Source: Vera Hembitskaya

CS 2731 Introduction to Natural Language Processing

Session 22: Machine translation part 2

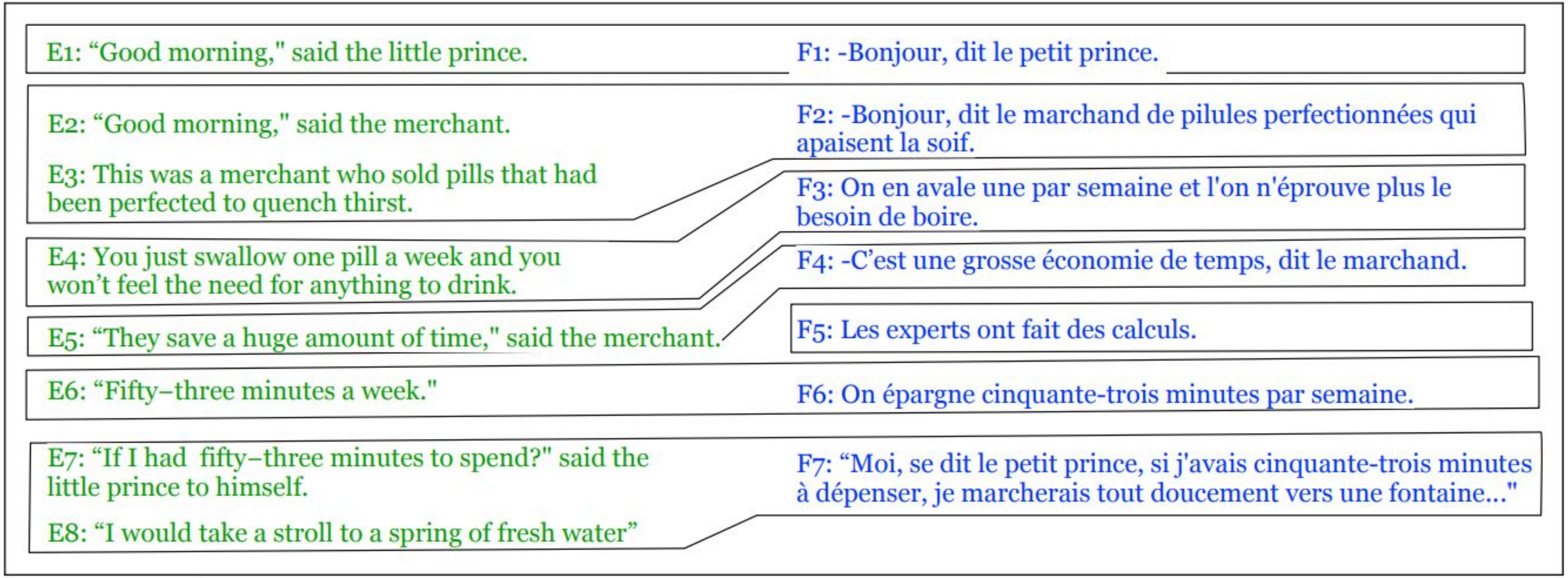
Sabit Hassan

November 15, 2023

University of Pittsburgh | School of Computing and Information

Last time: parallel corpora

Sentence alignment



E1: “Good morning," said the little prince.

E2: “Good morning," said the merchant.

E3: This was a merchant who sold pills that had been perfected to quench thirst.

E4: You just swallow one pill a week and you won’t feel the need for anything to drink.

E5: “They save a huge amount of time," said the merchant.

E6: “Fifty–three minutes a week."

E7: “If I had fifty–three minutes to spend?" said the little prince to himself.

E8: “I would take a stroll to a spring of fresh water”

F1: -Bonjour, dit le petit prince.

F2: -Bonjour, dit le marchand de pilules perfectionnées qui apaisent la soif.

F3: On en avale une par semaine et l'on n'éprouve plus le besoin de boire.

F4: -C’est une grosse économie de temps, dit le marchand.

