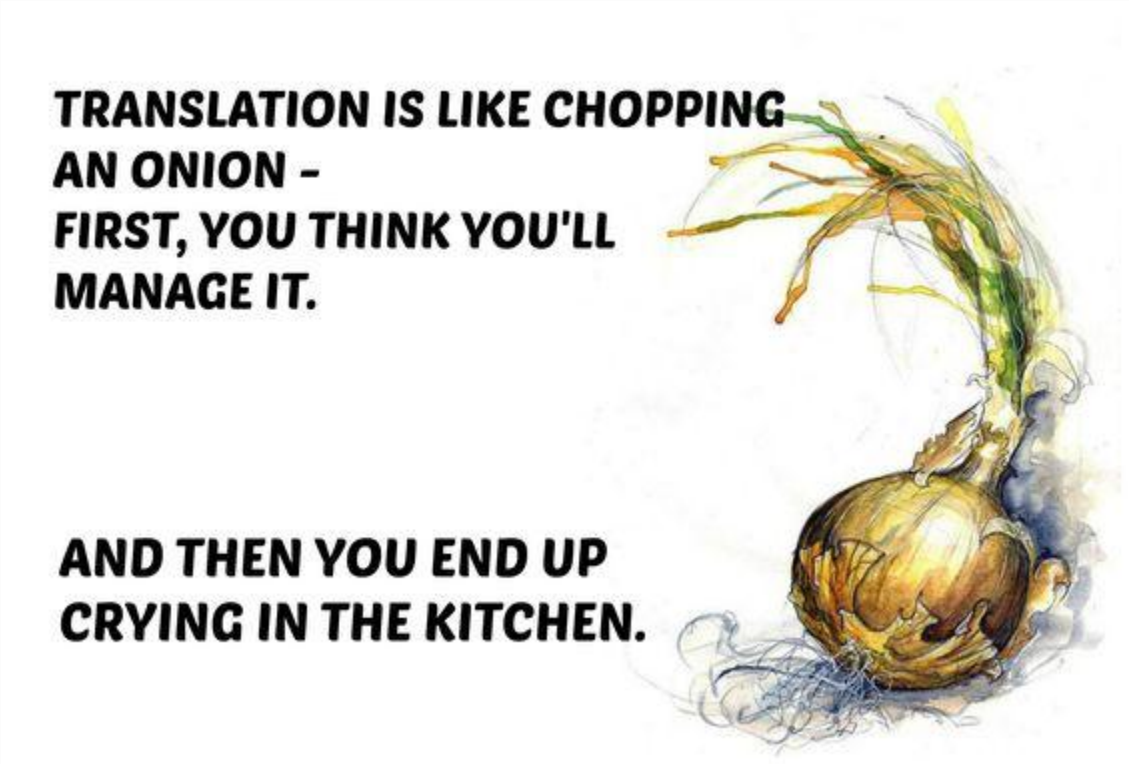



*Source: Vera Hembitskaya*

# CS 2731
# Introduction to Natural Language Processing

## Session 23: Machine translation

Michael Miller Yoder

November 11, 2025

University of Pittsburgh | School of Computing and Information

# Course logistics: project

- Project progress report **due tomorrow, Thu Nov 13**
- Part 1: Task and dataset
  - Address the questions on basic dataset statistics, as well as how you will use your dataset to address your task
  - If you do not have a “traditional” dataset, present rough equivalents
- Part 2: Some kind of a result
  - Options: Baseline system evaluation on your dataset, a result from your own system, an example output from your system
- Part 3: Open questions and challenges
  - Need any help or additional resources?

# Course logistics: homework

- Homework 4 is **due next Thu Nov 20**
    - More Hugging Face, this time BERT-based models for part-of-speech tagging

# Structure of this course

| | | | |
|---|---|---|---|
| MODULE 1 | Introduction and text processing | text normalization, machine learning, NLP tasks | |
| | **Approaches** | **How text is represented** | **NLP tasks** |
| MODULE 2 | statistical machine learning | n-grams | language modeling<br>text classification |
| MODULE 3 | neural networks | static word vectors | text classification |
| MODULE 4 | transformers and LLMs | contextual word vectors | language modeling<br>text classification |
| MODULE 5 | Sequence labeling and parsing | named entity recognition, dependency parsing | |
| MODULE 6 | NLP applications and ethics | machine translation, chatbots, search engines, bias | |

# Review

1. What is syntax?
2. What is the output format of syntactic parsing (like dependency parsing) tasks?

# Overview: Machine translation

- Translation in practice
- Why is translation difficult?
- Exercise: translate some Tajik
- Parallel corpora
- Encoder-decoder MT systems with transformers
- Beam search
- MT evaluation
- Bias and MT

