CS 2731 Introduction to Natural Language Processing

Session 23: Machine translation part 2

Michael Miller Yoder

April 3, 2024

University of Pittsburgh | School of Computing and Information

Course logistics: project

- Project basic working system **due this Thu Apr 4**
  - 1-2 pages, in ACL LaTeX format that final report will be in
- No in-person lecture on Mon Apr 8 (solar eclipse)
  - Video lecture to watch asynchronously on Canvas

Images credit: Newsweek, Lansing State Journal

Overview: Machine translation part 2

- Parallel corpora
  - Sentence alignment
- Encoder-decoder MT systems with transformers
- MT for low-resource languages
- MT evaluation
- Bias and MT

Parallel corpora and sentence alignment

Review: parallel corpora (bitext)

French	English
Qui contrôle strictement court le risque que ses ports restent encombrés d' épav ⋱.	Countries that impose stricter controls run the risk of being saddled with shipw ⋱.
Cela suppose que nous soyons capables de rehausser politiquement chacune des ins ⋱.	This presupposes our being able to raise the profile of each of the institutions ⋱.
La Lituanie dispose d'un potentiel appréciable de croissance économique durable.	Lithuania has considerable potential for long–term economic growth.
Enfin, les adultes incapables ne doivent participer qu'à des essais qui portent ⋱.	Finally, adults incapable of giving consent should only participate in trials th ⋱.
Par intérêt économique, l'Europe, les États–Unis et l'Australie ne demandaient q ⋱.	Out of economic self–interest, Europe, the United States and Australia wanted to ⋱.
J'ai reçu sept propositions de résolution , déposées sur la base de l'article 37 ⋱.	I have received seven motions for resolutions, tabled pursuant to Rule 37(2) of ⋱.
La Commission, une fois encore, n'a pas voulu s'engager dans des négociations in ⋱.	The Commission, again, has failed to commit itself to entering international neg ⋱.
L'entendre ainsi nier le fait que les aides d'État ont diminué durant la période ⋱.	Hearing him deny the fact that state aid was reduced in the period 1994–1998, ev ⋱.
L'avocat se voit interdire tout ce qui n'est pas permis par le strict respect de la légalité.	Lawyers are forbidden to do anything that is not strictly legal.
Les applaudissements qui l'ont salué montrent bien que lorsqu'il y a un objectif ⋱.	The applause that rounded it off clearly demonstrates that when there is a speci ⋱.