F5: Les experts ont fait des calculs.

F6: On épargne cinquante-trois minutes par semaine.

F7: “Moi, se dit le petit prince, si j'avais cinquante-trois minutes à dépenser, je marcherais tout doucement vers une fontaine..."

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

Multilingual embedding space

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

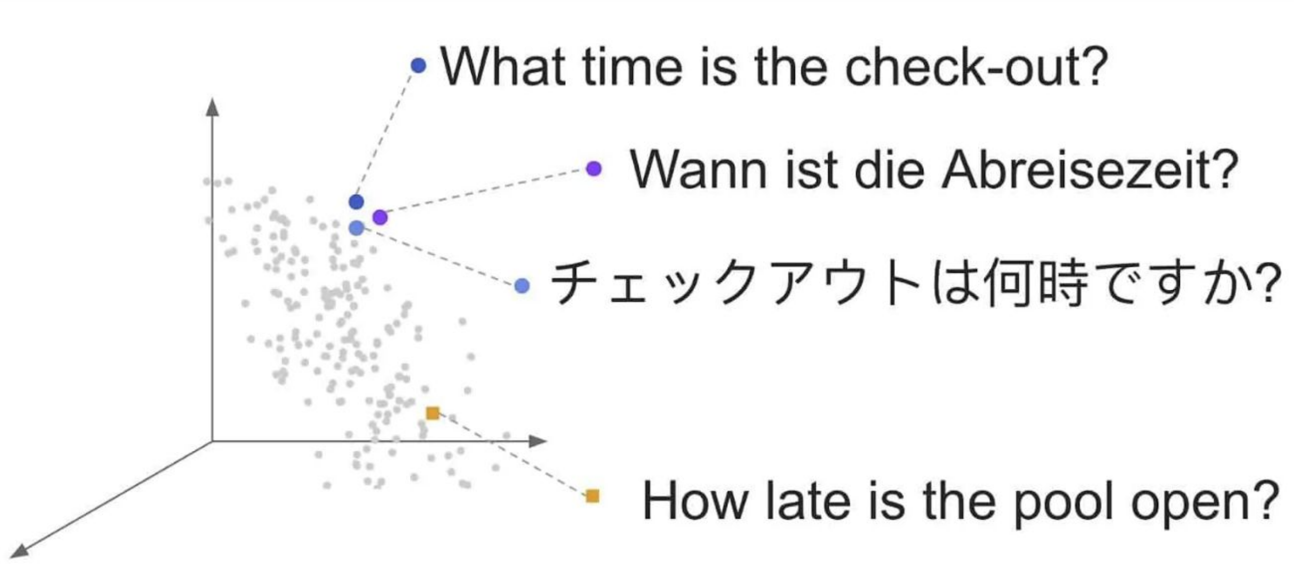
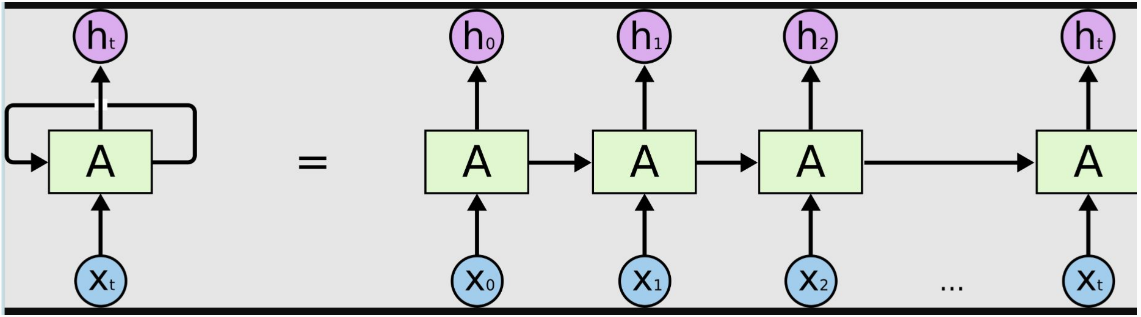

What time is the check-out?
Wann ist die Abreisezeit?
チェックアウトは何時ですか?
How late is the pool open?

