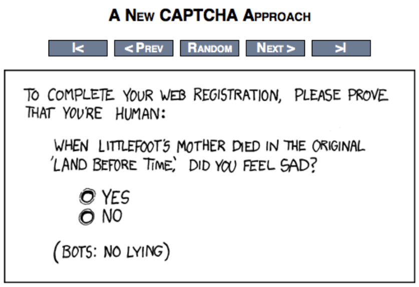
+



A New CAPTCHA Approach

TO COMPLETE YOUR WEB REGISTRATION, PLEASE PROVE THAT YOU'RE HUMAN:

WHEN LITTLEFOOT'S MOTHER DIED IN THE ORIGINAL 'LAND BEFORE TIME,' DID YOU FEEL SAD?

○ YES
○ NO

(BOTS: NO LYING)

+

Natural Language Processing

CS311, Spring 2013
David Kauchak

+

Administrivia

- Status report 1 due tomorrow
- Look at written problems 3 by Monday
  - Recall we do have quizzes for this class
- Exam 2 will be available early next week
- Upcoming schedule
  - Today: NLP
  - Tuesday: Robotics
  - Thursday: Computer Vision
  - Last week: Philosophy and ethics of AI

+

What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia

What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language…

- Dan Jurafsky

Key: Natural text

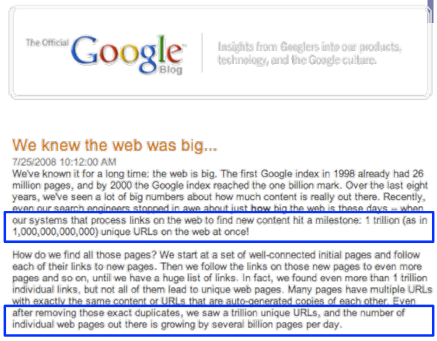
ALL NATURAL

"A growing number of businesses are making Facebook an indispensible part of hanging out their shingles. Small businesses are using …"

Natural text is written by people, generally for people

Why do we even care about natural text in computer science?

Why do we need computers for dealing with natural text?



The Official Google Blog | Insights from Googlers into our products, technology, and the Google culture.

We knew the web was big...

7/25/2008 10:12:00 AM

We've known it for a long time: the web is big. The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, we've seen a lot of big numbers about how much content is really out there. Recently, even our search engineers stopped in awe about just **how** big the web is these days -- when our systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once!

How do we find all those pages? We start at a set of well-connected initial pages and follow each of their links to new pages. Then we follow the links on those new pages to even more pages and so on, until we have a huge list of links. In fact, we found even more than 1 trillion individual links, but not all of them lead to unique web pages. Many pages have multiple URLs with exactly the same content or URLs that are auto-generated copies of each other. Even after removing those exact duplicates, we saw a trillion unique URLs, and the number of individual web pages out there is growing by several billion pages per day.

Why do we need computers?

amazon mechanical turk™
Artificial Artificial Intelligence

~7 billion people in the world

Hire them all:

- Task that takes 10 seconds per page
  - 1 trillion * 10 seconds = 23 minutes
- Task that takes 1 min per page
  - 2 hours
- Task that takes 5 min per page
  - 10 hours

+ Web is just the start…

e-mail

247 billion e-mails a day

27 million tweets a day

corporate
databases

Blogs: **126 million** different blogs

http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/

+ Why is NLP hard?

Iraqi Head Seeks Arms

Juvenile Court to Try Shooting Defendant

Stolen Painting Found by Tree

Kids Make Nutritious Snacks

Local HS Dropouts Cut in Half

Obesity Study Looks for Larger Test Group

British Left Waffles on Falkland Islands

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

+ Why is NLP hard?

User: Where is Contagion playing in the Middlebury Area?

System: Contagion is playing at the Marquis Theatre.

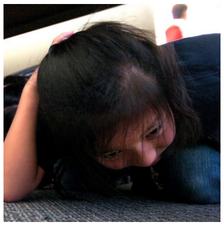
User: When is it playing there?