# Translation

- Mapping a “text” in a source language to a target language

“I went to the store to buy eggs” → “Eu fui à loja comprar ovos”


→


*Slide adapted from Lori Levin*

# Translation in practice

# Most translation is still done by human translators

## Translation and Localization Industry Grows 11.8% in 2021 to USD 26.6bn


https://slator.com/translation-localization-industry-grows-in-2021-to-usd-26bn/
*Image credit: AMN Healthcare*
*Slide adapted from Lori Levin*

# Post-editing and computer-assisted translation

- Checking and correcting of machine translation by humans is called **post-editing**


Evacuation Ladder


Do not yell

*Slide adapted from Lori Levin*

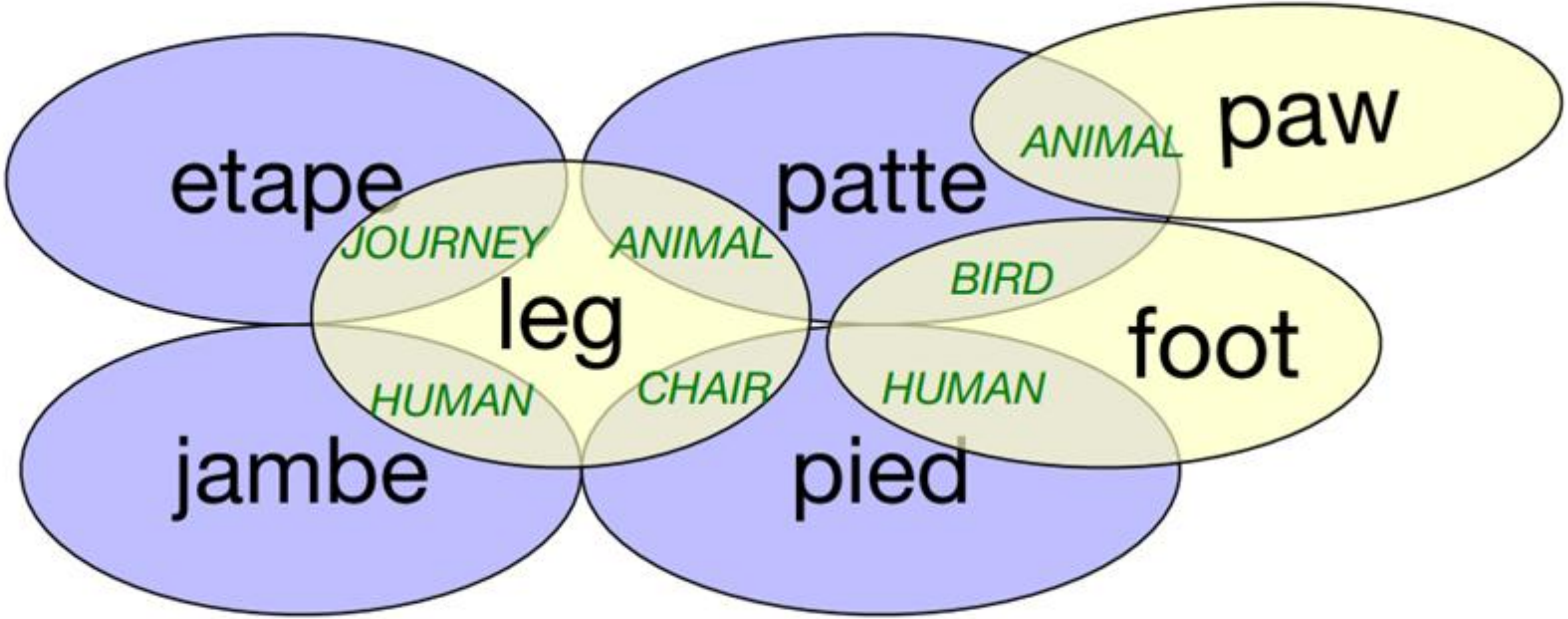


*Images credit:* https://www.languageconnections.com/blog/7-hilarious-machine-translation-mistakes/

# Why is translation difficult?

# Why not just look up each word in a dictionary and translate word-for-word?

Many-to-many mappings of words




Slide adapted from Lori Levin, Jurafsky & Martin, Hutchins & Somers

# Why not translate word-for-word: grammar distinctions

The grammars of some languages make distinctions that other languages don't make:

- Russian *kniga* translates to English as *the book* or *a book*.
  - English grammar makes a distinction in definiteness
  - Russian grammar does not.
- English *it* translates to French *il*/*le* (masculine) or *elle*/*la* (feminine).
- English *a* translates to French as *un* (masculine) or *une* (feminine).
  - *Une chaise* (a chair) vs *un livre* (a book)
  - French grammar makes a distinction in gender
  - English grammar does not.