Figure credit: Wolfram Data Repository

Sentence alignment

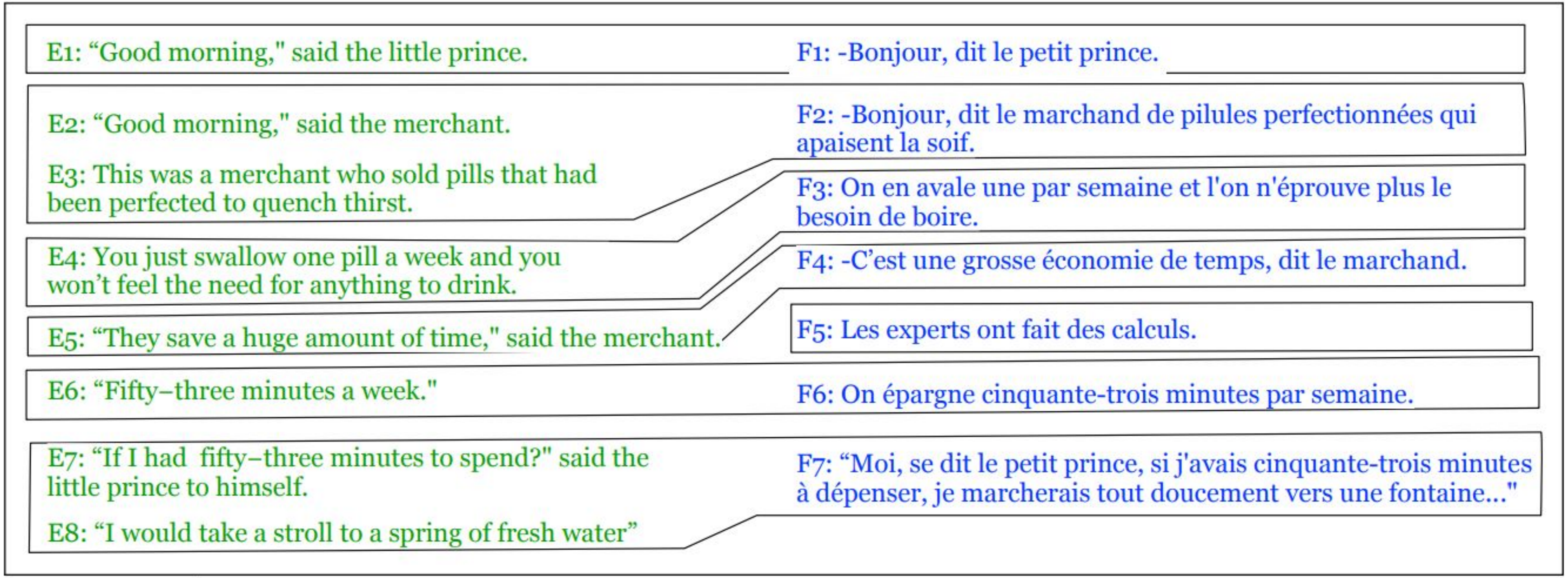


Figure 10.17 A sample alignment between sentences in English and French, with sentences extracted from Antoine de Saint-Exupery’s *Le Petit Prince* and a hypothetical translation. Sentence alignment takes sentences $e_1, \ldots, e_n$, and $f_1, \ldots, f_n$ and finds minimal sets of sentences that are 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

How to align sentences

Need:

1. Cost function: how likely are a source language span and a target language span to be translations (matching sentences)?
2. Alignment algorithm: uses scores between spans to find a good alignment between documents

Multilingual embedding space

1. Cost function: score similarity of sentences across languages with cosine similarity of embeddings in **multilingual embedding space**