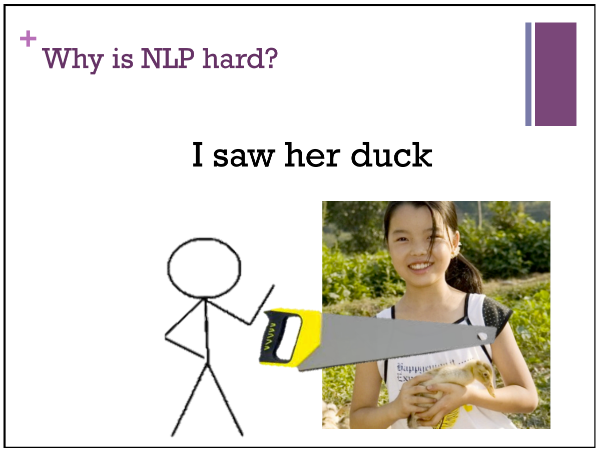
System: It's playing at 2pm, 5pm and 8pm

User: I'd like 1 adult and 2 children for the first show. How much would that cost?

+ Why is NLP hard?

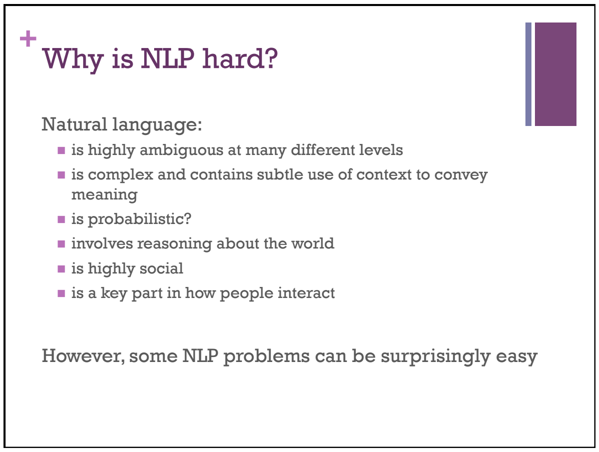
I saw her duck





Why is NLP hard?

I saw her duck

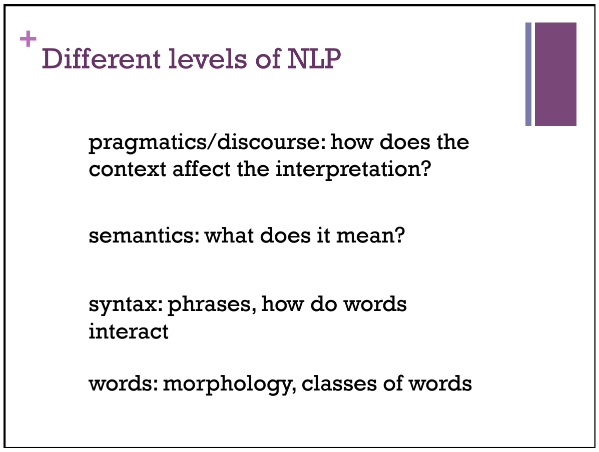


Why is NLP hard?

Natural language:

- is highly ambiguous at many different levels
- is complex and contains subtle use of context to convey meaning
- is probabilistic?
- involves reasoning about the world
- is highly social
- is a key part in how people interact

However, some NLP problems can be surprisingly easy



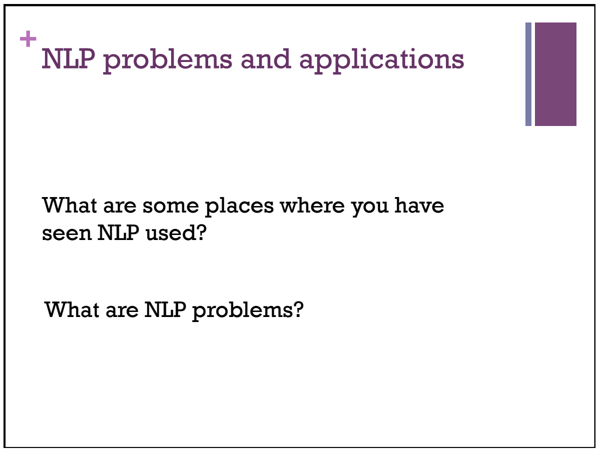
Different levels of NLP

pragmatics/discourse: how does the context affect the interpretation?

semantics: what does it mean?

syntax: phrases, how do words interact

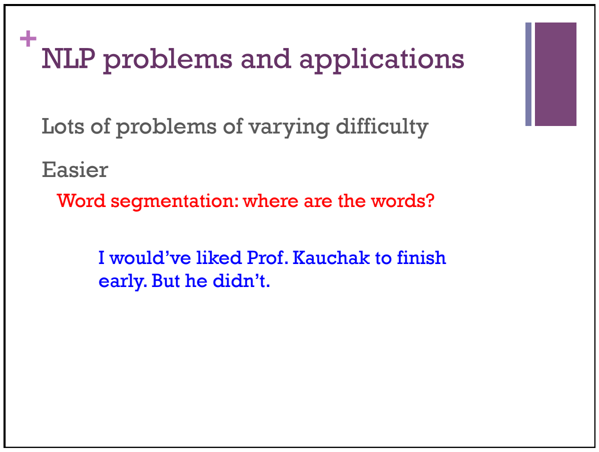
words: morphology, classes of words



NLP problems and applications

What are some places where you have seen NLP used?

