Multilingual and Parallel Corpora

Machine Translation (ctd)

Amir Zeldes

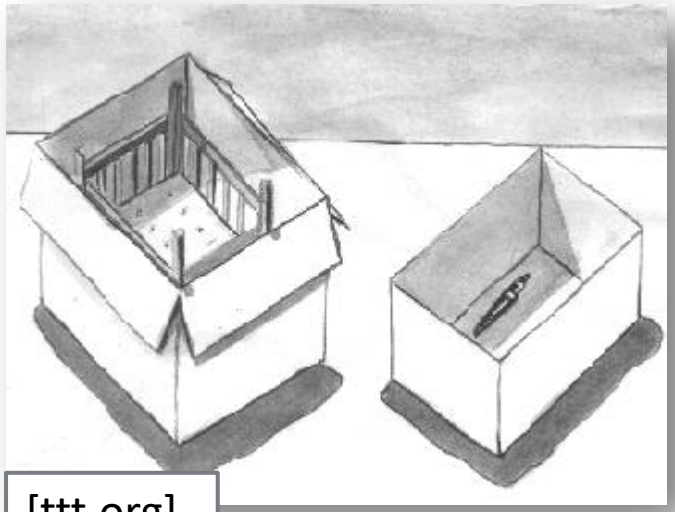
amir.zeldes@georgetown.edu

Homework assignment – discussion

- Votes for most similar text:
  - 7/9 in favor of **text 2**
  - Reasons:
    - V3 omits ‘crucial information’: 5
    - V2 matches feel of V1 better, “direct translations”: 4
    - Paragraph split, length: 3
    - Number of shifts (stable?): 2
  - Caveats:
    - V3 has “more direct translations”: 2
    - V2 “more explicit”: 2

Bar Hillel – Nonfeasibility of FAHQT

- Basic problem:
  - Word Sense Disambiguation (WSD)
- Approaches:
  - Idioglossaries – ambiguity and specialized lexicon
  - Context
- But consider:
  - A ‘pen’ is a type of burglary.
  - Adam committed a pen.



[ttt.org]

Disambiguation

- WSD is not uniquely a problem of MT
- Many approaches
- Example: (non-)named entity recognition (NER)
- System: https://corpling.uis.georgetown.edu/xrenner/

Rule based approaches to MT

- We could make a (large) list of rules:
  - “not to mention” + NP >
  - “ganz zu Schweigen von” + NP >

- If the language pair behaves regularly we can **map** their words and even word orders (a.k.a. **transfer**)
- Would we need to understand the content?

Example Japanese - Chinese

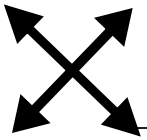
- Chinese is SVO, Japanese is SOV
- Japanese is agglutinative, Chinese is isolating
- Both languages use an ideographic script (in Japanese also two syllabaries)
- Often symbols are identical for the same concepts

➢For simple cases, can we just use symbol reordering?

Example Japanese - Chinese

- Simple declarative transitive sentence:

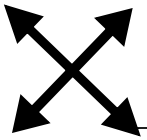
S	O	V
Student	English	learns
大学生は	英語を	学ぶ
大学生	学	英語
Student	learns	English
S	V	O



Example Japanese - Chinese

- Simple declarative transitive sentence:

S	O	V
Student	English	learns
大学生は	英語を	学ぶ
大学生	学	英語
Student	learns	English
S	V	O



Problems

1. Extensibility (new domains, new languages)
2. …

Interlanguage

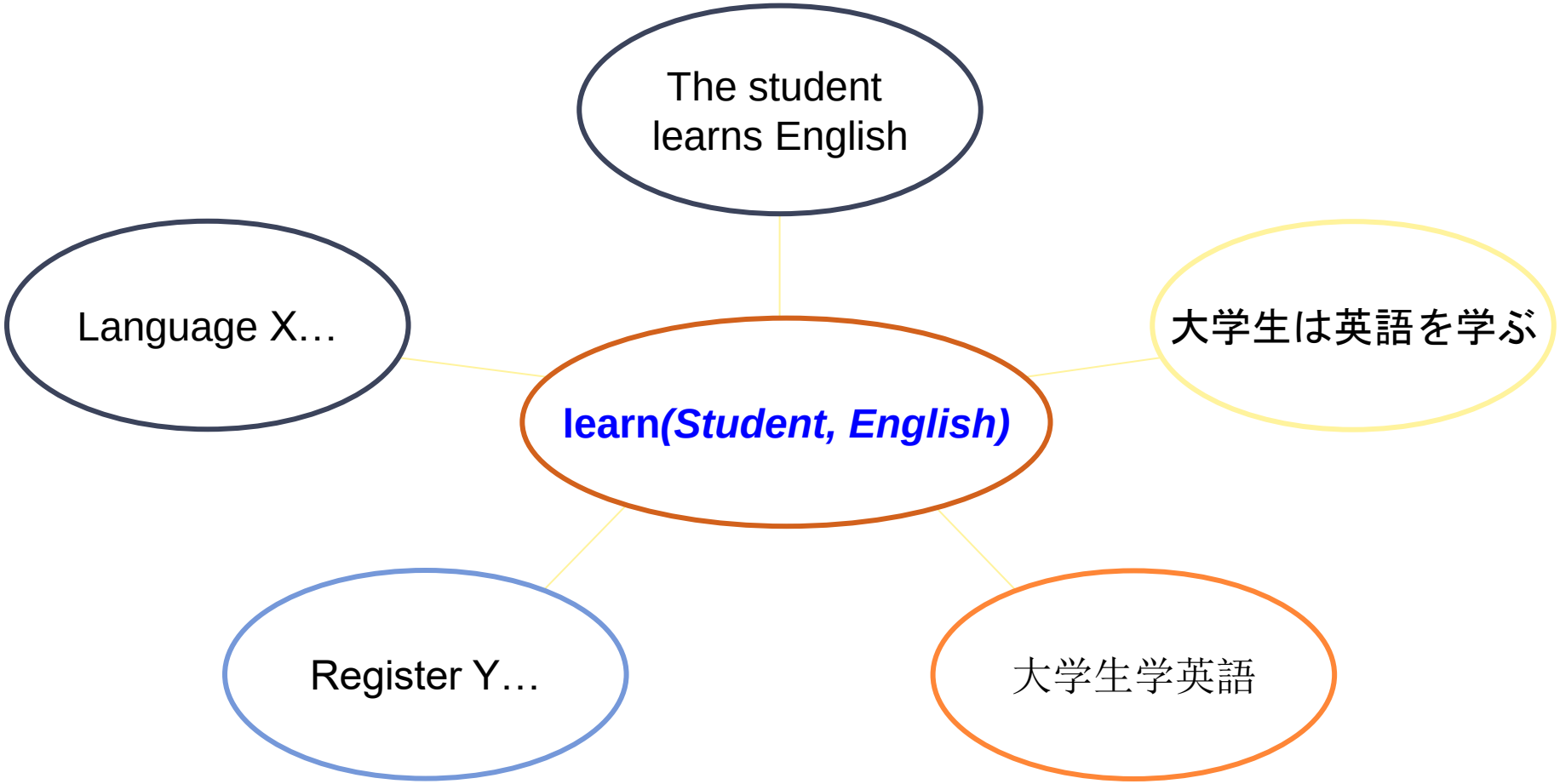
- We can try to translate the ST into a **semantic** representation which only contains language independent information
- Allows us to translate into any output TL via an intermediate – no pairwise-models, just a model of **content**

➢ Prerequisite: We need to be able to “understand” the text automatically

What is content?

- A formal semantic approach
- Utterances are understood as propositions
- Reduction to predicates and arguments, or functions and variables

Transfer via interlanguage



Problems

1. Extensibility (new domains, new languages)
2. Context dependence (the same input can have different translations in different situations)

➢ More rules

Problems

1. Extensibility (new domains, new languages)
2. Context dependence (the same input can have different translations in different situations)
3. Number of rules explodes exponentially
4. Disambiguation challenging (potentially a challenge for any approach)

Statistical approaches

- If we want to be able to translate an enormous amount of situations…
- … we need to look at an enormous amount of situations – a large parallel corpus

Direct translation model

- We match the largest possible substrings of an input
  - Not to mention
  - the problems
- Find translations
  - Ganz zu schweigen
  - die Probleme
- Assemble
  - Ganz zu schweigen + die Probleme ->

    Ganz zu schweigen **von** d**en** Problem**en**

Problems with direct translation

- Hard to exhaustively predict all cases that require modification in assembly
- We can notice regularities in training data:
  - Ganz zu schweigen always followed by ***von***
  - This is never followed by ***die***
  - …
- But the training data is small – just a parallel corpus
- Can't we learn what good German is like from a larger, monolingual corpus?