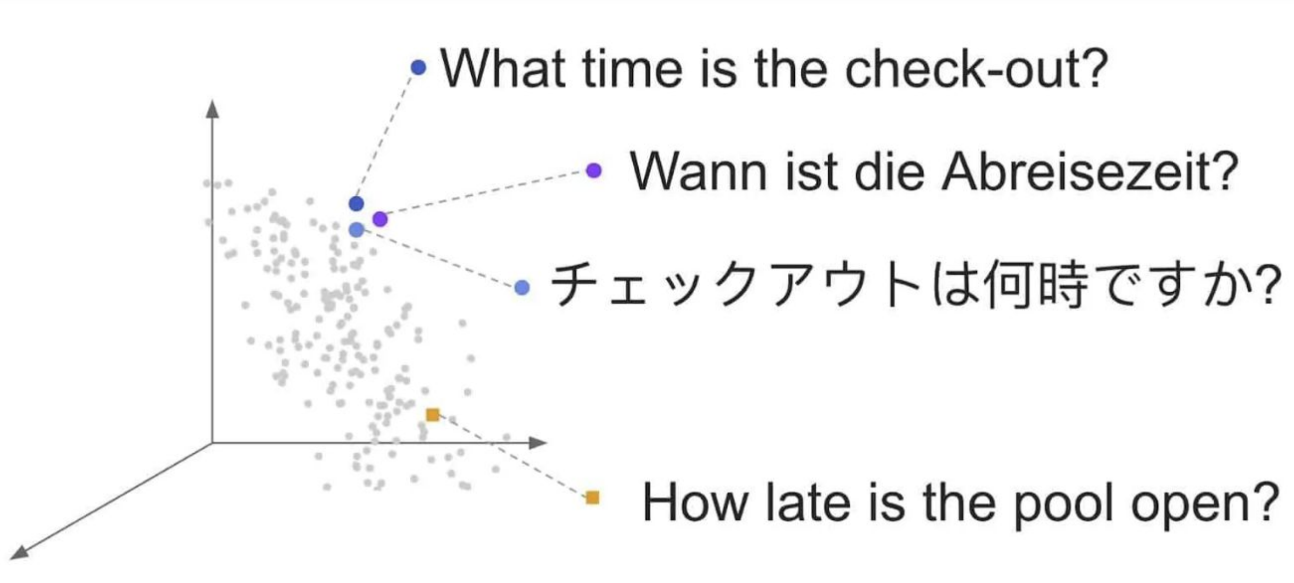


Figure credit: Megagon Labs

Sentence alignment: cost function and alignment alg

1. Cost function using cosine similarity of embeddings in multilingual embedding space [Thompson + Koehn 2019]

$$c(x,y) = \frac{(1-\cos(x,y))\,\mathrm{nSents}(x)\ \mathrm{nSents}(y)}{\sum_{s=1}^{S} 1-\cos(x,y_s) + \sum_{s=1}^{S} 1-\cos(x_s,y)}$$

2. Dynamic programming algorithm [Gale + Church 1993] as the alignment algorithm
   - Minimize cost over the entire sequence of spans

Subword tokenization review

- Create a shared vocabulary between source and target language with **subword tokenization**
- Example: Byte-pair encoding (BPE, Sennrich et al. 2016)
  - Merges frequently seen sequences of characters together into tokens
- More powerful alternatives
  - Wordpiece
    - Merge tokens based on what increases language model probability of a training corpus
  - SentencePiece/unigram
    - Start with huge vocabulary of all frequent sequences of characters, remove sequences that don't have a high probability in the training corpus iteratively

Slide adapted from Jurafsky & Martin

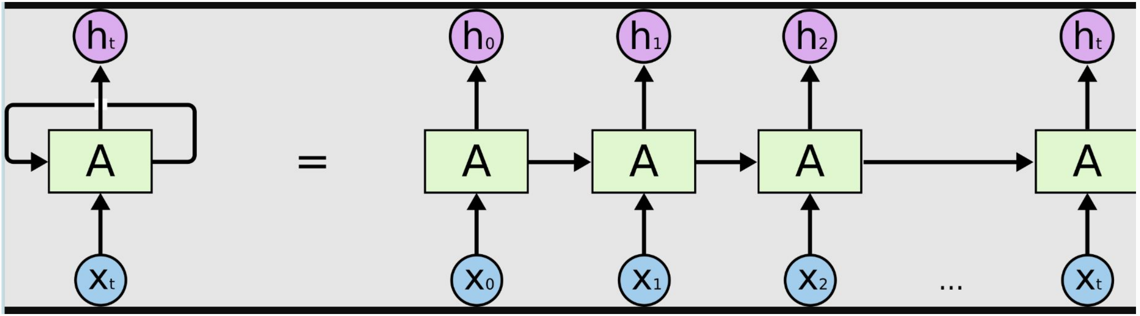
Encoder-decoder MT systems

Which model to train?

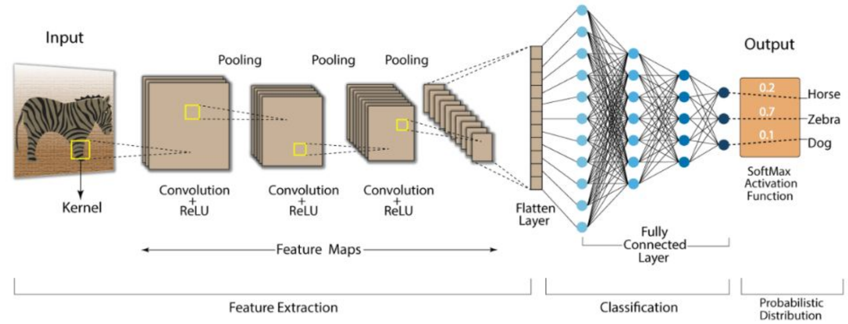
 of course. But why?

