TRANSLATION IS LIKE CHOPPING AN ONION –
FIRST, YOU THINK YOU'LL MANAGE IT.

AND THEN YOU END UP CRYING IN THE KITCHEN.

Source: Vera Hembitskaya

CS 2731 Introduction to Natural Language Processing

Session 22: Machine translation part 1

Michael Miller Yoder

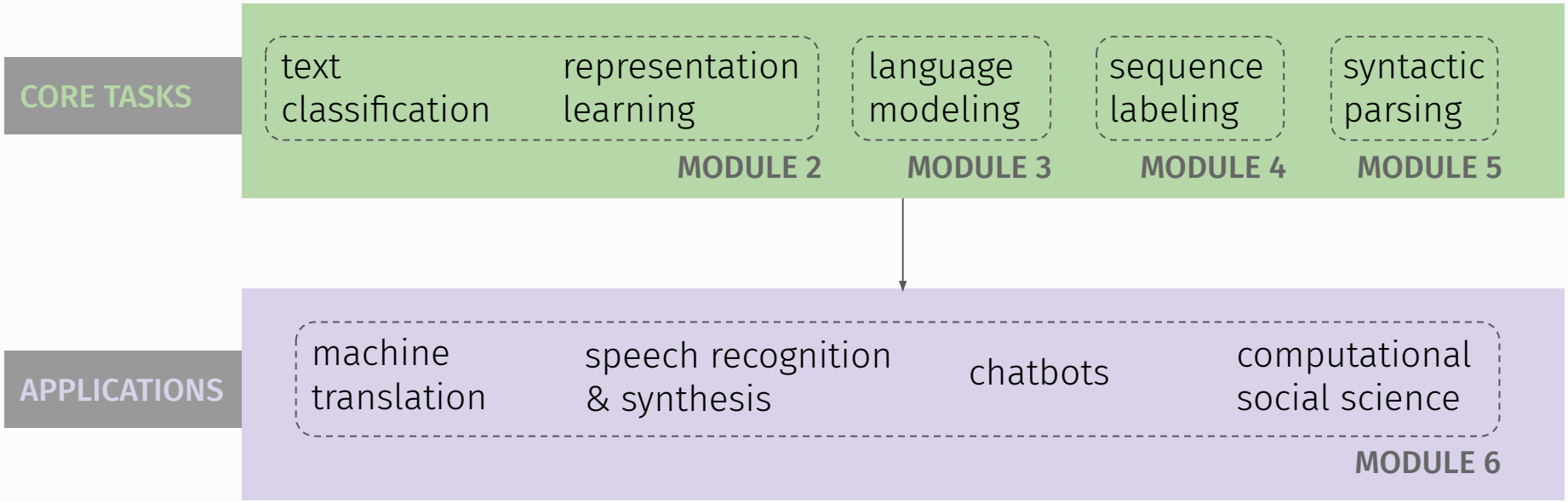
November 13, 2023

School of Computing and Information

Course logistics

- Basic working project system **due this Thu 11-16**
  - 1-2 pages, in ACL LaTeX format that final report will be in
- Office hours the same times, but switching instructor/TA
  - Michael's office hours this week:
    2:45-3:45pm Tue 11-14, Sennott Square 6505
  - Pantho's office hours this week:
    1:30-2:30pm Wed 11-15, Sennott Square 5106
- Pantho will be giving the lecture on Wed

Core tasks and applications of NLP



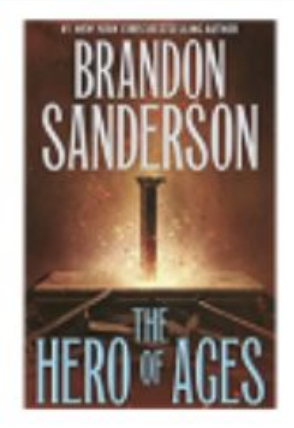
Overview: Machine translation part 1

- History of machine translation (MT)
- Translation in practice
- Why is translation difficult?
- Parallel corpora
  - Sentence alignment

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”