What are NLP problems?



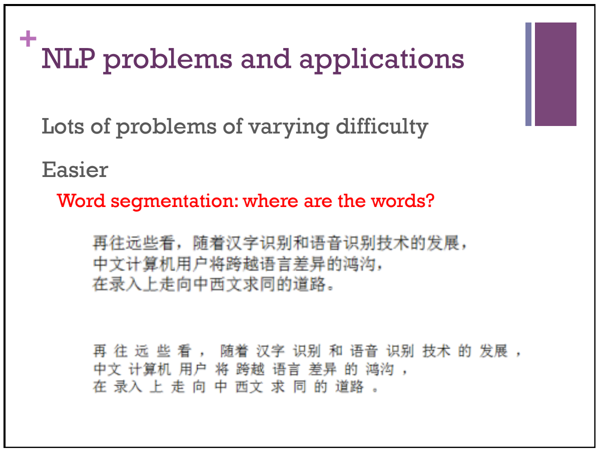
NLP problems and applications

Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

I would've liked Prof. Kauchak to finish early. But he didn't.



NLP problems and applications

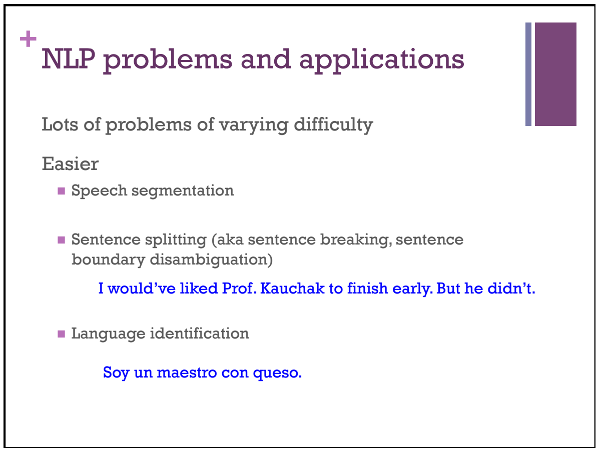
Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

再往远些看，随着汉字识别和语音识别技术的发展，
中文计算机用户将跨越语言差异的鸿沟，
在录入上走向中西文求同的道路。

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。



NLP problems and applications

Lots of problems of varying difficulty

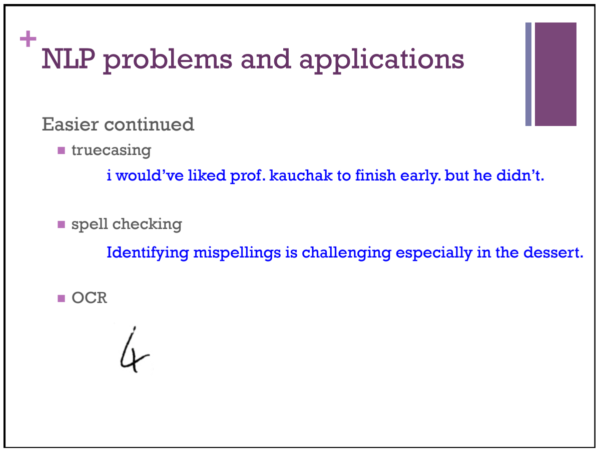
Easier

- Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)

  I would've liked Prof. Kauchak to finish early. But he didn't.

- Language identification

  Soy un maestro con queso.