*Slide adapted from Lori Levin*

# Why not translate word-for-word:
# Different numbers of words to say the same thing

uygarlaştıramadıklarımızdanmışsınızcasına

"(behaving) as if you are among those whom we were not able to civilize"

| | |
|---|---|
| uygar | "civilized" |
| +laş | "become" |
| +tır | "cause to" |
| +ama | "not able" |
| +dık | past participle |
| +lar | plural |
| +ımız | first person plural possessive ("our") |
| +dan | ablative case ("from/among") |
| +mış | past |
| +sınız | second person plural ("y'all") |
| +casına | finite verb → adverb ("as if") |

*Slide adapted from Lori Levin*

# Why not translate word-by-word: word order

English: *He wrote a letter to a friend* ⟵ SVO (verb-medial)

Japanese: *tomodachi ni tegami-o kaita* ⟵ SOV (verb-final)

friend to letter wrote

Arabic: *katab risāla li ṡadq* ⟵ VSO (verb-initial)

wrote letter to friend



*Slide adapted from Lori Levin, Jurafksy & Martin*

# Exercise: Tajik

There are 3,344,720 speakers of ***Tajik*** in Tajikistan (one of the Central Asian republics of the former Soviet Union) and another million speakers in surrounding countries.

| | |
|---|---|
| дуусти хуби ҳамсояй сумо | a good friend of your neighbor |
| ҳамсояй дуусти хуби сумо | a neighbor of your good friend |
| ҳамсояй хуби дуусти сумо | a good neighbor of your friend |

Above are three phrases in Tajik with their English translations. Your task is to give the English translations of all four Tajik words. The possibilities are simply "good," "friend," "neighbor," and "your." The order of the words – which is not the same order as in English! – does the rest.

дуусти ________

ҳамсояй ________

хуби ________

сумо ________

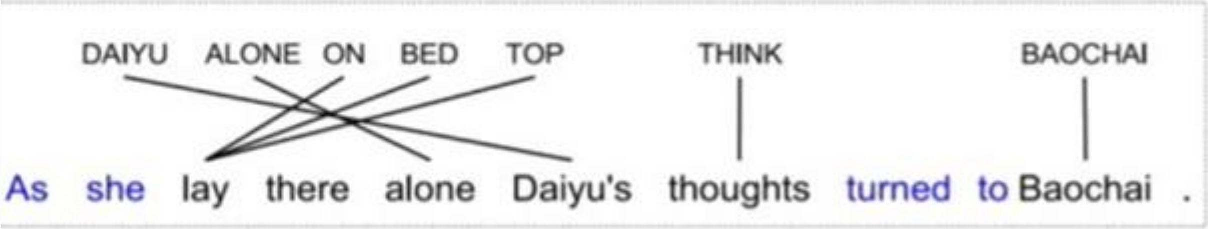
*Problem by Adriana Solovyova*
*Slide adapted from Lori Levin*

# Why is translation difficult? Style and genre


From “Dream of the Red Chamber”, Cao Xue Qin (1792)

Chinese:

English:




Parallel data is more likely to match styles (like literary style) than be an “exact” translation



*Slide adapted from Patrick Fernandez, Graham Neubig, Xinyi Wang, Lori Levin*

# Preparing for machine translation

1. Collect a parallel corpus
2. Align sentences

*Slide adapted from Lori Levin*

# Parallel corpora

Buy Gift Cards    **Pickup** ASAP (20-25 minutes) **from** 114 Atwood St.    0

## Bao - Pitt Campus

# Food

## Appetizers 头台

**Tea Egg 茶叶蛋**
$4.00

**Pork Belly Slider 五花肉刈包**
$7.95

**Popcorn Chicken 盐酥鸡**
$8.95

**Cantonese Style Chicken Feet 广式风爪**
$8.95

**Rolled Pancakes w/Roast Beef 牛肉卷饼**
$12.95

**Pan Fried Radish Cake 萝卜糕**
$7.95

**Crab Rangoon 蟹角**
$7.95

**Indian Pan Fried Pancake 印度薄煎饼**
$6.95

# Parallel corpora examples

- Europarl: Proceedings of the European Parliament; 21 languages; up to 2 million sentences
- United Nations Parallel Corpus: 10 million sentences in Arabic, Chinese, English, French, Russian, Spanish
- OpenSubtitles: movie and TV subtitles
- ParaCrawl: 223 million sentences in 23 EU languages



*Slide adapted from Lori Levin, Jurafsky & Martin*

# What about parallel corpora for the other 7000 languages?

- For many languages, the only parallel text is the Christian Bible.
- Low-resource MT is a large area of research
  - How to leverage monolingual texts (backtranslation)
  - Humans in the loop
  - Leverage multilingual models