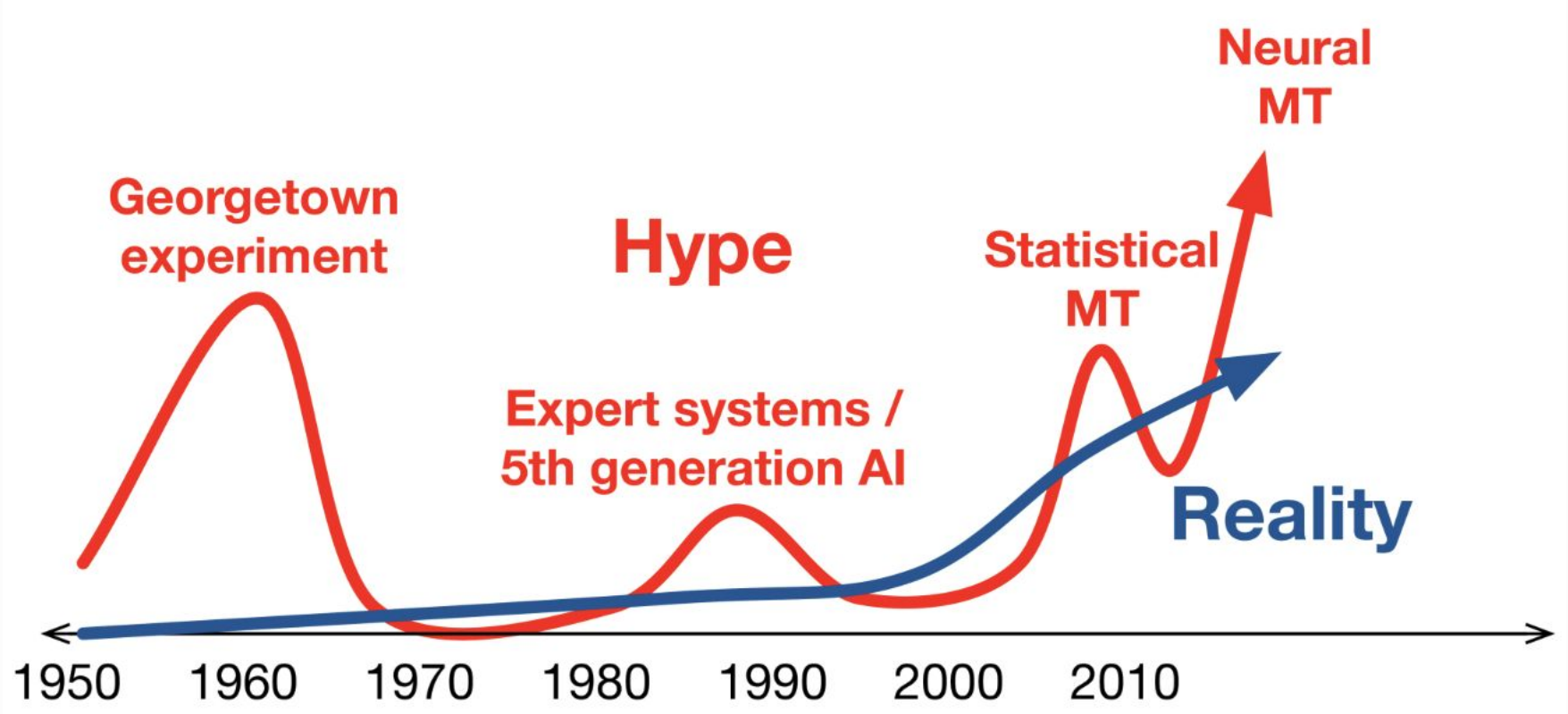


→

Slide adapted from Lori Levin

History of machine translation

MT history: hype vs reality



Slide adapted from Mohit Iyyer, Richard Socher

When did people start using computers to translate?

- Roughly 1950s
- Research stopped in the US for about 15-20 years after a 1967 report deemed it impossible
- Research resumed in the US in the early 1980s

Slide adapted from Lori Levin, Chris Manning

What did early MT systems look like?

