6.863J Natural Language Processing
Lecture 17: Machine translation I

Robert C. Berwick
berwick@ai.mit.edu

The Menu Bar

- Administrivia:
  - Start w/ final projects, unless there are objections
- *Agenda:*
  - Machine Translation (MT) as a ‘litmus test’ or ‘sandbox’ (graveyard?) for putting together all of NLP
  - Practical systems: Phraselator; Systran (Babelfish); Logos,...

Submenu bar

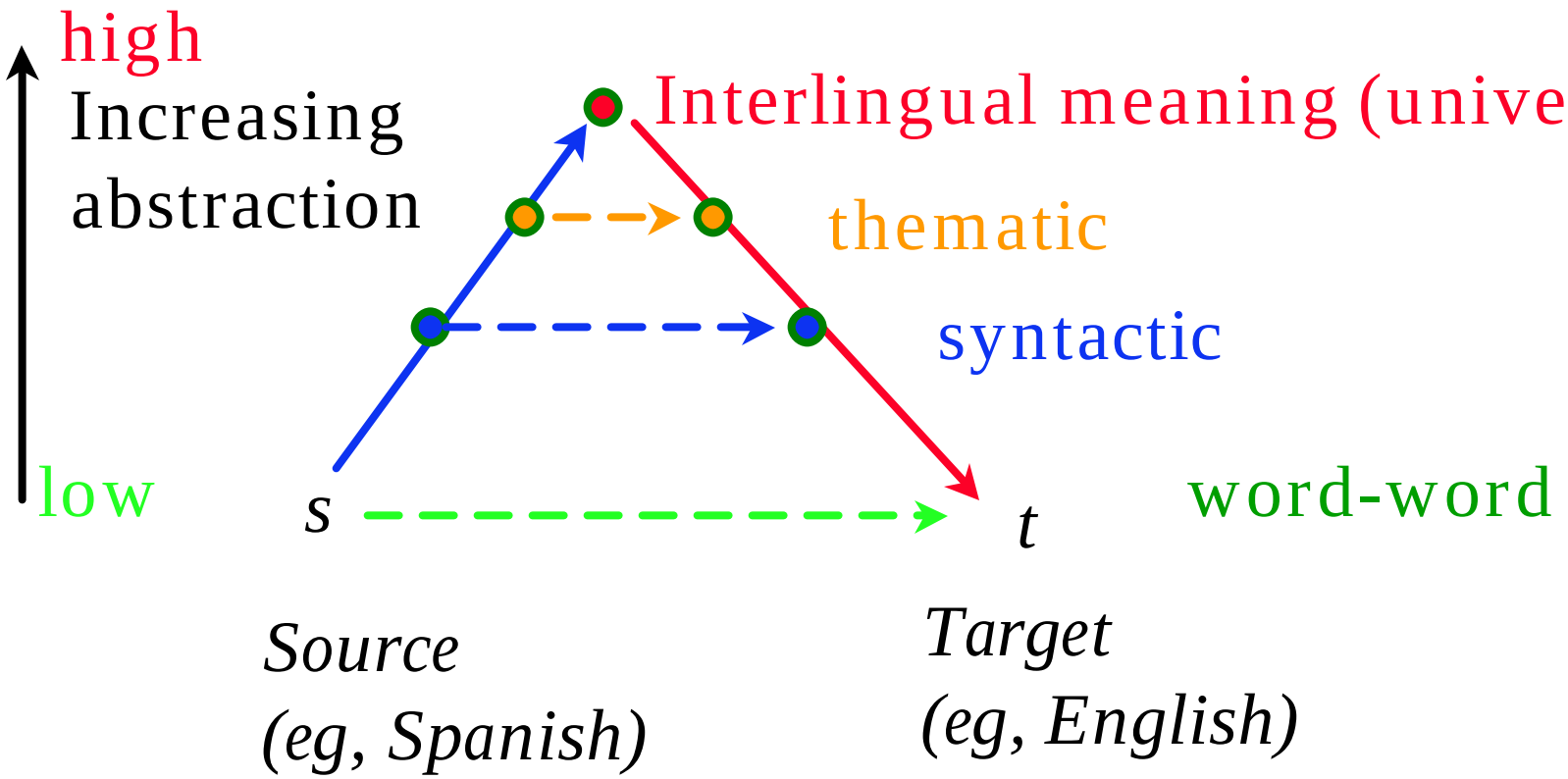
- What is MT?
- Why MT as litmus test?
- A brief history of time
- Getting in the sandbox (nitty gritty)
- The current methods: the great triangle
  - Word-word
  - Transfer
  - Interlingual
  - (Statistical methods used in all)

Why study this?