*Slide adapted from Lori Levin*

# Sentence alignment

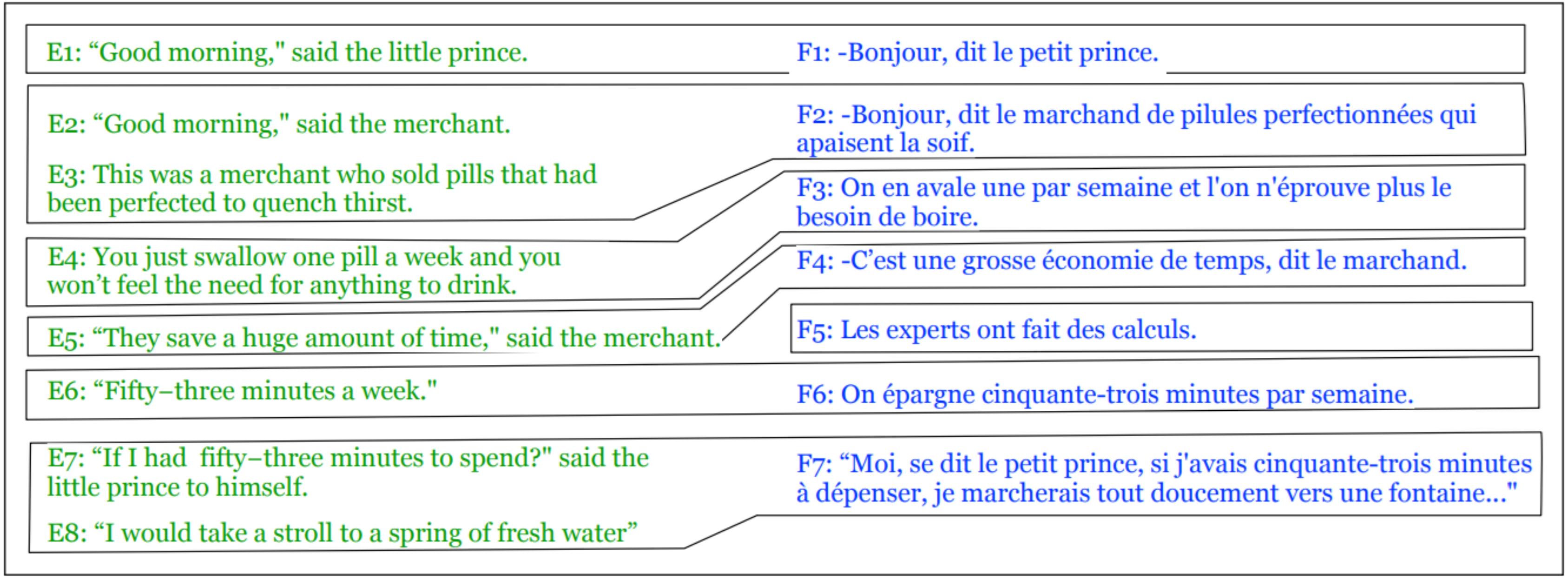



**Figure 10.17** A sample alignment between sentences in English and French, with sentences extracted from Antoine de Saint-Exupery’s *Le Petit Prince* and a hypothetical translation. Sentence alignment takes sentences $e_1, \ldots, e_n$, and $f_1, \ldots, f_n$ and finds minimal sets of sentences that are translations of each other, including single sentence mappings like ($e_1$,$f_1$), ($e_4$,$f_3$), ($e_5$,$f_4$), ($e_6$,$f_6$) as well as 2-1 alignments ($e_2$/$e_3$,$f_2$), ($e_7$/$e_8$,$f_7$), and null alignments ($f_5$).

*Figure credit: Jurafsky & Martin*

# Encoder-decoder MT systems

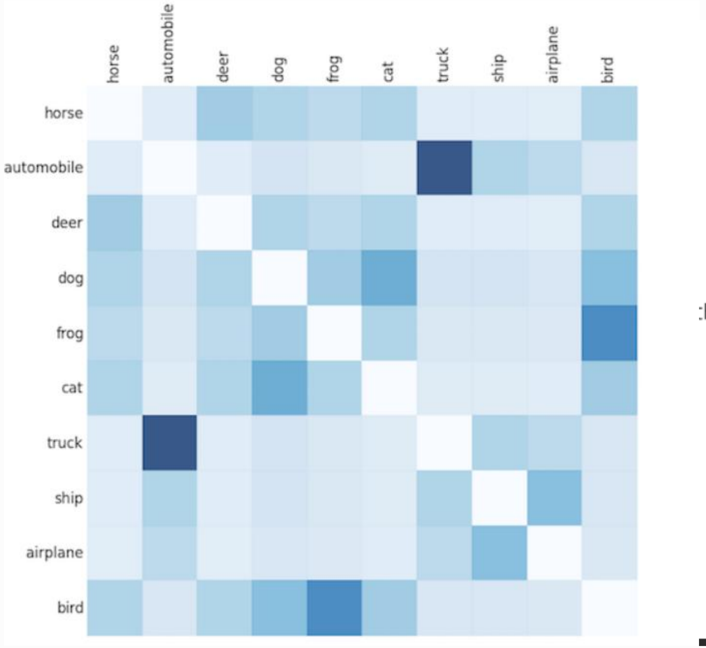
Which model to train?