NLP problems and applications

Easier continued

- truecasing

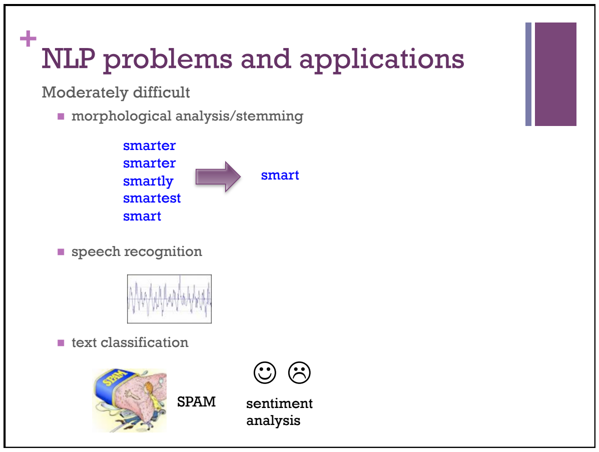
  i would've liked prof. kauchak to finish early. but he didn't.

- spell checking

  Identifying mispellings is challenging especially in the dessert.

- OCR

  4



+ NLP problems and applications

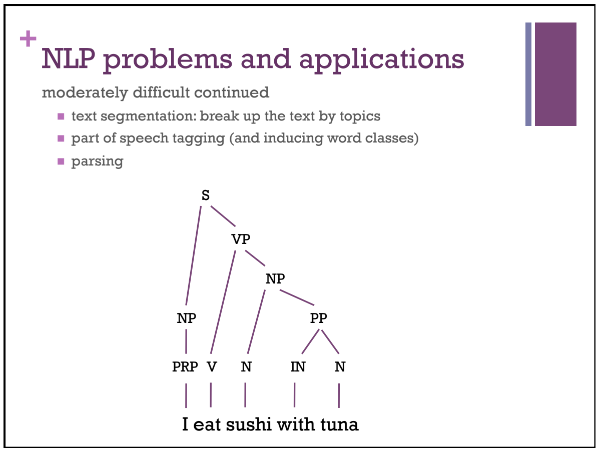
Moderately difficult

- morphological analysis/stemming

smarter
smarter
smartly
smartest
smart
smart

- speech recognition
- text classification

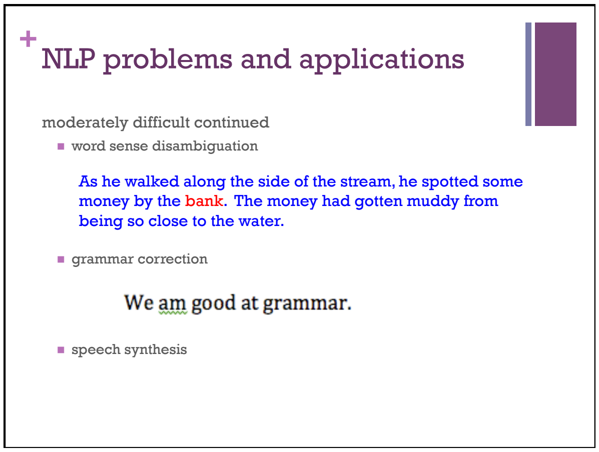
SPAM
☺ ☹
sentiment analysis



+ NLP problems and applications

moderately difficult continued

- text segmentation: break up the text by topics
- part of speech tagging (and inducing word classes)
- parsing



+ NLP problems and applications

moderately difficult continued

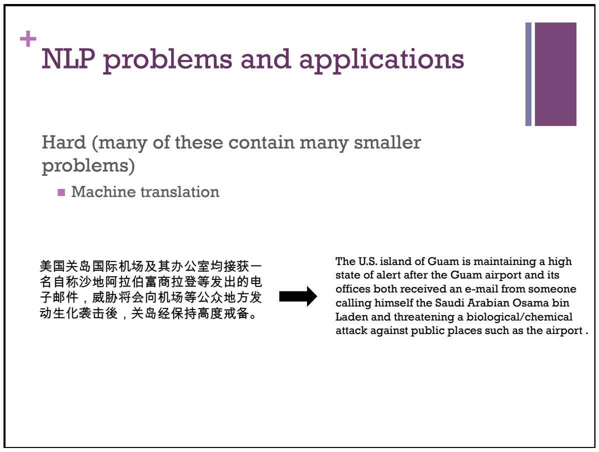
- word sense disambiguation

As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

- grammar correction

We am good at grammar.