- Contains *all* parts of NLP
- Famously hard: more or less a Turing test – have computer fool you that there's a human translator behind the curtain
- Current applications & trends
  - Web pages
  - High quality semantics-based in restricted domains – weather reports; equipment manuals
  - Software assistants for MT
  - Automatic knowledge acquisition for improving MT

The golden (Bermuda?) triangle

The golden (Bermuda?) triangle



Then too

- We all have our favorite Monty Python episodes...

The Full Monty

- “My hovercraft...

  is full of eels”
- Hungarian: “Can you direct me to the railway station?”
- [...censored...]
- Mi aerodeslizador es lleno de anguilas
- Where is the men’s room?
- ¿Dónde está el cuarto de los hombres?

A few more idioms...

- Out of sight, out of mind
- ? ? ? ? ? ? ? ? ,
- From vision to heart
- Famous MT – on mag tape – to Russian:

  ? ? ??????????, ?? ?????

  From the sighting, from the reason

What is MT?

- Use of computer
- Translate text (speech) from source to target language (semi)automatically

- Can have humans in the loop
- Holy Grail: FAHQT

Why MT?

- EU uses > 2000 translators for 11 languages
- What % of web is other than English?
- 10% done w/ Systran

- Professional translator gets 15-20 cents/word (Chinese 3x as much)

MT

- Given a sentence *s* in the source language *S*, return a sentence *t* in the target language *T* that conveys the same meaning as *s*
- 'conveys the same meaning' is left unspecified!

A brief history of time – the dawn age

- 1946/47: First discussions on the feasibility of Machine Translation (Warren Weaver and Andrew Booth – after Rockefeller Fdn turned down computer analysis of protein structure...)
- 1949: Weaver's memorandum (considered to be the single act which initiated MT R&D)
- 1950-52: MT studies at MIT (Weiner), Univ. of Washington, UCLA, National Bureau of Standards (NBS), and RAND Corporation.
- 1951: Yehosha Bar-Hillel becomes first full-time MT research person; his appointment was at MIT

The dawn age: the codebreakers

- 1952: First MT Conference, MIT
- 1952: Creation of the Georgetown University research team under Léon Dostert
- 1954: Georgetown-IBM experiment, IBM Technical Computing Bureau, NY; English-Russian MT (eventually: Systran)
- 1954: First English MT research team, Cambridge University
- 1954: First issue of Mechanical Translation
- 1955: First known Soviet MT research

And then came..