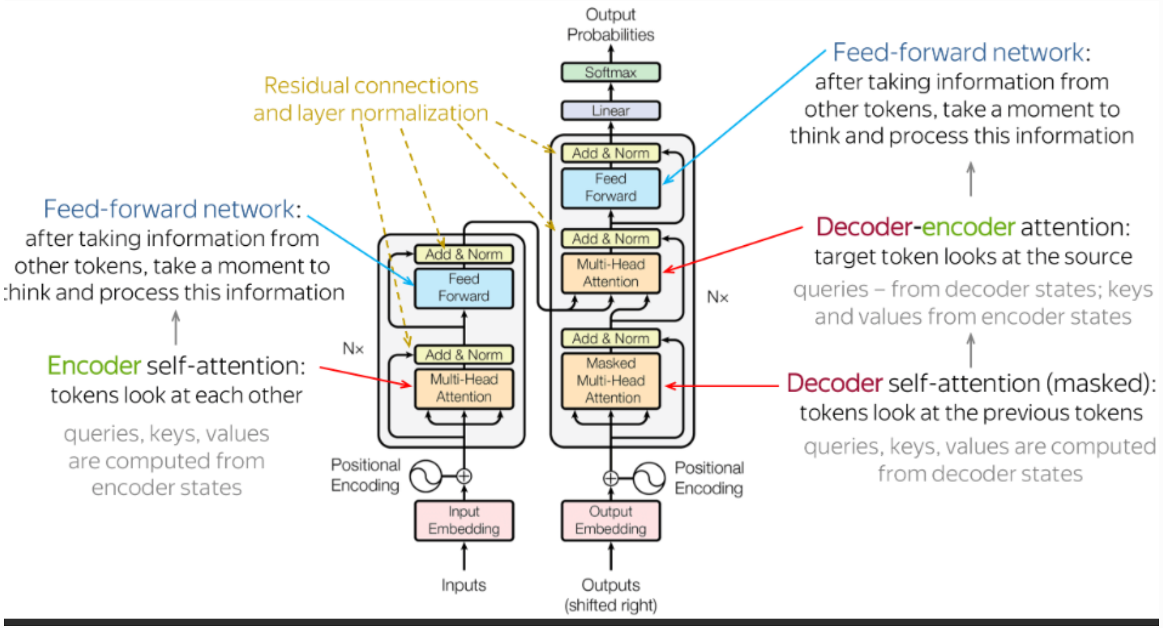
of course. But why?



Slide credit: Sabit Hassan

# Recap: Attention and Transformers




- Focus on different parts of input for each input and output
- Closer to how we humans may process language