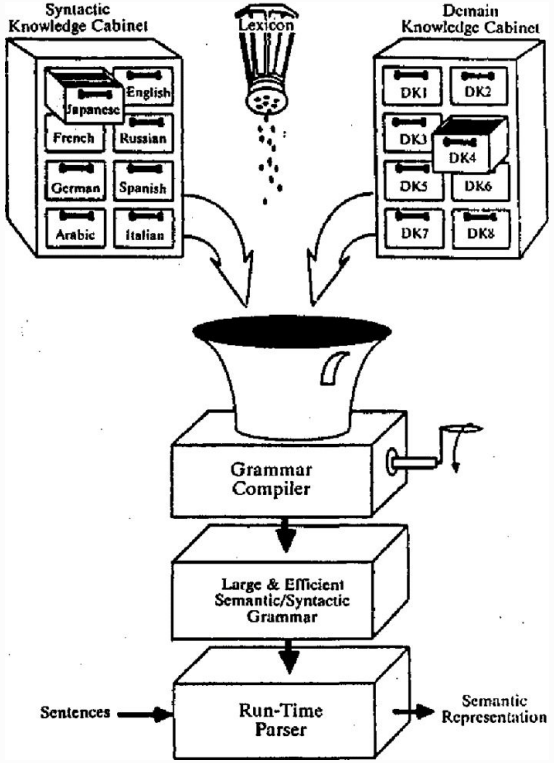
Human linguists wrote elaborate rules involving syntax, semantics, etc

```
(<S> <--> (<V>)
     ((x0 = x1)))

(<S> <--> (<NP> <S>)
     (((x2 subj-case) = *defined*)
      ((x2 subj-case) = (x1 case))
      (x0 = x2)
      ((x0 subj) = x1)))

(<S> <--> (<NP> <S>)
     (((x2 obj-case) = *defined*)
      ((x2 obj-case) = (x1 case))
      (x0 = x2)
      ((x0 obj) = x1)))
```

```
(emap *insert
   <=1=> insert ((CAT v) (SUBCAT trans))
   (role =sem (*physical-action))
   (:agent =syn (SUBJECT))
   (:theme =syn (DOBJECT))
   (:goal =syn (PPADJUNCT
               ((PREP into) (CAT n)))))
```



Slide adapted from Lori Levin

Learning to translate from data

Since the late 1980s, Machine Translation researchers have been using parallel corpora to train Machine Translation systems.

	ENGLISH	MANDARIN
1	i **wanna** live in a wes anderson world	我想要生活在Wes Anderson的世界里
2	Chicken soup, corn never truly digests. **TMI**.	鸡汤吧，玉米神马的从来没有真正消化过.恶心
3	To DanielVeuleman **yea iknw imma** work on that	对DanielVeuleman说，是的我知道，我正在向那方面努力
4	**msg 4** Warren G his **cday** is today 1 **yr** older.	发信息给Warren G，今天是他的生日，又老了一岁了。
5	Where **the hell** have you been all these years?	这些年你**TMD**到哪去了

	ENGLISH	ARABIC
6	It's **gonna** be a warm week!	الاسبو ع **الياي** حر
7	onni this gift only **4 u**	أوني هذة الهدية فقط لك
8	sunset in aqaba **:)**	غروب الشمس في العقبة:)
9	RT @MARYAMALKHAWAJA: there is a call for widespread protests in #bahrain **tmrw**	هناك نداء لظاهرات في عدة مناطق غدا

Figure credit: Wang Ling, Microblogs as parallel corpora, 2013
Slide adapted from Lori Levin

Statistical machine translation (1990s-2010s)

- Core idea: Learn a probabilistic model from data
- For French -> English, we want to find best English sentence *y*, given French sentence *x*
- Use Bayes Rule to break this down into two components to be learned separately:

$$\text{argmax}_y P(y|x) = \text{argmax}_y \underbrace{P(x|y)}\underbrace{P(y)}$$