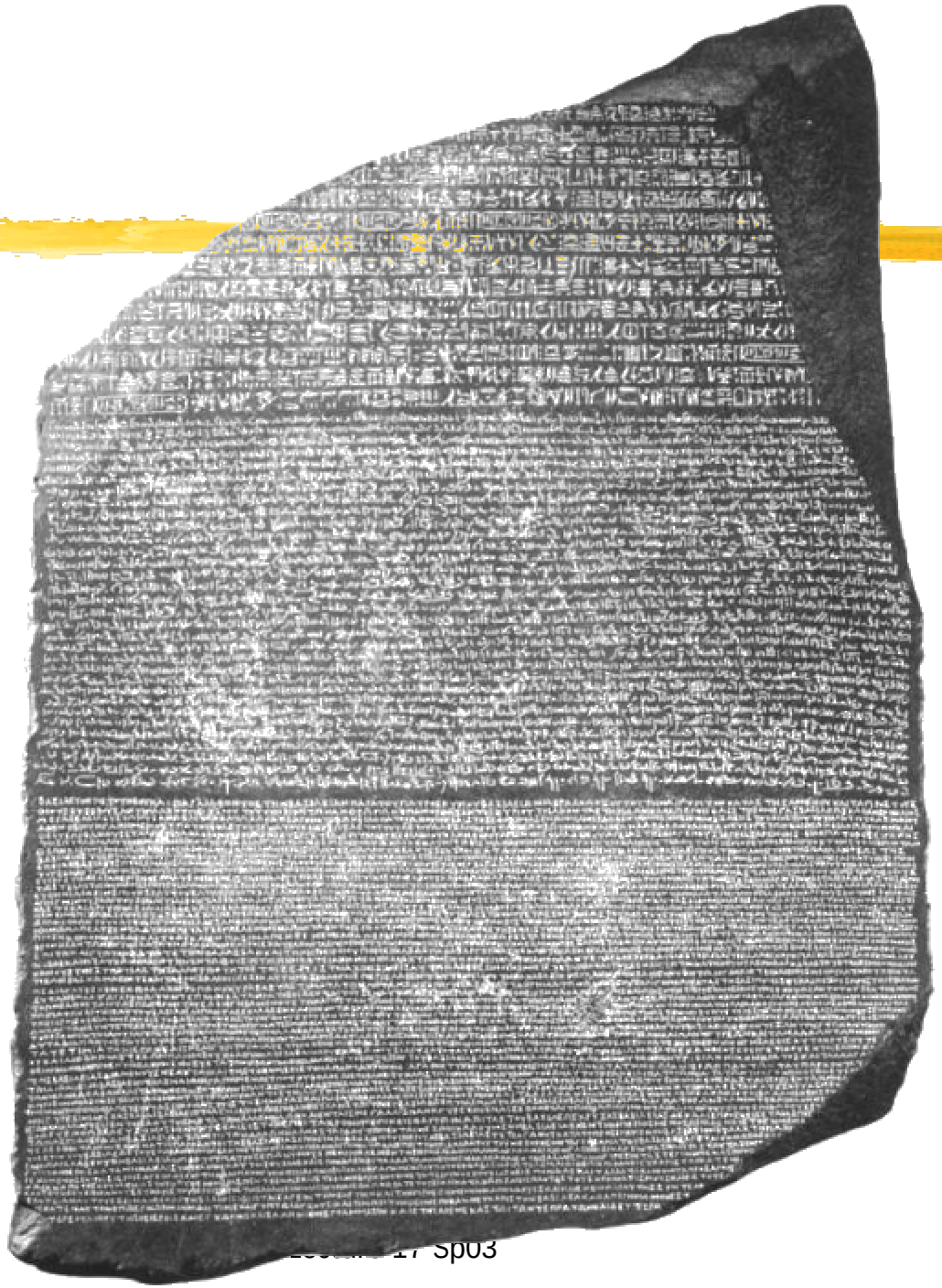
- 1956: First international conference on MT
- 1959: Bar-Hillel's Report on the state of machine translation in the United States and Great Britain: "pig in the pen" example
- 1956-1966: Continued US efforts in MT including: University of Washington, IBM's Watson Research Center; University of Texas; Georgetown University; RAND Corporation; University of Michigan; MIT; National Bureau of Standards, Harvard University ...

The Dark ages..(?)

- 1964: the Automatic Language Processing Advisory Committee (ALPAC) formed by the National Academy of Sciences to study the feasibility of machine translation
- 1966: the ALPAC published its Language and machines: computers in translation and linguistics report, known simply as The ALPAC Report
- The ALPAC Report essentially quashed MT research in the US and other parts of the world until the early 1980's with some exceptions
- Why?

Let’s see why…

- Approach it like a cryptographic problem
- Word-for-word cipher
- Here’s a sample from alien languages (courtesy K. Knight)



Alien languages: Alpha-centauri & Betelgeuse

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. iat 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 crrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .

We will build two things

- Assume word-word translation – though not same word order
- Use alignment of words to build translation dictionary
- Use translation dictionary to improve the alignment – because it eliminates some possibilities

To begin

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. iat 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 .

Translation dictionary:
- ghiork – hilat
- ok-drubel – at-drubel
- ok-voon – at-voon
- ok-yurp – at – yurp
- zananok - zanzanat

OK, what does pairing buy us?

- Sentence 1: 2 possibilities left...
  1. ororok ↔ bichat & sprok ↔ dat
  2. ororok ↔ dat & sprok ↔ bichat

(But also: what if ororok untrans aux v...?)

Which is more likely?

Look for sentence w/ sprok but not ororok

Sentence (2a)

Link throughout corpus (1, 2, 3, 6, 7)

Sentence (2) now looks like a good place to crack...

Sentences 2, 3...

