Learning to Translate

Machine Translation in the 1950’s

“We’ll have this up and running in a few years, it’ll be great, give us lots of $$$”

Oops! Foundered on word-sense disambiguation.

Nearly sank funding for all of AI.

Currently available technology

(L&H translator, via Japanese)
At the beginning a god created Hajime for the sky and the earth. The earth is frozen as missing, formlessly, darkness was frozen as ceasing, and superficially, deeply, then a divine mind moved on a surface of water.

(Babelfish translator, via Japanese)
God drew up the heaven and the earth with beginning. The earth the formless and was invalid, as for the darkness there was a surface being deep, mind of God was moving to the surface of the water.

The Rosetta Stone (196 BC)
found 1799; hieroglyphs decoded in 1822 by Champollion

Egyptian: hieroglyphs (used from 3300 BC – 400 AD)

Egyptian: Demotic (a late cursive script)

Greek (the language of Ptolemy V, ruler of Egypt)

6 feet tall
The online Bible as Rosetta Stone

English: In the beginning God created the heavens and the earth.
Spanish: En el principio crió Dios los cielos y la tierra.
French: Au commencement Dieu créa les cieux et la terre.
Haitian: Nan konmansman, Bondye kreye syël laak latēa.
Danish: Begyndelsen skabte Gud Himmelen og Jorden.
Swedish: I begynnelsen skapade Gud himmel och jord.
Finnish: Alussa loi Jumala taivaan ja maan.
Greek: ἐν αρχῇ ἐσχῆσαι τὸν θεὸν καὶ τὴν γῆν.
Latin: in principio creavit Deus caelum et terram
Vietnamese: Ban đầu Đức Chúa Trời dung nền trời đất.
Where’s “heaven” in Vietnamese?

English: In the beginning God created the heavens and the earth.
Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.

English: God called the expanse heaven.
Vietnamese: Đức Chúa Trời dat tên khoang không la trời.

English: … you are this day like the stars of heaven in number.
Vietnamese: … các người dông nhu sao trên trời.

“Created” in Vietnamese?

English: In the beginning God created the heavens and the earth.
Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.

English: God created the great sea monsters …
Vietnamese: Đức Chúa Trời dung nên các loại cá lón …

English: God created man in His own image …
Vietnamese: Đức Chúa Trời dung nên loại người nhu hình Ngài …

“God” has a stronger claim …

English: In the beginning God created the heavens and the earth.
Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.

English: God created the great sea monsters …
Vietnamese: Đức Chúa Trời dung nên các loại cá lón …

English: God created man in His own image …
Vietnamese: Đức Chúa Trời dung nên loại người nhu hình Ngài …

… “created” makes do with rest

English: In the beginning God created the heavens and the earth.
Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.

English: God created the great sea monsters …
Vietnamese: Đức Chúa Trời dung nên các loại cá lón …

English: God created man in His own image …
Vietnamese: Đức Chúa Trời dung nên loại người nhu hình Ngài …
What’s “bathroom” in Vietnamese?

- Bible only gives you “begat,” not “bathroom” – but web is much bigger
- Find bilingual web pages automatically
  - “Click for English / Français”
  - Government, tourist, commercial, tech ...
- Run this strategy on them automatically
- Get a dictionary
- Uses: multilingual search, translation aid ...

Competitive Linking Algorithm

1. Link words that look alike or often go together.
2. Make a tentative French-English dictionary of linked words.
   (or if such a dictionary exists already, maybe you can convince the publisher to give you the typesetting files – will work better)

Noisy Channel Model:
Chinese as Garbled English

Given input C, software chooses E that maximizes $p(\text{English}=E) \times p(\text{Chinese}=C \mid \text{English}=E)$

Latin as Garbled English

E = Topmost with praise?
  - high $p(L\mid E)$ but low $p(E)$
  - E = Burger with fries?
    - high $p(E)$ but low $p(L\mid E)$

L = With highest honors
  - maximizes $p(E) \times p(L\mid E)$

Translingual Knowledge Projection and Statistical Machine Translation

National laws applying in Hong Kong

The urgent response to...

Statistical
Model

24 hours!

New Statistical
Language Software

summa cum laude
**What are the models?**

- **Source model** $p(E)$ could be trigram model
  - Guarantees semi-fluent English
- **Channel model** $p(C|E)$ or $p(L|E)$ could be finite-state transducer
  - Stochastically translates each word + allows a little random rearrangement - with high prob, words stay more or less put
  - Maximizing $p(C|E)$ would give really lousy Chinese translation of English
    - Random word translation is stupid - need word *sense* from context
    - Random word rearrangement is stupid - phrases rearrange!
  - This channel has no idea what fluent Chinese looks like
- But maximizing $p(E)^*p(C|E)$ gives a better English translation of Chinese because $p(E)$ knows what English should look like.

- **Currently trying to make these models less stupid.**