Slide adapted from Chris Manning

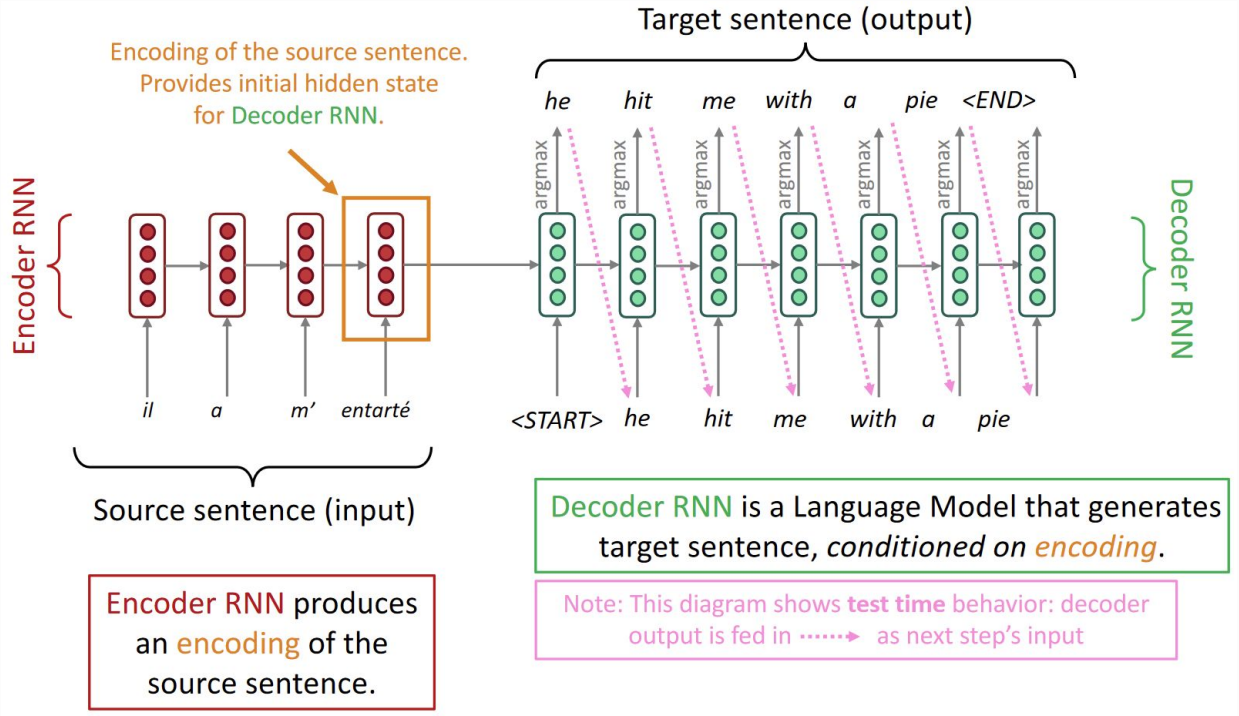
Statistical machine translation (1990s-2010s)

- The best SMT systems were extremely complex
  - Hundreds of important details
- Systems had many separately-designed subcomponents
  - Lots of feature engineering
  - Need to design features to capture particular language phenomena
- Required compiling and maintaining extra resources, like tables of equivalent phrases
  - Lots of human effort to maintain
- Repeated effort for each language pair

Slide adapted from Chris Manning

Neural machine translation (2010s on)

- Single end-to-end neural network
- Encoder-decoder (sequence-to-sequence, seq2seq) framework