Slide credit: Sabit Hassan

Recap: Neural Networks



RNNs: Sequential. Good for time-series data



CNNs: focuses on “patches”. Good for images

Slide credit: Sabit Hassan

Try processing this text like a CNN/RNN:

"Life will make you do crazy things. That's why it's fun!"

Slide credit: *Sabit Hassan*

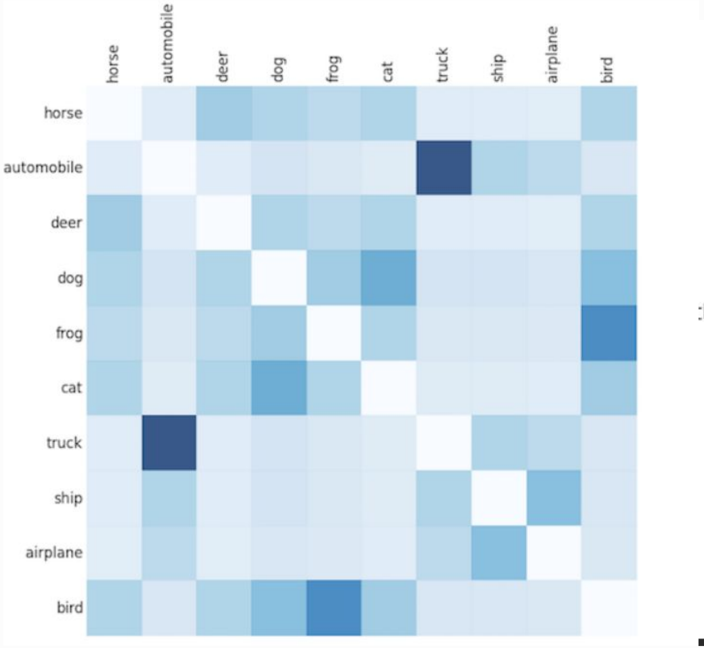
RNN/CNN - not how humans process text.

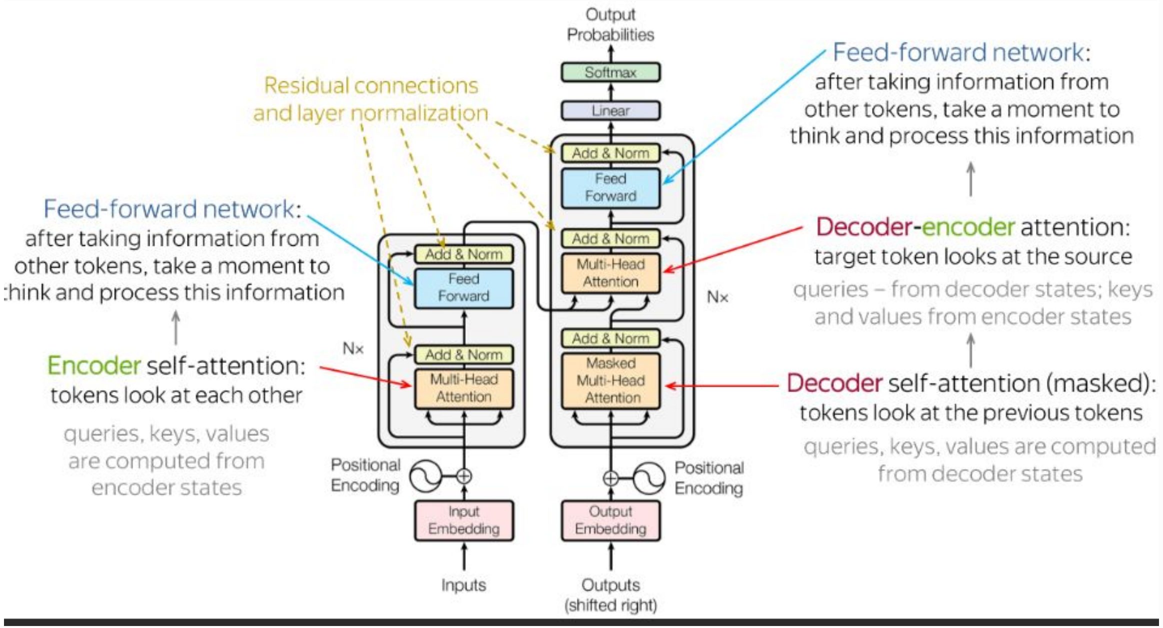
We make sense of text as a whole, focusing on different parts.