*Slide credit: Sabit Hassan*

# Encoder-decoder transformer architecture

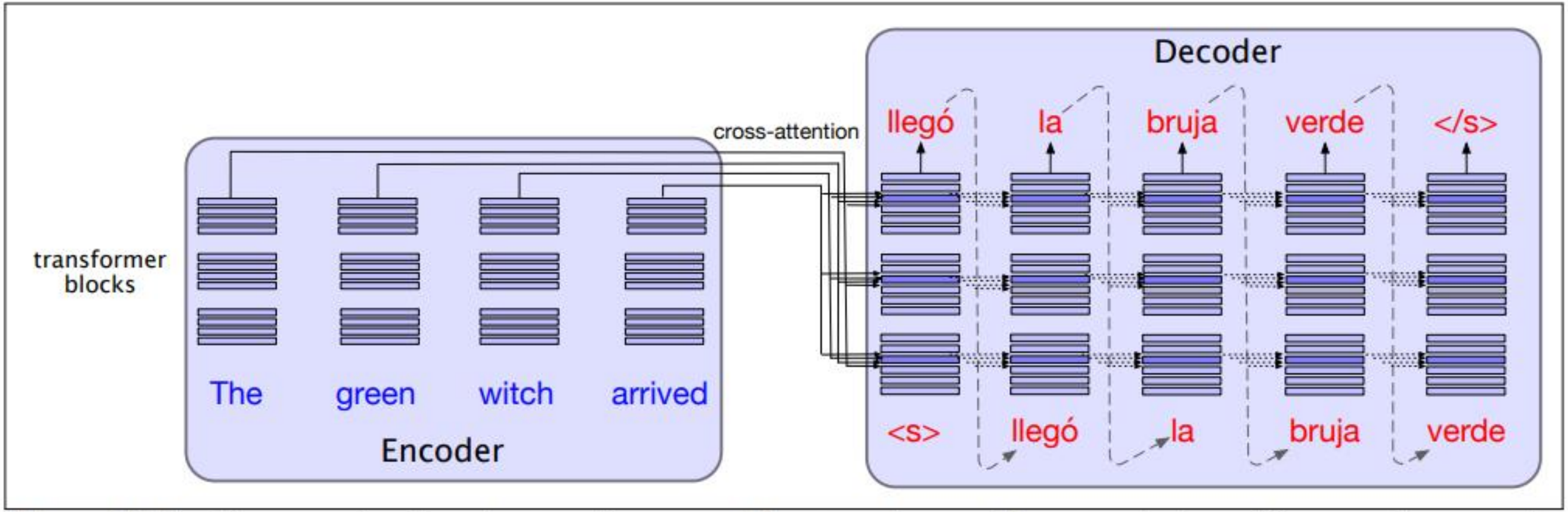


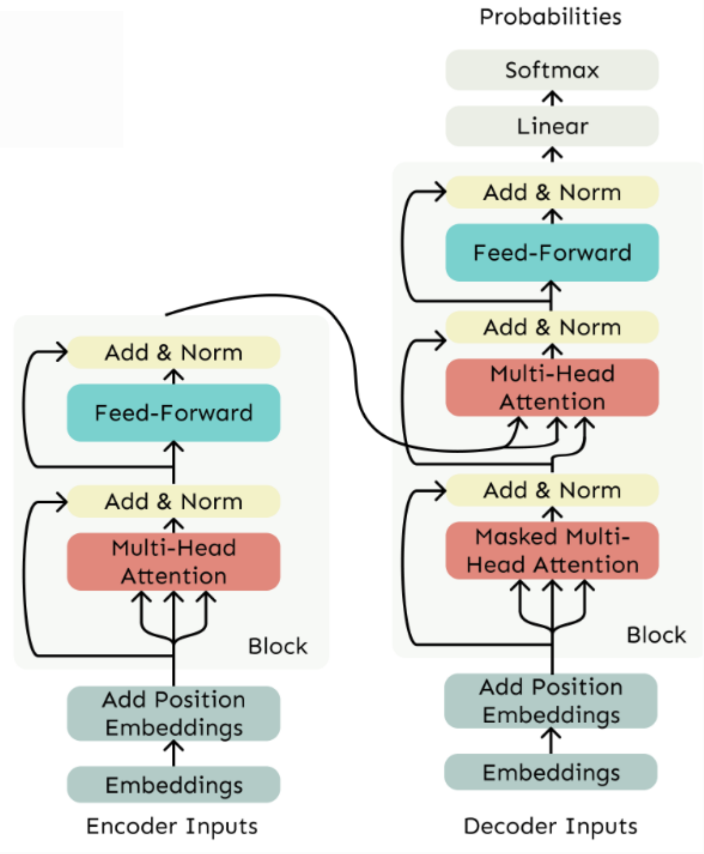

**Figure 10.5** The encoder-decoder transformer architecture for machine translation. The encoder uses the transformer blocks we saw in Chapter 9, while the decoder uses a more powerful block with an extra **cross-attention** layer that can attend to all the encoder words. We'll see this in more detail in the next section.



*Slide adapted from Jurafsky & Martin*

# The transformer encoder-decoder

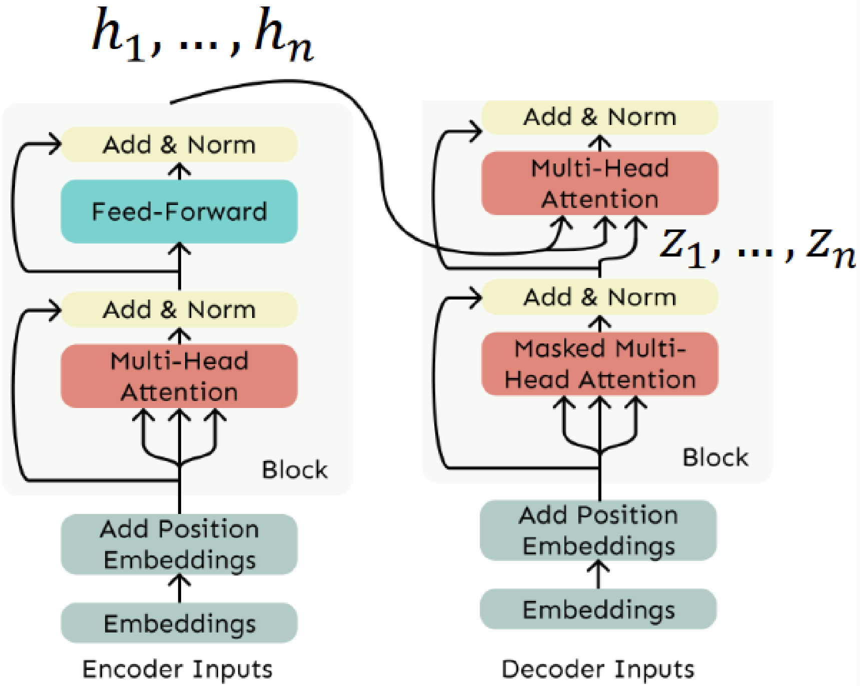
- Can use transformers for encoder-decoder (seq2seq) framework
- Transformer decoder modified to perform cross-attention to the output of the encoder