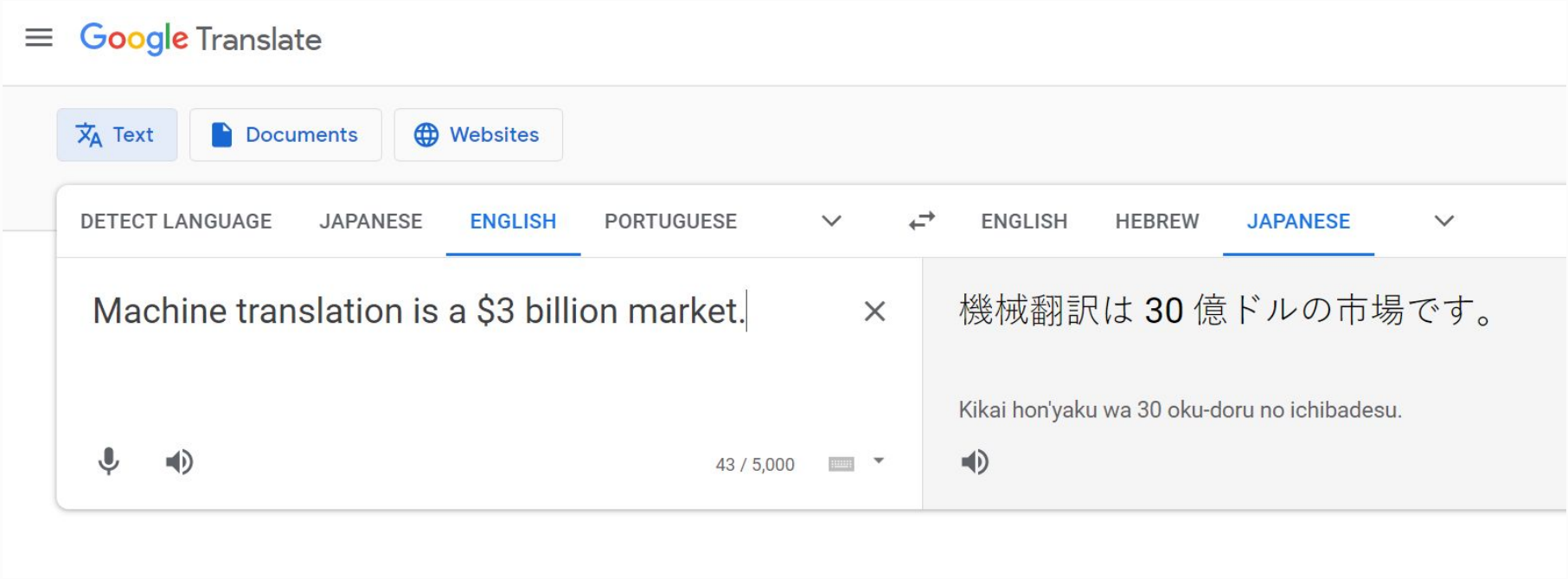


Slide adapted from Chris Manning

Translation in practice

Machine translation is a $3 billion market

Translation of text



Slide adapted from Lori Levin

Machine translation is a $3 billion market

Translation of speech

Person: Alexa, how do you say, “I hate this movie” in Japanese.

Alexa: “I hate this movie” in Japanese is “Kono eiga wa kirai da.”

Person : Alexa, how do you say, “I hate this movie in Japanese” in Japanese.

Alexa: “I hate this movie in Japanese” in Japanese is “Kono eiga wa nihongo de wa kirai da.”

Real time translation of meetings is also now viable.