- S2: anok plok/pippat rrat
- S4: **4a. ok-voon anok drok brok jok .**
  4b. at-voon krat pippat sat lat .

Ok, anok ↔ pippat & plok ↔ rrat

S3: So far we have:

erok sprok izok hihok ghirok
totat dat arrat dat hilat

Look at 8; 11; 3 & 12; 5, 6, 9

This suggests

Note:

- Aligning builds the translation dictionary
- Building the translation dictionary aids alignment
- “Decipherment”
- We shall see how this can be automated next time

The dictionary so far...

anok - pippat
erok - total
ghirok - hilat
hihok - arrat
izok – vat/quat
ok-drubel - at-drubel
ok-yurp - at-yurp
ok-voon - at-voon
ororok - bichat
plok - rrat
sprok - dat
zanzanok - zanzanat

If you work through it you'll get all the pairs here, save 1: crrrok

- But you are suddenly abducted to the Federation Translation Center & presented with this sentence from Betelgeuse to translate into Alpha-Centaurian:

- **`iat lat pippat eneat hilat oloat at-yurp .`**

You are given this fragment of Alpha-C text & its bigrams

For actual translation...

- More ambiguous words
- Sentence lengths different
- Sentences longer
- Words translated differently depending on context
- Output word order depends on input order
- Phrasal dictionary: for idioms, etc
- Prounouns; inflections; structural ambiguity

In reality

- 40-50% of English words difft position than French
- For English-Japanese – nearly 100%
- Idioms: ‘got out’, ‘got by’, ‘got even’
- French: sorti, passé, obtenu même

English-French

- The world’s largest living lizard
- Le plus grand lézard vivant du monde

- Book him, Danno
- Le réserver, Danno

And does it scale?

- Is there a large bilingual corpus for (any) pair of natural languages?
- Can we get the bigram data (Yes – see Google)
- Can it be converted to sentence pairs?
- Can we automate decipherment?
- Can we automate translation?
- Are translations good?

(What are alternatives?)

In the words of Babelfish

- If you cannot strike it, connect them

MT: the classical problems

- A challenge: all aspects of NLP

Ch. 18 *The story of stone, 1792, Cao Xue Qin*

"As she lay there alone, Dai-yu's thoughts turned to Bao-chai…Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry." (trans. Hawkes)

Literal

Dai-yu zi zai chuang shang gan nian Bao chai

Dai-yu alone on bed top think-of-w/gratitude

You tinjian chuang wai zhu shao xiang ye zhe

Again listen to window outside bambop tip plaintain leaf of

How is this done???

- Names of servants by meanings
- Verbal tense & aspect rarely marked; so *tou* trans. as *penetrated.*
- Possessive pronoun *her* chosen – better than *the window*
- *Ma (‘curtain’)* as ‘curtains of her bed’
- *Bamboo tip plaintain leaf* – elegant in Chinese, not in English
- This is called *High Quality Full Translation* (HQFT)
- *Not* yet achievable

Rough sublanguage translation

- Eg, on web: various methods use what we shall see is called a *transfer approach*
- Rough enough to give idea of thematic roles
- *Au sortir de la saison 97/98 et surtout au debut de cette saison 98/99,*
- *With leaving season 97/98 and especially at the beginning of this season 98/99...*

Challenges

- Capture variation and similarities amongst languages
- Dimensions not so clear
- Morphologically: # morphemes/word:
  - *Isolating* languages (Vietnamese, Cantonese) – 1 word/ 1 morpheme
  - *Polysynthetic languages* (Siberian Yupik), 1 word = a whole sentence
- Another dimension: degree to which morphemes are segmentable
  - *Agglutinative* (Turkish)
  - *Fusion* (Russian) – *om* in *stolom* (table-sg-instr-decl)

Challenges, II

- Syntax: head first/final
  - *To Yukio; Yukio ne*
- Head marking vs. dependent marking
- English vs. Hungarian:
  - *The man's(*affix) *house*(head)
  - *Az ember haz*(head)-*a*(affix)
- This is related to lexical-semantic analysis: manner of motion marked by verb or on satellite particles like PPs, adverb phrases
- Example:
  - *The bottle floated out*
  - *La botella salio flotando* (direction marked on verb)