“Life will make you do crazy things. That's why it's fun!”

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

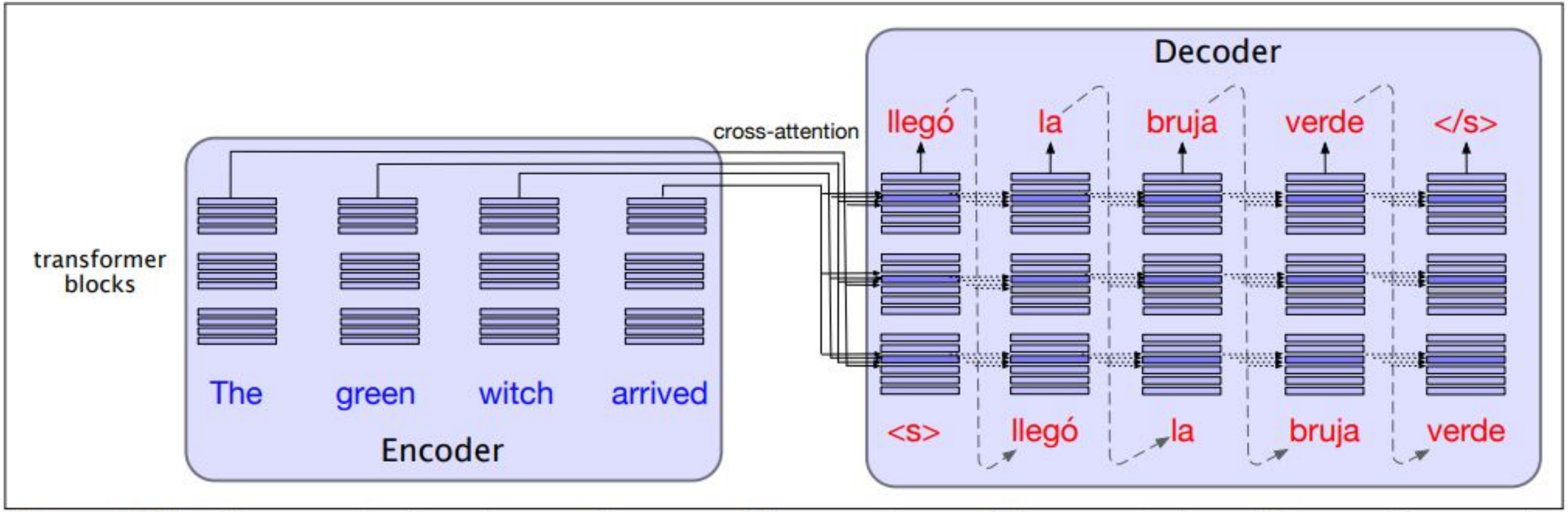


Figure 10.5 The encoder-decoder transformer architecture for machine translation. The encoder uses the transformer blocks we saw in Chapter 9, while the decoder uses a more powerful block with an extra **cross-attention** layer that can attend to all the encoder words. We'll see this in more detail in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