Slide adapted from Lori Levin

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

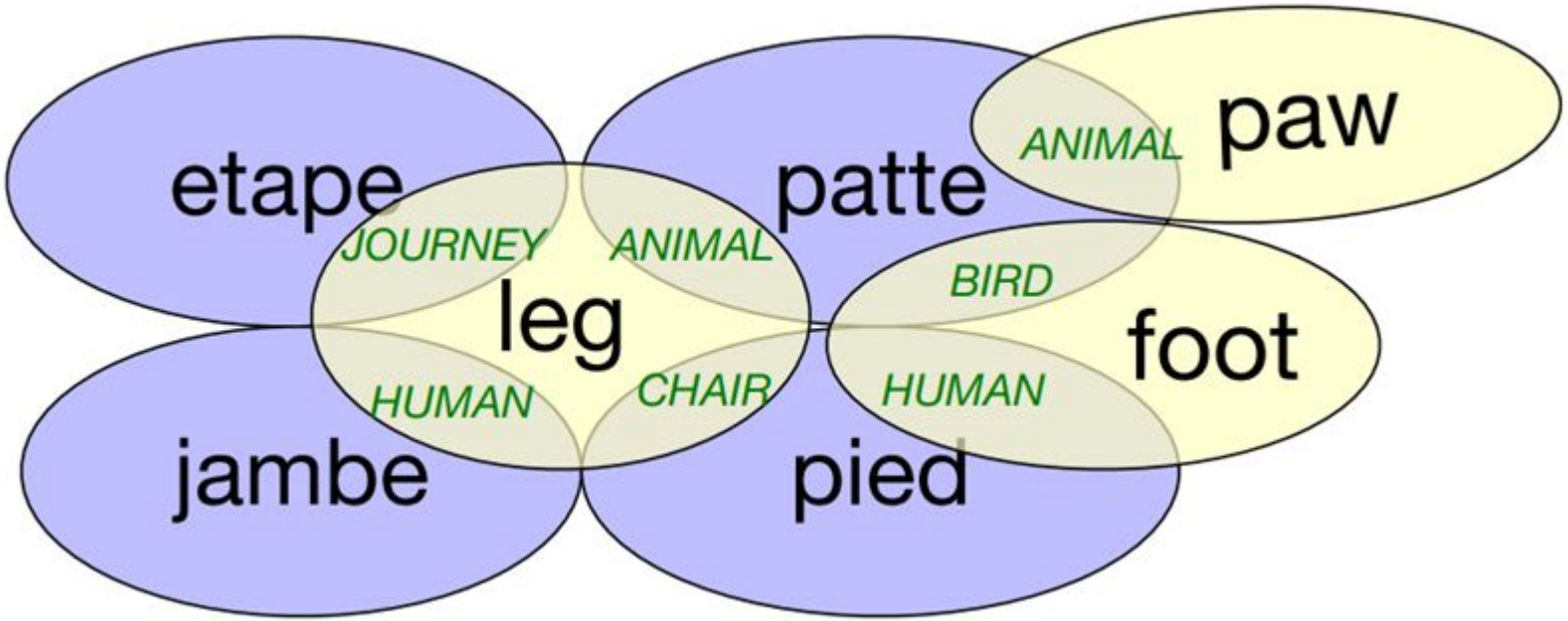
Slide adapted from Lori Levin

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: *katabt 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 _____

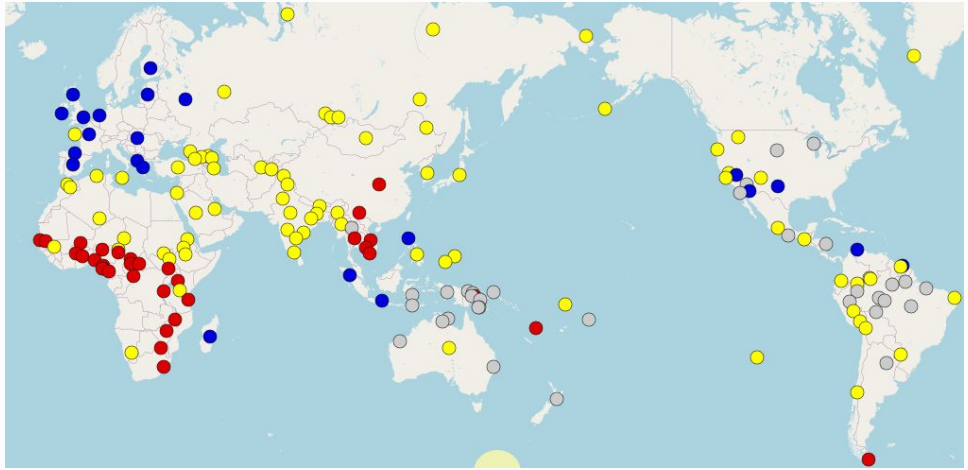
сумо _____

Problem by Adriana Solovyova
Slide adapted from Lori Levin

What is difficult about translation?