Homework assignment: Centauri/Arcturan [Knight 1997]

farok	**crrrok**	**hihok**	**yorok**	**clok**	**kantok**	**ok-yurp .**

1a. ok-voon ororok sprok .	1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .	2b. at-drubel at-voon pippat rrat dat .
3a. erok sprok izok hihok ghirok .	3b. totat dat arrat vat hilat .
4a. ok-voon anok drok brok jok .	4b. at-voon krat pippat sat lat .
5a. wiwok farok izok stok .	5b. totat jjat quat cat .
6a. lalok sprok izok jok stok .	6b. wat dat krat quat cat .
7a. lalok farok ororok lalok sprok izok enemok .	7b. wat jjat bichat wat dat vat eneat .
8a. lalok brok anok plok nok .	8b. wat lat pippat rrat nnat .
9a. wiwok nok izok kantok ok-yurp .	9b. totat nnat quat oloat at-yurp .
10a. lalok mok nok yorok ghirok clok .	10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .	11b. wat nnat arrat mat zanzanat .
12a. lalok rarok nok izok hihok mok .	12b. wat nnat forat arrat vat gat .

Homework assignment: Centauri/Arcturan [Knight 1997]

farok	**crrrok**	**hihok**	**yorok**	**clok**	**kantok**	**ok-yurp .**

1a. ok-voon ororok sprok .	1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .	2b. at-drubel at-voon pippat rrat dat .
3a. erok sprok izok hihok ghirok .	3b. totat dat arrat vat hilat .
4a. ok-voon anok drok brok jok .	4b. at-voon krat pippat sat lat .
5a. wiwok **farok** izok stok .	5b. totat jjat quat cat .
6a. lalok sprok izok jok stok .	6b. wat dat krat quat cat .
7a. lalok **farok** ororok lalok sprok izok enemok .	7b. wat jjat bichat wat dat vat eneat .
8a. lalok brok anok plok nok .	8b. wat lat pippat rrat nnat .
9a. wiwok nok izok kantok ok-yurp .	9b. totat nnat quat oloat at-yurp .
10a. lalok mok nok yorok ghirok clok .	10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .	11b. wat nnat arrat mat zanzanat .
12a. lalok rarok nok izok hihok mok .	12b. wat nnat forat arrat vat gat .

Homework assignment: Centauri/Arcturan [Knight 1997]

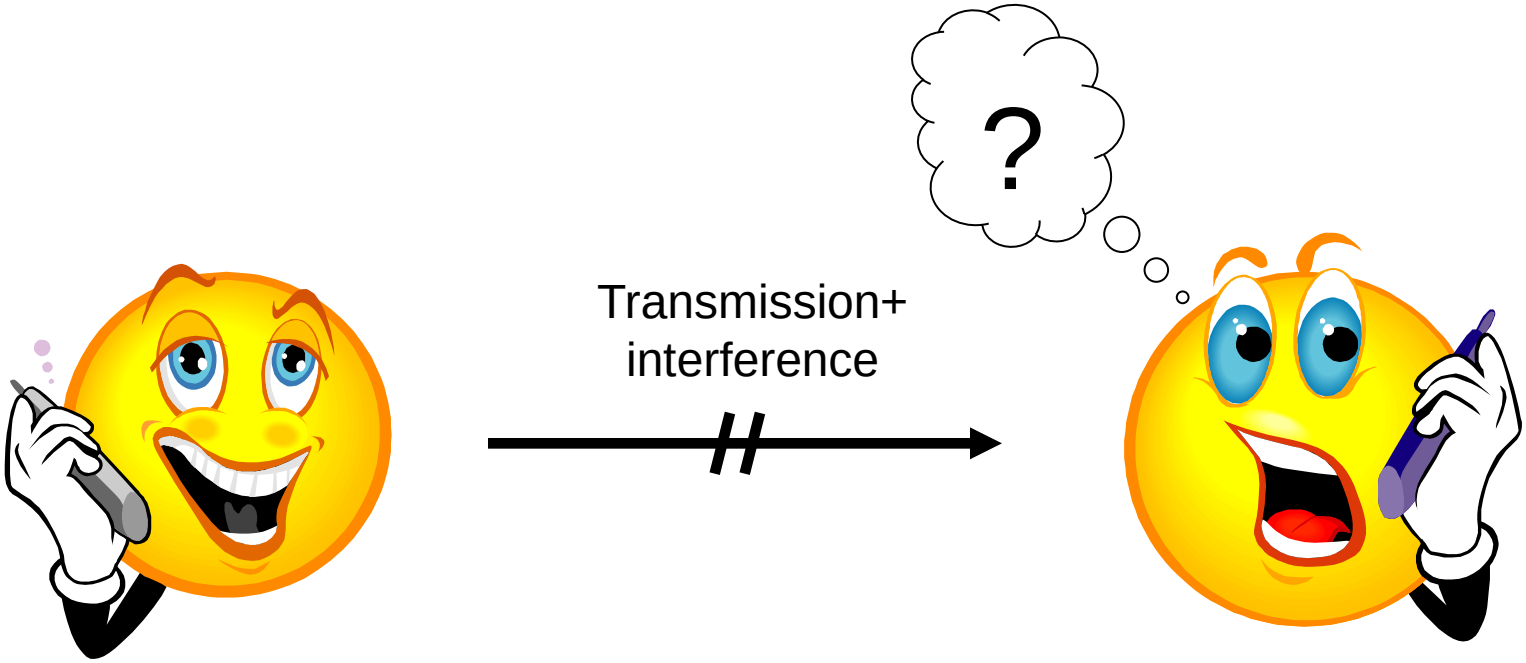
farok	**crrrok**	**hihok**	**yorok**	**clok**	**kantok**	**ok-yurp .**