Slide adapted from Jurafsky & Martin

MT for low-resource languages

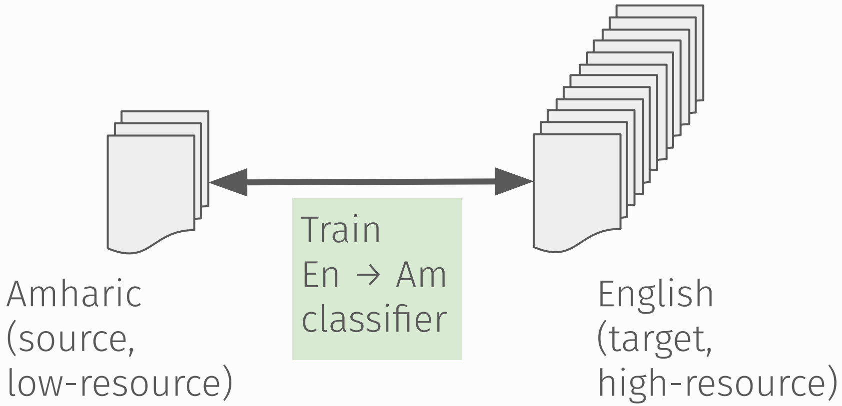
Low-resource MT

- No large-scale parallel corpora for many languages
- **Method 1: Backtranslation**
    - If have large corpora in target lang

Slide adapted from Sabit Hassan

Low-resource MT

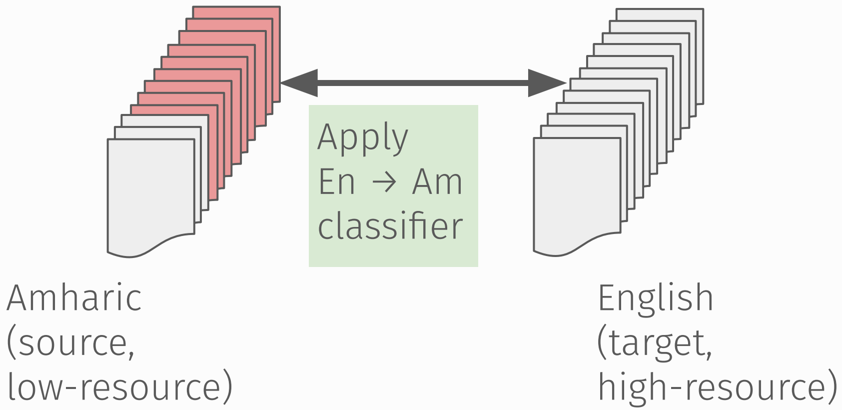
- No large-scale parallel corpora for many languages
- **Method 1: Backtranslation**
  - If have large corpora in target lang
  - Train reverse translation engine: target-source on small parallel corpus



Slide adapted from Sabit Hassan

Low-resource MT

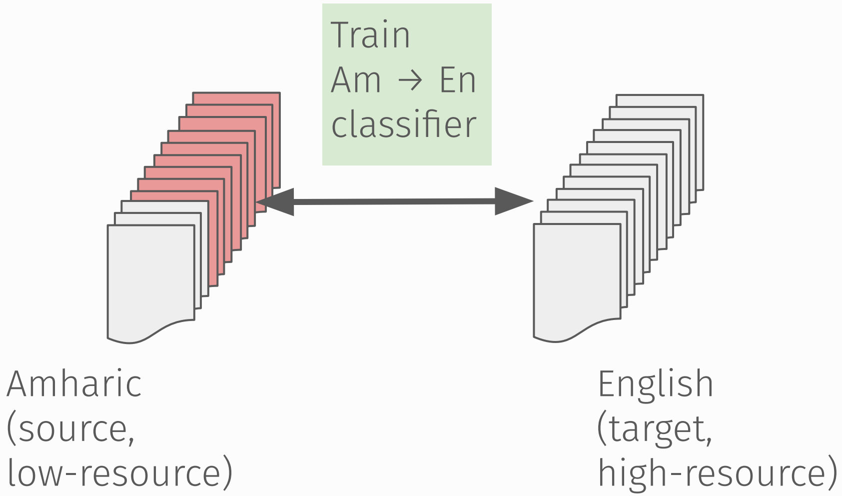
- No large-scale parallel corpora for many languages
- **Method 1: Backtranslation**
  - If have large corpora in target lang
  - Train reverse translation engine: target-source on small parallel corpus
  - Translate the large corpora and add that (lower-quality) data



