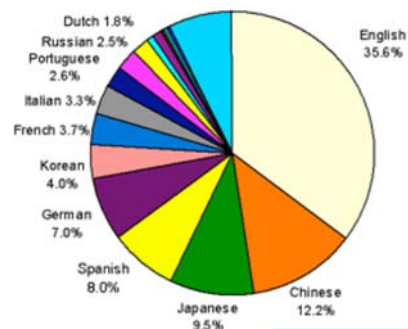




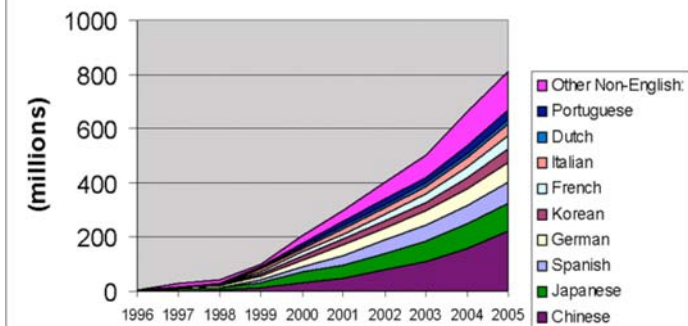
Learning to Translate



Online Language Populations (native speakers)
Total: 680 Million
(Sept., 2003)



Evolution of non-English-speaking online population
(number of people online in each "language zone", I think)



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

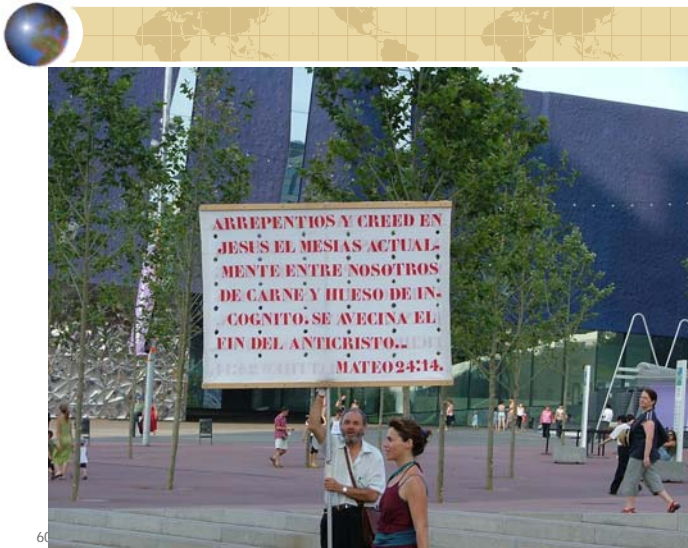


Egyptian:
hieroglyphs
(used from 3300
BC – 400 AD)

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 syèl laak latèa.
 Danish: Begyndelsen skabte Gud Himmelen og Jorden.
 Swedish: I begynnelsen skapade Gud himmel och jord.
 Finnish: Alussa loi Jumala taivaan ja maan.
 Greek: *Ἐν ἀρχῇ ἐποίησεν ὁ Θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.*
 Latin: in principio creavit Deus caelum et terram
 Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.



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 syèl laak latèa.
 Danish: Begyndelsen skabte Gud Himmelen og Jorden.
 Swedish: I begynnelsen skapade Gud himmel och jord.
 Finnish: **Alussa loi Jumala taivaan ja maan.**
 Greek: *Ἐν ἀρχῇ ἐποίησεν ὁ Θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.*
 Latin: in principio creavit Deus caelum et terram
 Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.



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 syèl laak latèa.
 Danish: **Begyndelsen skabte Gud Himmelen og Jorden.**
 Swedish: I begynnelsen skapade Gud himmel och jord.
 Finnish: Alussa loi Jumala taivaan ja maan.
 Greek: *Ἐν ἀρχῇ ἐποίησεν ὁ Θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.*
 Latin: in principio creavit Deus caelum et terram
 Vietnamese: Ban đầu Đức Chúa Trời dung nên trời đất.



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 syèl laak latèa.
 Danish: Begyndelsen skabte Gud Himmelen og Jorden.
 Swedish: I begynnelsen skapade Gud himmel och jord.
 Finnish: Alussa loi Jumala taivaan ja maan.
 Greek: *Ἐν ἀρχῇ ἐποίησεν ὁ Θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.*
 Latin: in principio creavit Deus caelum et terram
 Vietnamese: **Ban đầu Đức Chúa Trời dung nên 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 đặt tên khoảng không là trời.

English: ... you are this day like the stars of **heaven** in number.
Vietnamese: ... các ngôi đồng như sao trên trời.



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 đặt tên khoảng không là **trời**.