- speech synthesis



+ NLP problems and applications

Hard (many of these contain many smaller problems)

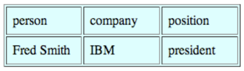
- Machine translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。 → The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

NLP problems and applications

Information extraction

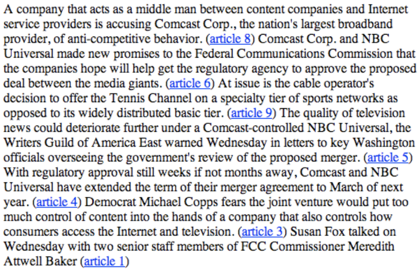
IBM hired Fred Smith as president.



person	company	position
Fred Smith	IBM	president

NLP problems and applications

Summarization



A company that acts as a middle man between content companies and Internet service providers is accusing Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior. (article 8) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 6) At issue is the cable operator's decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 9) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials overseeing the government's review of the proposed merger. (article 5) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the term of their merger agreement to March of next year. (article 4) Democrat Michael Copps fears the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 3) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Attwell Baker (article 1)

NLP problems and applications

Natural language understanding

- Text => semantic representation (e.g. logic, probabilistic relationships)

Information retrieval and question answering

"How many programmers in the child care department make over $50,000?"

"Who was the fourteenth president?"

"How did he die?"

NLP problems and applications

Text simplification

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

↓

Alfonso Perez is a former Spanish football player.

+ Where are we now?

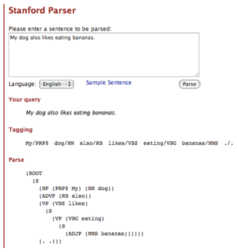
Many of the "easy" and "medium" problems have reasonable solutions

- spell checkers
- sentence splitters
- word segementers/tokenizers

+ Where are we now?

Parsing

- Stanford Parser (http://nlp.stanford.edu:8080/parser/)



Stanford Parser

Please enter a sentence to be parsed:

My dog also likes eating bananas.

Your query

My dog also likes eating bananas.

Tagging

```
My/PRP$  dog/NN  also/RB  likes/VBZ  eating/VBG  bananas/NNS  ./.
```

Parse

```
(ROOT
  (S
    (NP (PRP$ My) (NN dog))
    (ADVP (RB also))
    (VP (VBZ likes)
      (S
        (VP (VBG eating)
          (S
            (ADJP (NNS bananas))))))
    (. .)))
```

+ Where are we now?

Machine translation

- Getting better every year
- enough to get the jist of most content, but still no where near a human translation
- better for some types of text

http://translate.google.com

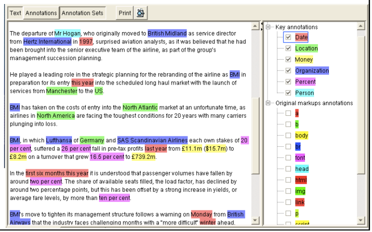
Many commercial versions…

- systran
- language weaver

+ Where are we now?

Information extraction

- Structured documents (very good!)
  - www.dealtime.com
  - www.froogle.com
- AKT technologies
- Lots of these
  - FlipDog
  - WhizBang! Labs
  - …
  - work fairly well



The departure of Mr Hogan, who originally moved to British Midland as service director from Hertz International in 1997, surprised aviation analysts, as it was believed that he had been brought into the senior executive team of the airline, as part of the group's management succession planning.

He played a leading role in the strategic planning for the rebranding of the airline as BMI in preparation for its entry this year into the scheduled long haul market with the launch of services from Manchester to the US.

BMI has taken on the costs of entry into the North Atlantic market at an unfortunate time, as airlines in North America are facing the toughest conditions for 20 years with many carriers plunging into loss.

BMI, in which Lufthansa of Germany and SAS Scandinavian Airlines each own stakes of 20 per cent, suffered a 26 per cent fall in pre-tax profits last year from £11.1m ($15.7m) to £8.2m on a turnover that grew 16.5 per cent to £739.2m.

In the first six months this year it is understood that passenger volumes have fallen by around two per cent. The share of available seats filled, the load factor, has declined by around two percentage points, but this has been offset by a strong increase in yields, or average fare levels, by more than ten per cent.

BMI's move to tighten its management structure follows a warning on Monday from British Airways that the industry faces challenging months with a "more difficult" winter ahead.

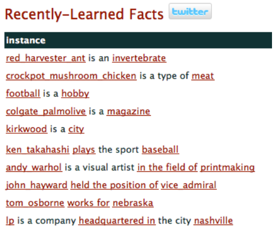
Key annotations
- ☑ Date
- ☑ Location
- ☑ Money
- ☑ Organization
- ☑ Percent
- ☑ Person

Original markups annotations
- ☐ a
- ☐ b
- ☐ body
- ☐ br
- ☐ font
- ☐ head
- ☐ html
- ☐ img
- ☐ link
- ☐ p

Where are we now?