*Slide adapted from John Hewitt*

# Cross-attention

- We saw that self-attention is when keys, queries, and values come from the same source.
- In the decoder, we have attention that looks more like what we saw last week.
- Let $h_1, \ldots, h_n$ be **output** vectors **from** the Transformer **encoder**; $x_i \in \mathbb{R}^d$
- Let $z_1, \ldots, z_n$ be input vectors from the Transformer **decoder**, $z_i \in \mathbb{R}^d$
- Then keys and values are drawn from the **encoder** (like a memory):
  - $k_i = Kh_i, v_i = Vh_i$.
- And the queries are drawn from the **decoder**, $q_i = Qz_i$.




*Slide adapted from John Hewitt*

# Beam search

# Beam search improves on greedy decoding

- Traditional encoder-decoder framework involves generating highest probability word (argmax) at each timestep in the decoding
- But this greedy approach suffers from issues if choosing early high-probability tokens leads to low-probability sequences!
- **Solution**: Don't commit to just the 1 highest probability word, but keep multiple options in a "beam"
- Prune to *k* highest-probability sequences after each timestep

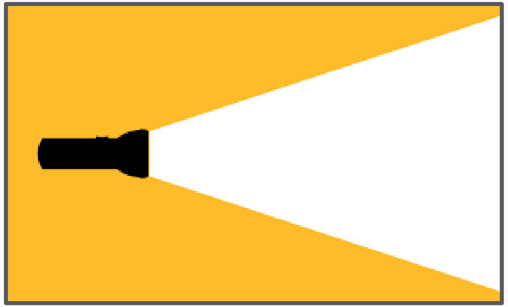
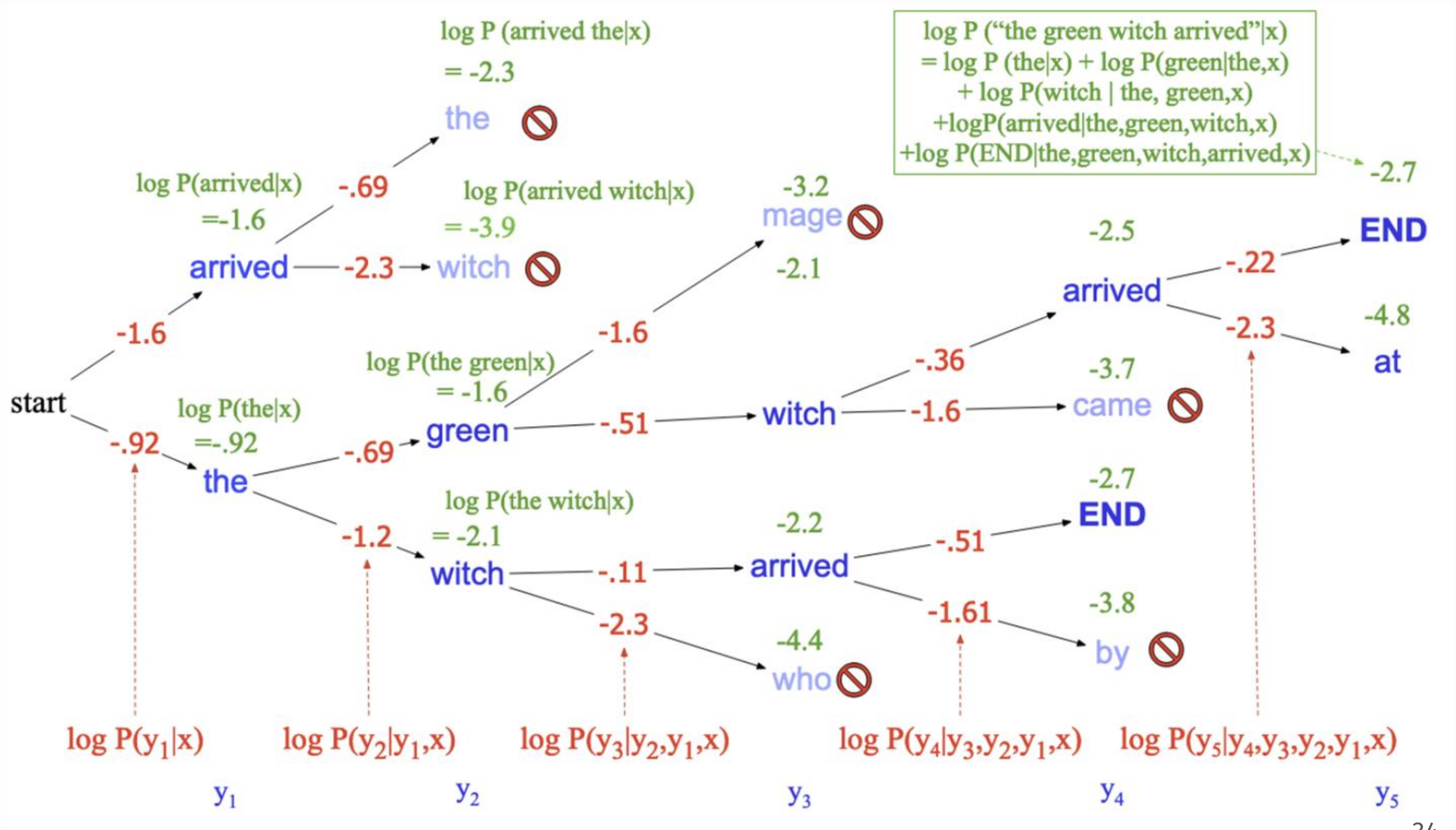