Figure credit: Megagon Labs

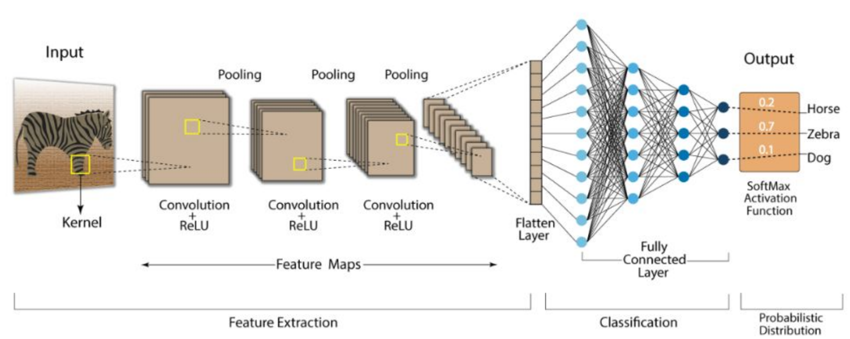
Which model to train?

of course. But why?

Recap: Neural Networks



RNNs: Sequential. Good for time-series data



CNNs: focuses on “patches”. Good for images

Try processing this text like a CNN/RNN:

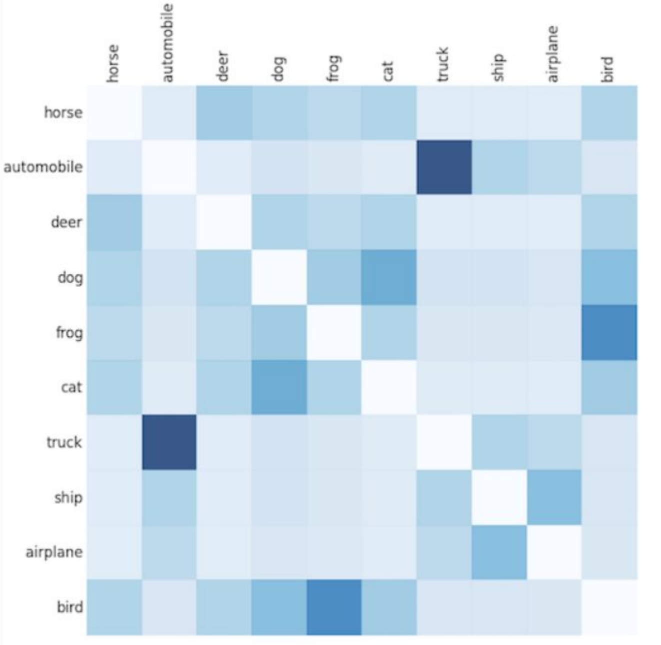
“Life will make you do crazy things. That's why it's fun!”