- People in NLP and MT have reduced “language divergences” to six major word order features from WALS, or seven lexical features
- But language typology is a system of “morphosyntactic strategies”, of which there are 1000s

Yellow: X is big from Y, or X is big to Y
Red: X is big, exceeds Y
Grey: X is big, Y is small
Blue: X is big than Y



Slide adapted from Lori Levin

But the picture is not so gloomy

- MT researchers have made much progress on handling language divergence
- Use data from typologically similar languages
- Use a multilingual model trained on many typologically different languages

Slide adapted from Lori Levin

Why is translation difficult? Style and genre

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

Chinese: Daiyu alone at bed top think baochai.

English: Daiyu alone on the bed thought about baochai.

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

Why is translation difficult? Style and genre

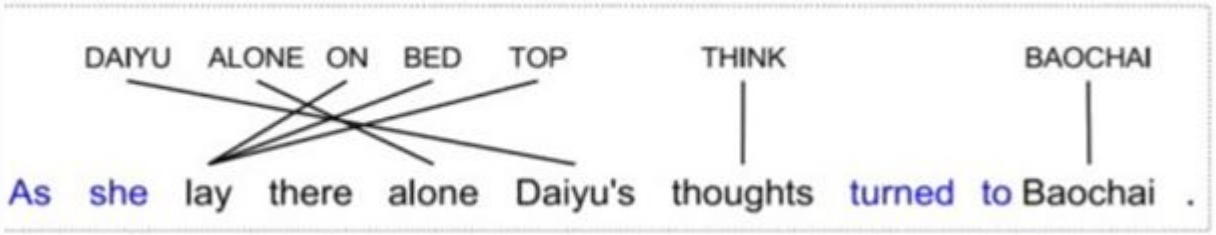
錨玉自在枕上感念寶釵

dai yu zi zai zhen shang gan nian bao chai

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
3. Tokenization
    - Split words into sub-word units, e.g., using BPE (Byte Pair Encoding)