*Image: iStock*

# Beam search example

Find highest probability English sentence for *x* = “llegó la bruja verde”



# MT evaluation

# Human evaluation of MT

**Human evaluation:** Rate/edit translations. Expensive but the best.

- Can ask bilingual raters to compare original source text with prediction
- Can ask monolingual raters to compare predicted translation with reference translation



*Slide adapted from Sabit Hassan*

# Two aspects of human evaluation of MT

- **Adequacy**: how well translation captures exact meaning of the source sentence
- **Fluency**: how fluent/readable/natural the translation is in the target language



Slide adapted from Sabit Hassan

# Automatic evaluation of MT

- Character or word **overlap-based**
  - chrF, BLEU
- **Embedding-based**: measure distance between embeddings of tokens
  - Trying to capture synonyms
  - METEOR, BERTScore
- **Classifier-based:** train a classifier to predict human ratings between predicted translations and reference translations
  - COMET, BLEURT



*Slide adapted from Sabit Hassan*

# chrF score

- **chrP**: percentage of character 1-grams, 2-grams, ..., k-grams in the hypothesis that occur in the reference, averaged.
- **chrR**: percentage of character 1-grams, 2-grams,..., k-grams in the reference that occur in the hypothesis, averaged.

$$\mathrm{chrF}\beta = (1+\beta^2)\frac{\mathrm{chrP}\cdot\mathrm{chrR}}{\beta^2\cdot\mathrm{chrP}+\mathrm{chrR}}$$



*Slide adapted from Sabit Hassan*

# Bias in MT

# Example: gender bias in pronoun translation

| Hungarian (gender neutral) source | English MT output |
|---|---|
| ő egy ápoló | she is a nurse |
| ő egy tudós | he is a scientist |
| ő egy mérnök | he is an engineer |
| ő egy pék | he is a baker |
| ő egy tanár | she is a teacher |
| ő egy esküvőszervező | she is a wedding organizer |
| ő egy vezérigazgató | he is a CEO |

**Figure 13.12** When translating from gender-neutral languages like Hungarian into English, current MT systems interpret people from traditionally male-dominated occupations as male, and traditionally female-dominated occupations as female (Prates et al., 2019).

*Figure from Jurafsky & Martin*

# Fixing MT: bias

- Expand definitions of bias
  - Bias is multifaceted. Gender, racial, cultural, linguistic
- Identify existence of bias
- Identify sources of bias: annotations? Embedding space?
- Involve native speakers in evaluation



*Slide credit: Sabit Hassan*

# Conclusion

- MT is often used in conjunction with human translators
- Language divergences (in word meaning, syntax structure, etc) make MT difficult
- Parallel corpora are used for training MT systems
- Encoder-decoder transformer MT systems use cross-attention to attend to the source language input when generating the target language output
- Automatic overlap methods (chrF, BLEU) are popular MT evaluations, though can be poor proxies for adequacy and fluency ratings by humans
- Like any NLP task, social biases (e.g. gender in pronouns) must be considered in MT

*Questions?*