Slide adapted from Sabit Hassan

Low-resource MT

- No large-scale parallel corpora for many languages
- **Method 1: Backtranslation**
  - If have large corpora in target lang
  - Train reverse translation engine: target-source on small parallel corpus
  - Translate the large corpora and add that (lower-quality) data



Slide adapted from Sabit Hassan

Low-resource MT

- No large-scale parallel corpora for many languages
- **Method 2: Multilingual model**
  - Train model with many language pairs
  - Can draw information from similar language

Slide adapted from Sabit Hassan

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 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

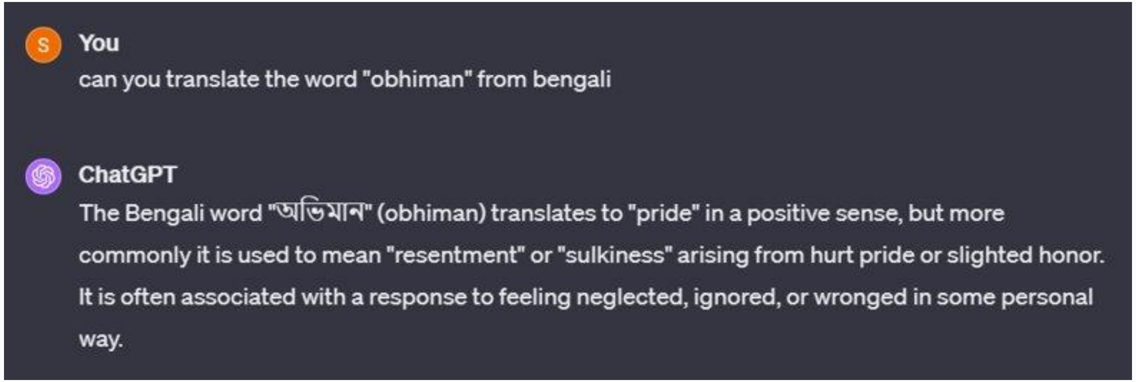
- **chrP**: percentage of character 1-grams, 2-grams, ..., k-grams in the hypothesis that occur in the reference, averaged.
- **chrR**: percentage of character 1-grams, 2-grams,..., k-grams in the reference that occur in the hypothesis, averaged.

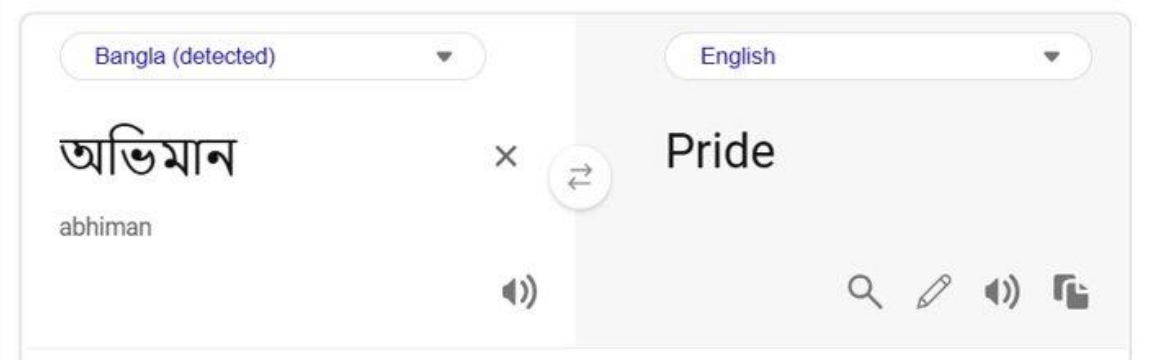
$$\mathrm{chrF}\beta = (1+\beta^2)\frac{\mathrm{chrP}\cdot\mathrm{chrR}}{\beta^2\cdot\mathrm{chrP}+\mathrm{chrR}}$$

Slide adapted from Sabit Hassan

Are we done?