CMU's NELL (Never Ending Language Learner)

http://rtw.ml.cmu.edu/rtw/



Recently-Learned Facts

instance

red_harvester_ant is an invertebrate

crockpot_mushroom_chicken is a type of meat

football is a hobby

colgate_palmolive is a magazine

kirkwood is a city

ken_takahashi plays the sport baseball

andy_warhol is a visual artist in the field of printmaking

john_hayward held the position of vice_admiral

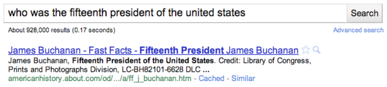
tom_osborne works for nebraska

lp is a company headquartered in the city nashville

Where are we now?

Information retrieval/query answering

- search engines:
  - pretty good for some things



who was the fifteenth president of the united states

About 928,000 results (0.17 seconds)

James Buchanan - Fast Facts - **Fifteenth President** James Buchanan

James Buchanan, **Fifteenth President of the United States**. Credit: Library of Congress, Prints and Photographs Division, LC-BH82101-6628 DLC ...

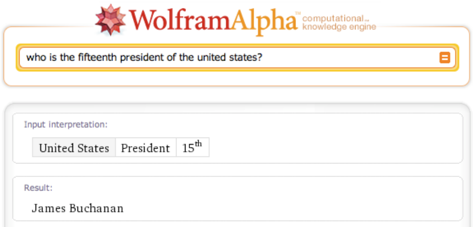
americanhistory.about.com/od/.../a/ff_j_buchanan.htm - Cached - Similar

  - does mostly pattern matching and ranking
    - no deep understanding
    - still requires user to "find" the answer

Where are we now?

Question answering

- wolfram alpha



who is the fifteenth president of the united states?

Input interpretation:

United States	President	15th

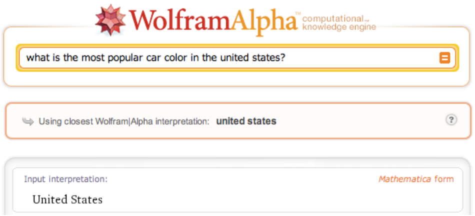
Result:

James Buchanan

Where are we now?

Question answering

- wolfram alpha



what is the most popular car color in the united states?

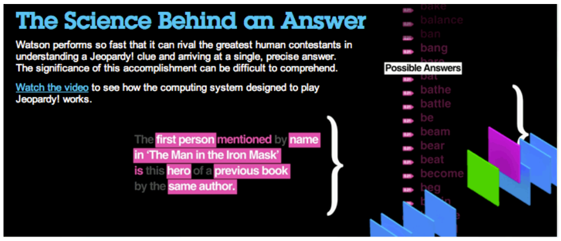
Using closest Wolfram|Alpha interpretation: **united states**

Input interpretation:

United States

+ Question answering

IBM WATSON



The Science Behind an Answer

Watson performs so fast that it can rival the greatest human contestants in understanding a Jeopardy! clue and arriving at a single, precise answer. The significance of this accomplishment can be difficult to comprehend.

Watch the video to see how the computing system designed to play Jeopardy! works.

The first person mentioned by name in 'The Man in the Iron Mask' is this hero of a previous book by the same author.

+ Where are we now?

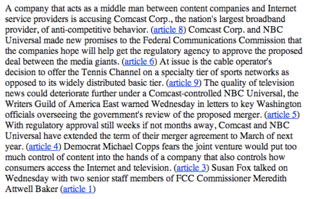
Question answering

- Many others…
  - TREC question answering competition
  - language computer corp
  - answerbus
  - …

+ Where are we now?

Summarization

- NewsBlaster (Columbia)
  - http://newsblaster.cs.columbia.edu/



A company that acts as a middle man between content companies and Internet service providers is accusing Comcast Corp., the nation's largest broadband provider, of anti-competitive behavior. (article 8) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 6) At issue is the cable operator's decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 9) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials overseeing the government's review of the proposed merger. (article 5) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the term of their merger agreement to March of next year. (article 4) Democrat Michael Copps fears the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 3) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Attwell Baker (article 1)

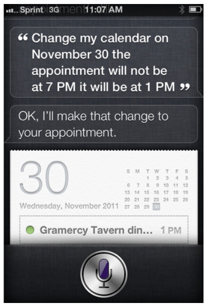
+ Where are we now?