1a. ok-voon ororok sprok .	1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .	2b. at-drubel at-voon pippat rrat dat .
3a. erok sprok izok hihok ghirok .	3b. totat dat arrat vat hilat .
4a. ok-voon anok drok brok jok .	4b. at-voon krat pippat sat lat .
5a. wiwok **farok** izok stok .	5b. totat **jjat** quat cat .
6a. lalok sprok izok jok stok .	6b. wat dat krat quat cat .
7a. lalok **farok** ororok lalok sprok izok enemok .	7b. wat **jjat** bichat wat dat vat eneat .
8a. lalok brok anok plok nok .	8b. wat lat pippat rrat nnat .
9a. wiwok nok izok kantok ok-yurp .	9b. totat nnat quat oloat at-yurp .
10a. lalok mok nok yorok ghirok clok .	10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .	11b. wat nnat arrat mat zanzanat .
12a. lalok rarok nok izok hihok mok .	12b. wat nnat forat arrat vat gat .

The Noisy Channel Model

- Originally developed in information theory for telecommunications
- Basic problem:
  - What do you do if your transmission channel (e.g. a radio or phone) has interference?

The Noisy Channel Model

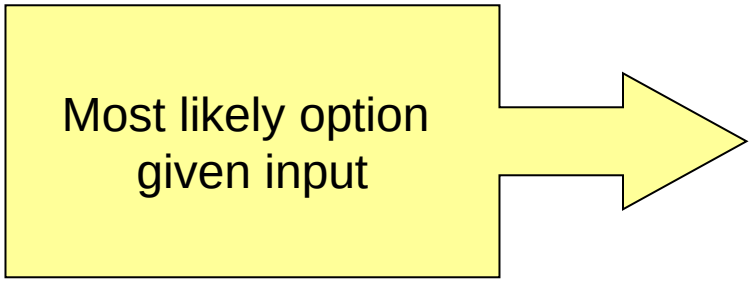


- I'm running a little late
- I-… ru-… … li-…l late

What did he mean?

Input

- I-… ru-… … li-…l late



Possible outputs

- I’ve run into Lyle late
- Isle runners alive all late
- I’m running a little late
- It’s me running …

What is most likely?

- Generally in English:

  P(It's me) >

  P(I'm running) >

  …

  P(Acapulco trips germinate)

- And given an input I-… ru-… li-…l late:

  P(I'm running a little late) >

  P(I ran a little late) >

  …

Suppose German is just English with interference…

To recreate a perturbed English message we need:

- Probability of any sequence in the TL – the **Language Model**
- Probability of each translation from SL -> TL – The **Translation Model**