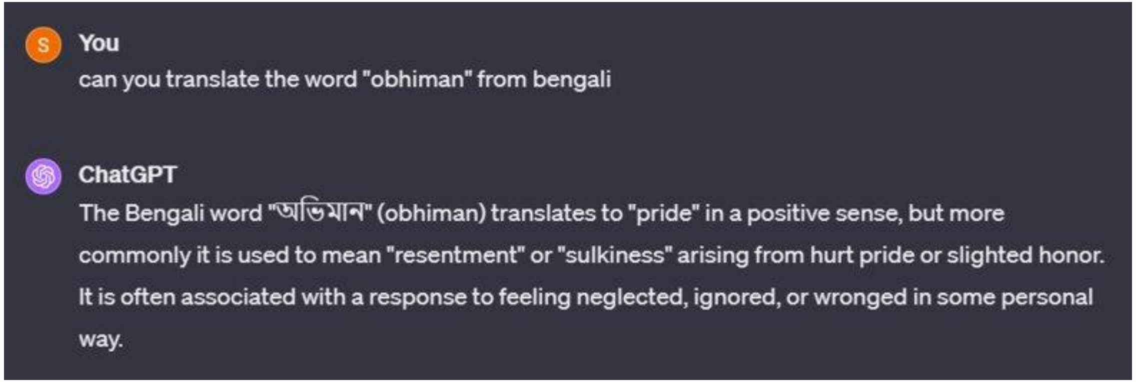
MT still fails to capture nuances of language

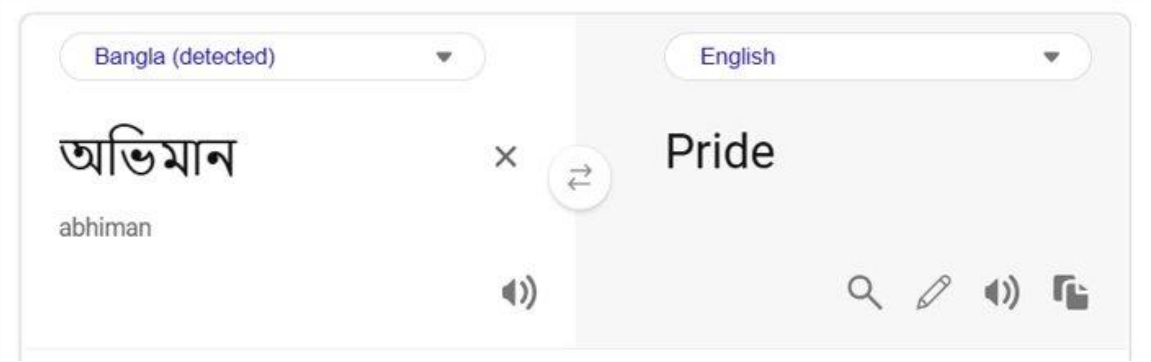




Slide credit: Sabit Hassan

MT still fails to capture nuances of language





অভিমান (Obhiman) is the feeling of being hurt by someone close to you. But you are not going to say anything about it. There is no exact word in English for this.

Slide credit: Sabit Hassan

What to do?

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

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

We shouldn’t just think about loss functions, model architecture etc.

We need cross-cultural, cross-disciplinary research

Slide credit: Sabit Hassan

Conclusion

- Sentences must be aligned in parallel corpora
- Subword tokenization is used for a shared vocabulary between languages
- Encoder-decoder transformer MT systems use cross-attention to attend to the source language input when generating the target language output
- Backtranslation and multilingual models are methods for handling a lack of parallel data (low-resource languages)
- 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?