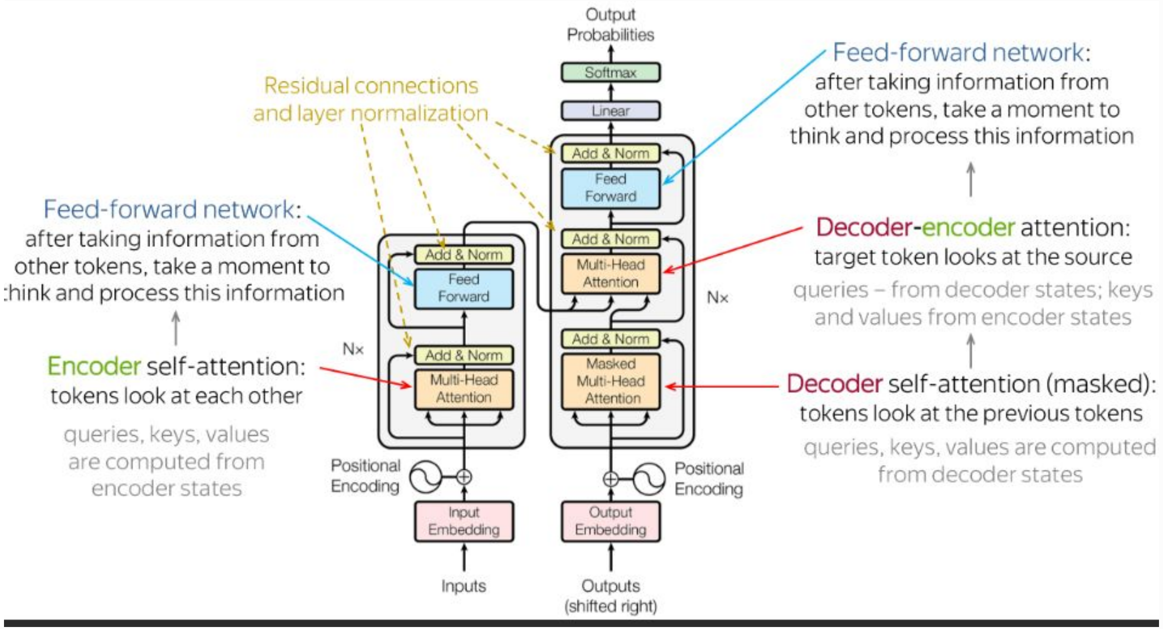
RNN/CNN - not how humans process text.

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!”

Recap: Attention and Transformers





Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Feed-forward network: after taking information from other tokens, take a moment to hink and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