Slide adapted from Lori Levin

Parallel corpora

芋见 BAO

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

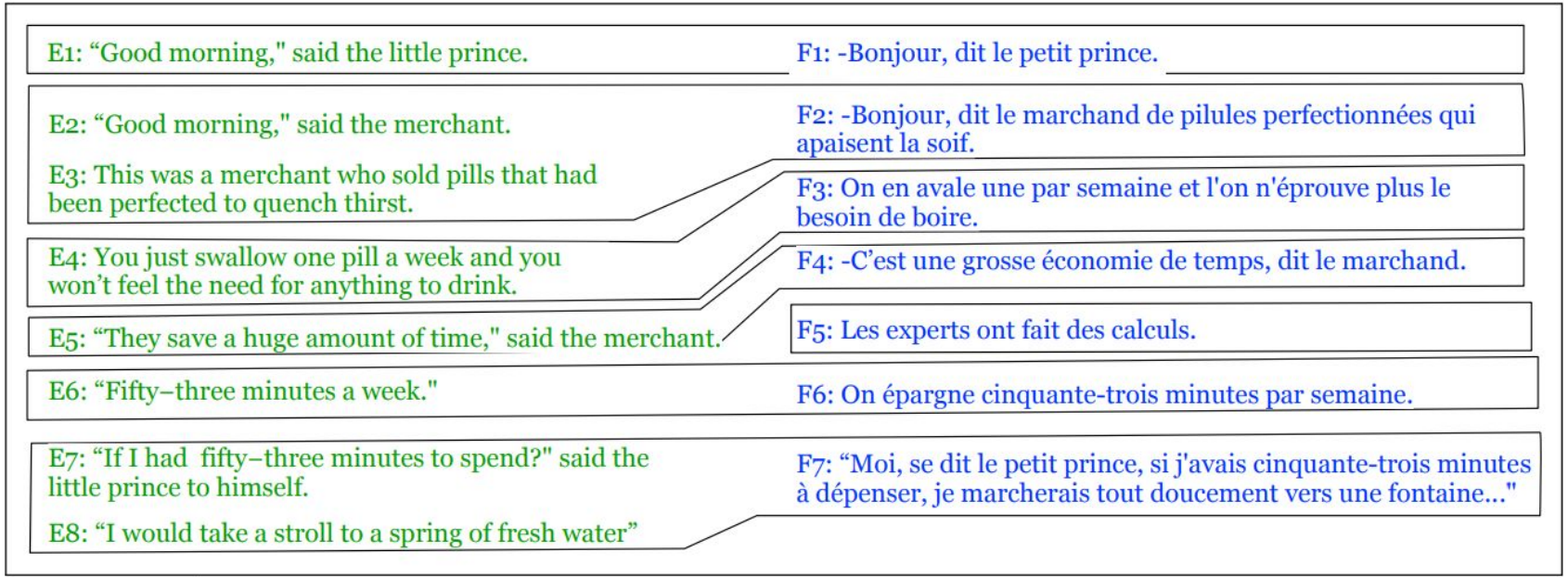


Figure 10.17 A sample alignment between sentences in English and French, with sentences extracted from Antoine de Saint-Exupery’s *Le Petit Prince* and a hypothetical translation. Sentence alignment takes sentences $e_1,...,e_n$, and $f_1,...,f_n$ and finds minimal sets of sentences that are translations 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?
2. Alignment algorithm: uses scores between spans to find a good alignment between documents

Multilingual embedding space

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

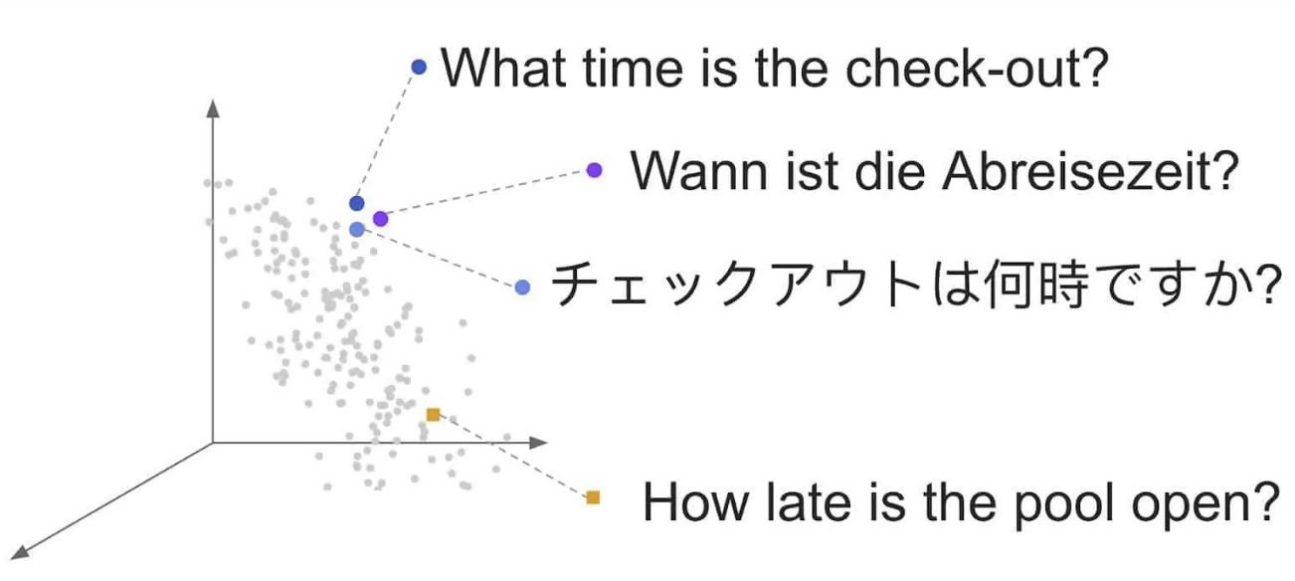


Figure credit: Megagon Labs

Sentence alignment: cost function and alignment alg

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

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

- 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

Wrapping up

- Modern machine translation methods use the neural encoder-decoder framework
- 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
- Sentences must be aligned in parallel corpora
- Subword tokenization is used for a shared vocabulary between languages

Questions?