Challenges III

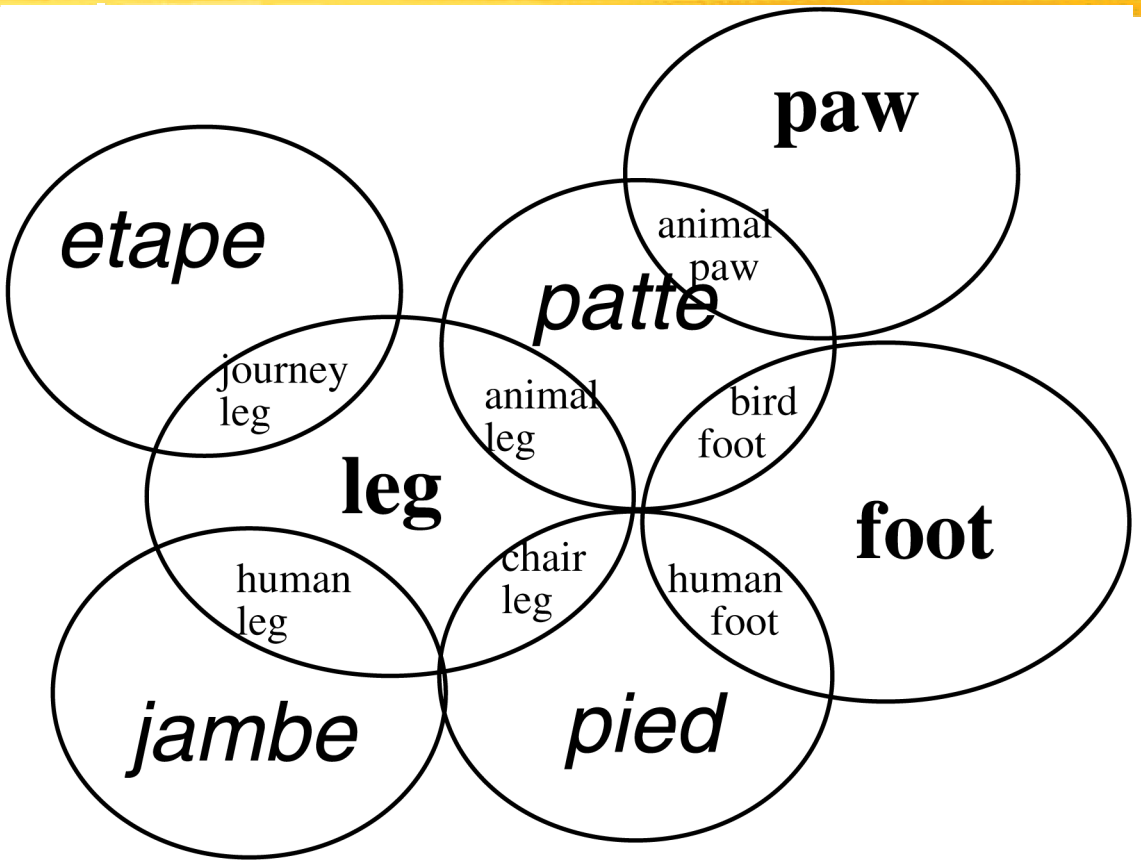
English	*brother*	Japanese	*otooto* (younger)
		Japanese	*oniisan* (older)
		Mandarin	*gege* (older)
		Mandarin	*didi* (older)
English	*wall*	German	*Wand* (inside)
		German	*Mauer* (outside)
English	*know*	French	*connaître* (be acquainted with)
		French	*savoir* (know a proposition)
English	*they*	French	*ils* (masculine)
		French	*elles* (feminine)
German	*berg*	English	*hill*
		English	*mountain*
Mandarin	*tā*	English	*he*, *she*, or *it*

Differences in specificity

Challenges III

- Summarize as *divergences:*
  - Morphological, syntactic, thematic, semantic...
  - Try to impedance match

Dividing up conceptual space



Dividing up conceptual space

- Lexical gap: Jp, no word for *privacy;* Eng: no word for *oyakoko* (filial piety)

The areas

1. Language understanding
2. Language generation
3. Mapping between language pairs

Language understanding

- Argued both for and against
- Example: language savants, 25 languages w/ IQs 50-60
- Linguistic problem: nondeterminism and ambiguity – lexical, syntactic, semantic, context
- Examples of each