- Focus of different parts of input.
- How we humans process language

Multilingual embedding space

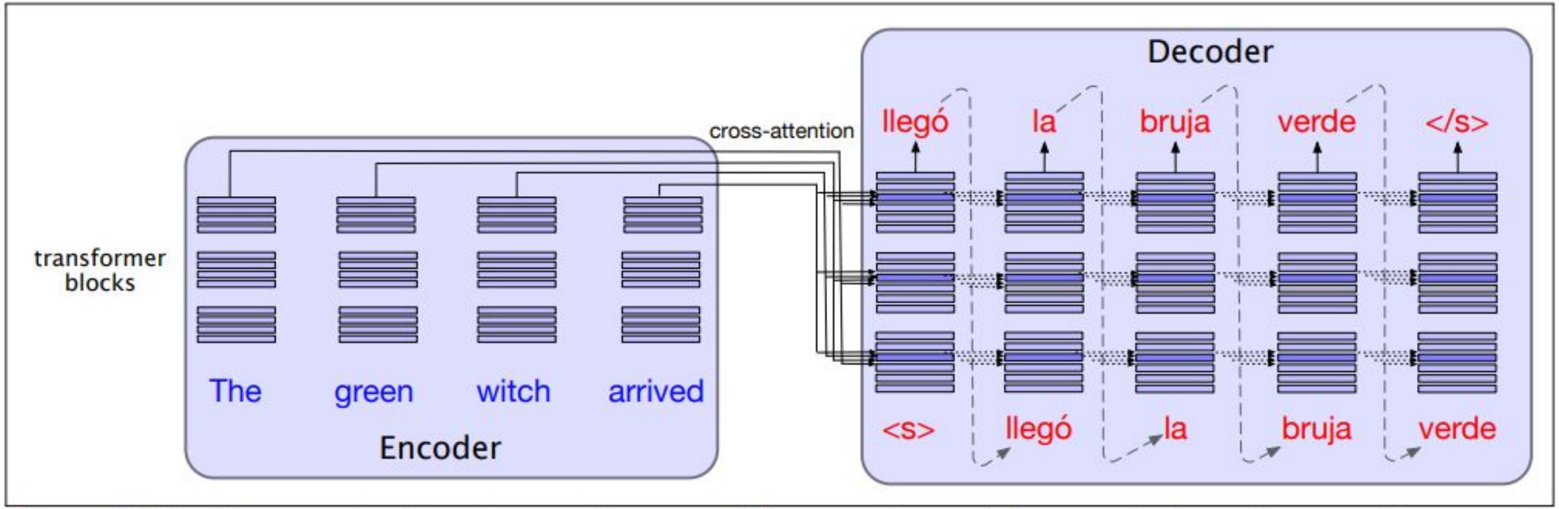


Figure 10.5 The encoder-decoder transformer architecture for machine translation. The encoder uses the transformer blocks we saw in Chapter 9, while the decoder uses a more powerful block with an extra **cross-attention** layer that can attend to all the encoder words. We'll see this in more detail in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

Low-resource MT

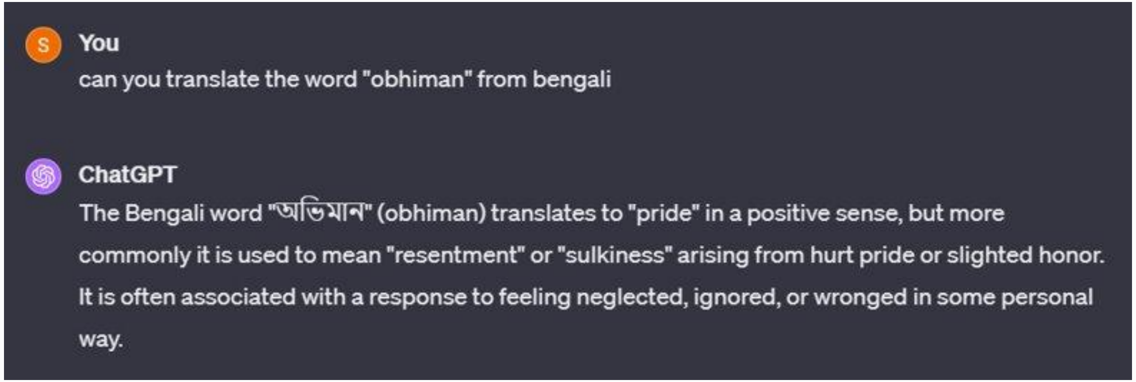
- No large-scale parallel corpora for many languages
- **Method 1: Backtranslation**
  - Assume large corpora in target lang
  - Small parallel corpora between source-target
  - Train reverse translation engine: target-source
  - Translate the large corpora
- **Method 2: Multilingual model**
  - Train model with many language pairs
  - Draw information from similar language

MT Evaluation

- **Adequacy**: how well translation captures exact meaning of the source sentence
- **fluency**: how fluent the translation is in the target language
- **Human evaluation:** Rate/edit translations. Expensive
- **Automatic evaluation:** character/word overlap, embedding-space
  - chrF, BLEU - overlap-based
  - METEOR, BERTScore - embedding-based

Are we done?

MT still fails to capture nuances of language

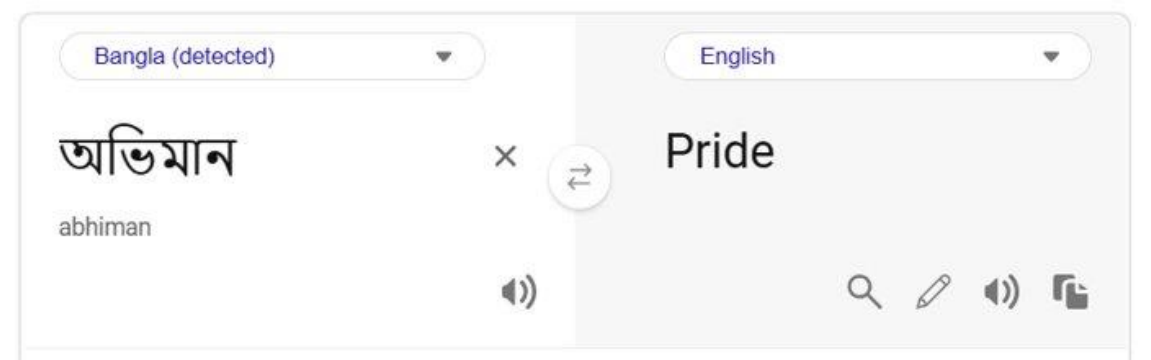


You

can you translate the word "obhiman" from bengali

ChatGPT

The Bengali word "অভিমান" (obhiman) translates to "pride" in a positive sense, but more commonly it is used to mean "resentment" or "sulkiness" arising from hurt pride or slighted honor. It is often associated with a response to feeling neglected, ignored, or wronged in some personal way.