English: ... you are this day like the stars of **heaven** in number.
Vietnamese: ... các ngôi đồng như 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 loài cá lớn ...

English: God **created** man in His own image ...
Vietnamese: Đức Chúa Trời dung nên loài người như hình Ngài ...



"Created" in Vietnamese? Uh-oh

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 loài cá lớn ...

English: God **created** man in His own image ...
Vietnamese: **Đức Chúa Trời dung nên** loài người như 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 loài cá lớn ...

English: **God created** man in His own image ...
Vietnamese: **Đức Chúa Trời dung nên** loài người như 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 loài cá lớn ...

English: **God created** man in His own image ...
Vietnamese: **Đức Chúa Trời dung nên** loài người như 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

... nod your head ... wag your tail ... head of the class ... swollen head ...
 ... hochez la tête ... hochez la queue ... en tête de la classe ... bouffant d'orgueil ...

Head ≠ hochez ... but often paired
 head = tête ... though not always
 nod = hochez ... though not always

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)



Competitive Linking Algorithm

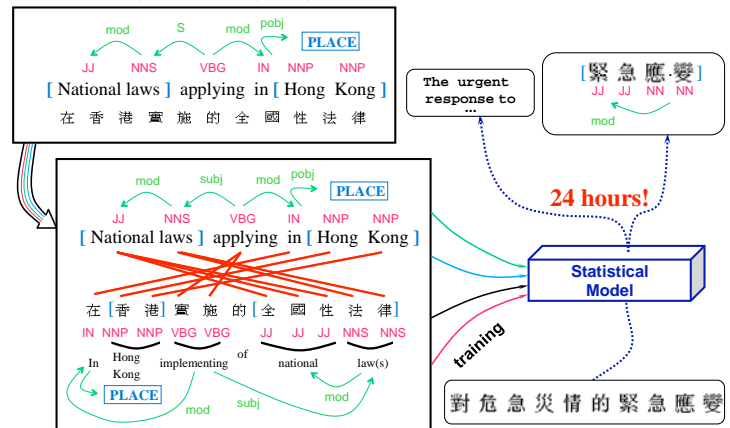
... nod your head ... wag your tail ... head of the class ... swollen head ...
 ... hochez la tête ... hochez la queue ... en tête de la classe ... bouffant d'orgueil ...

Head ≠ hochez ... but often paired
 head = tête ... though not always
 nod = hochez ... though not always

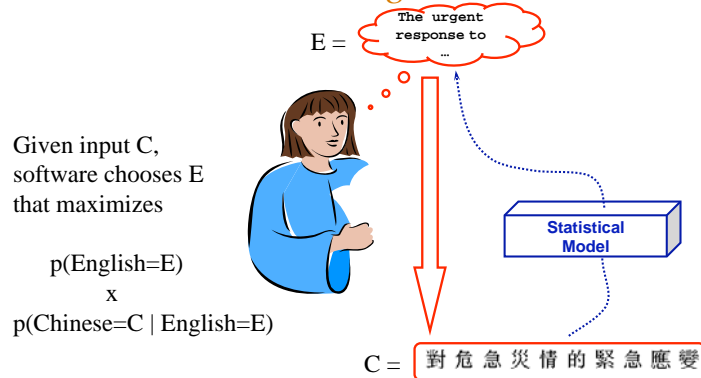
1. Link words that look alike or often go together.
2. Make a tentative French-English dictionary of linked words.
3. Use the dictionary to greedily guess each word's best link.
4. Use the links to get a better dictionary.
5. Repeat!



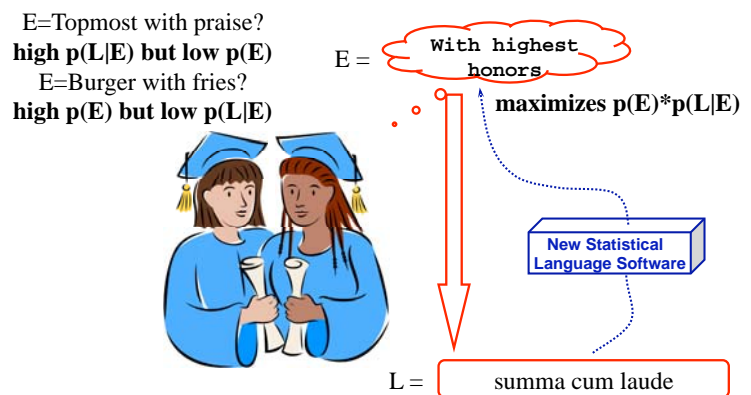
Translingual Knowledge Projection and Statistical Machine Translation



Noisy Channel Model: Chinese as Garbled English



Latin as Garbled English





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.