Voice recognition

- pretty good, particularly with speaker training
  - Apple OS has one built in:
    - "What time is it?"
    - "Switch to finder"
    - "Hide this application"
  - IBM ViaVoice
  - Dragon Naturally Speaking

- Speech generation
  - The systems can generate the words, but getting the subtle nuances right is still tricky
    - Apple OS
    - translate.google.com

+ A combination of problems…



“Change my calendar on November 30 the appointment will not be at 7 PM it will be at 1 PM”

OK, I'll make that change to your appointment.

Gramercy Tavern din… 1 PM

+ Other problems

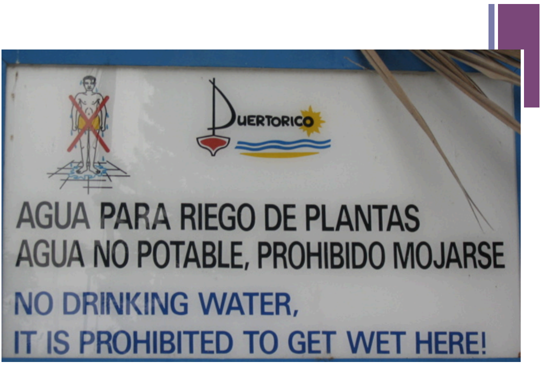
- Many problems untackled/undiscovered
- “That's What She Said: Double Entendre Identification”
  - ACL 2011
  - http://www.cs.washington.edu/homes/brun/pubs/pubs/Kiddon11.pdf

+ Language translation

+ MT Systems

Where have you seen machine translation systems?

+


PUERTORICO
AGUA PARA RIEGO DE PLANTAS
AGUA NO POTABLE, PROHIBIDO MOJARSE
NO DRINKING WATER,
IT IS PROHIBITED TO GET WET HERE!

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

→

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

People around the world stubbornly refuse to write everything in English.

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

Machine translation is becoming very prevalent

Even PowerPoint has translation built into it!

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

United States Guam International Airport and the Office received one claiming to be a wealthy Saudi Arabia an email such as Osama bin Laden, threats to the airport after biochemical attacks in public places such as Guam remain on high alert.

+

Which is the human?

Beijing Youth Daily said that under the Ministry of Agriculture, the beef will be destroyed after tests.

The Beijing Youth Daily pointed out that the seized beef would be disposed of after being examined according to advice from the Ministry of Agriculture.

?

+ Which is the human?

Pakistan President Pervez Musharraf Wins Senate Confidence Vote

Pakistani President Musharraf Won the Trust Vote in Senate and Lower House

?

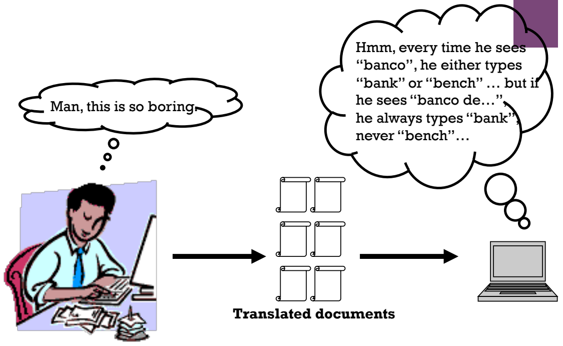
+ Which is the human?

There was not a single vote against him.

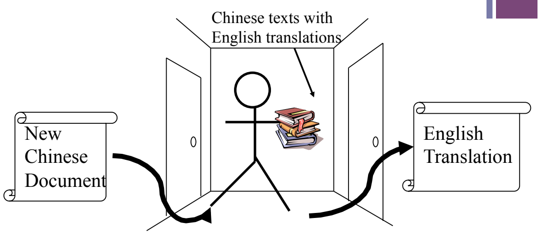
No members vote against him. "

?

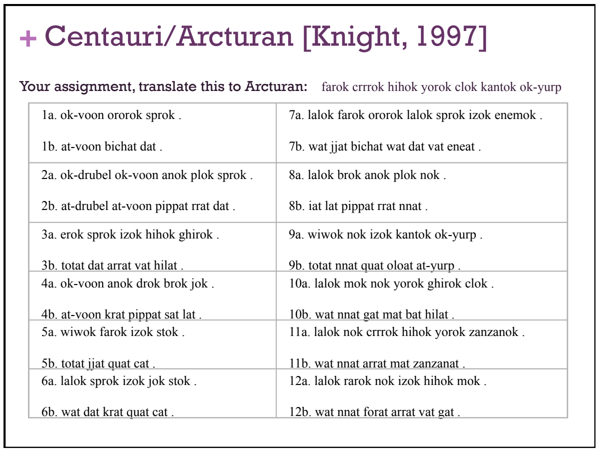
+ Data-Driven Machine Translation


Man, this is so boring.
Hmm, every time he sees "banco", he either types "bank" or "bench" … but if he sees "banco de…", he always types "bank", never "bench"…
Translated documents