Bangla (detected)

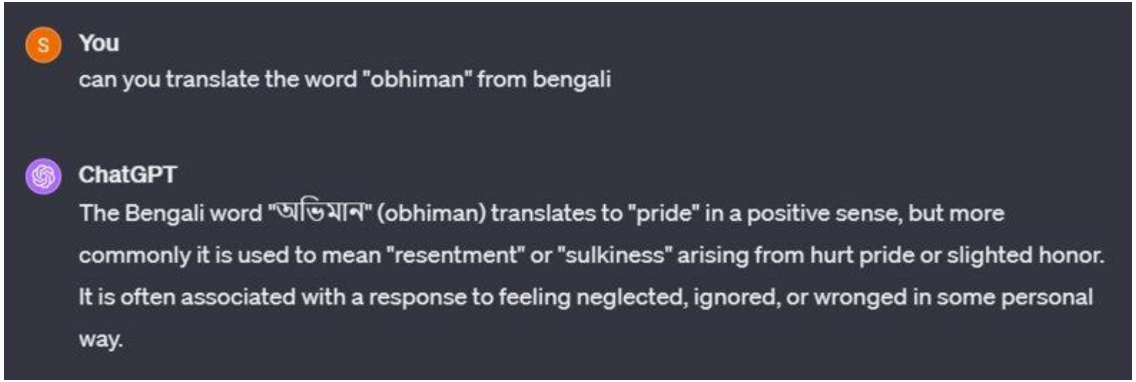
অভিমান

abhiman

English

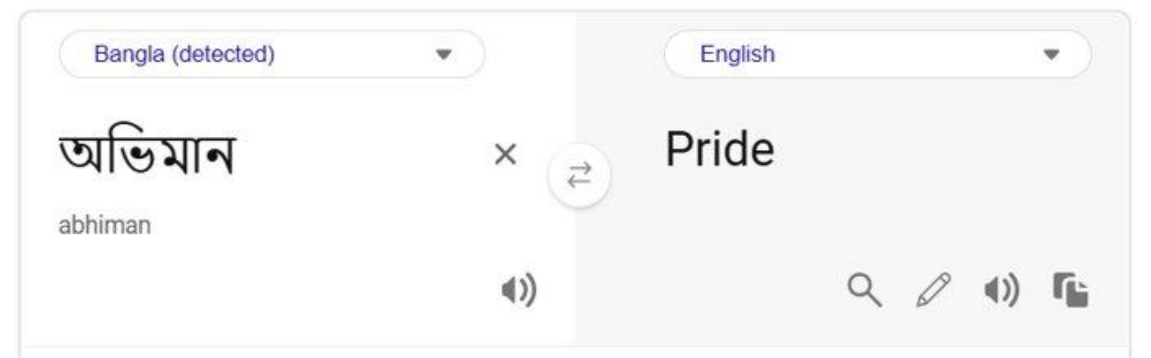
Pride

MT still fails to capture nuances of language



You
can you translate the word "obhiman" from bengali

ChatGPT
The Bengali word "অভিমান" (obhiman) translates to "pride" in a positive sense, but more commonly it is used to mean "resentment" or "sulkiness" arising from hurt pride or slighted honor. It is often associated with a response to feeling neglected, ignored, or wronged in some personal way.



Bangla (detected)

অভিমান

abhiman

English

Pride

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.

What to do?

Fixing MT

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

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

We need cross-culture, cross-disciplinary research

Questions?