+ Welcome to the Chinese Room



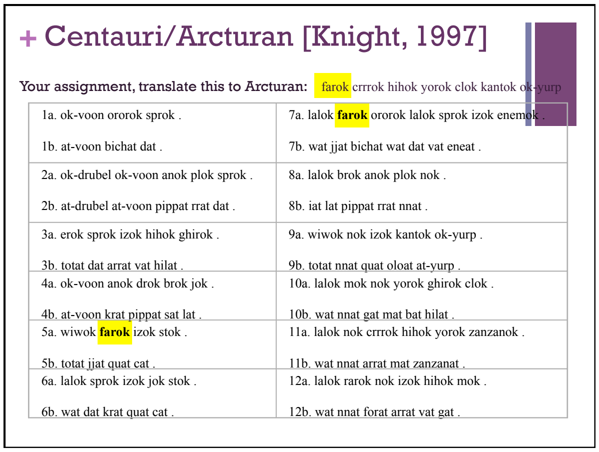
You can teach yourself to translate Chinese using *only* bilingual data (without grammar books, dictionaries, any people to answer your questions…)



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

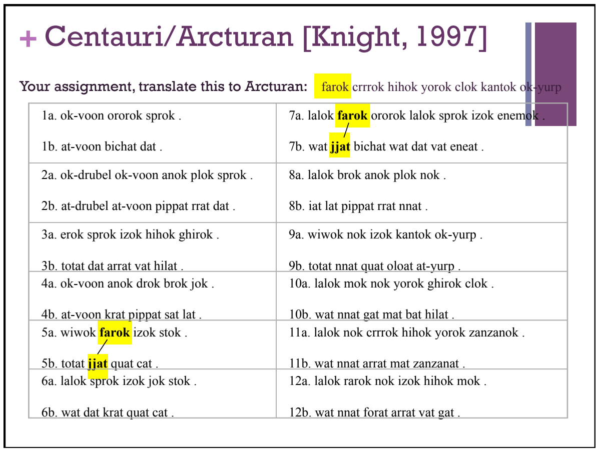
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok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



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

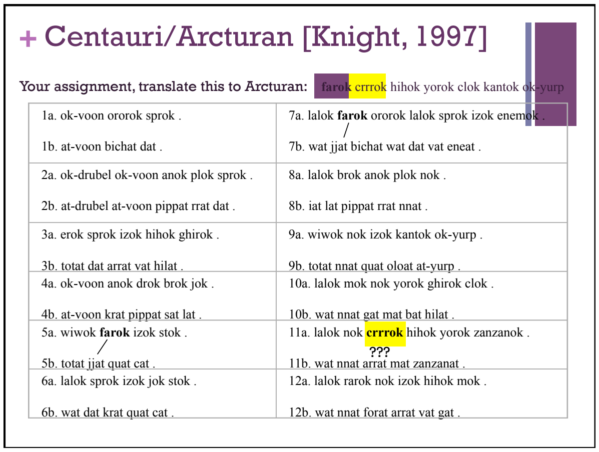
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok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



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

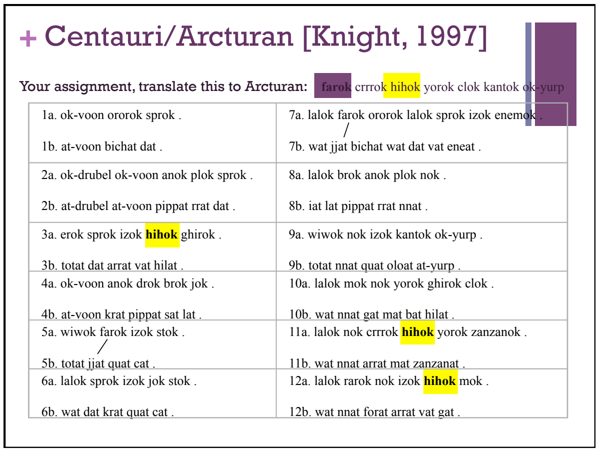
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok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



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

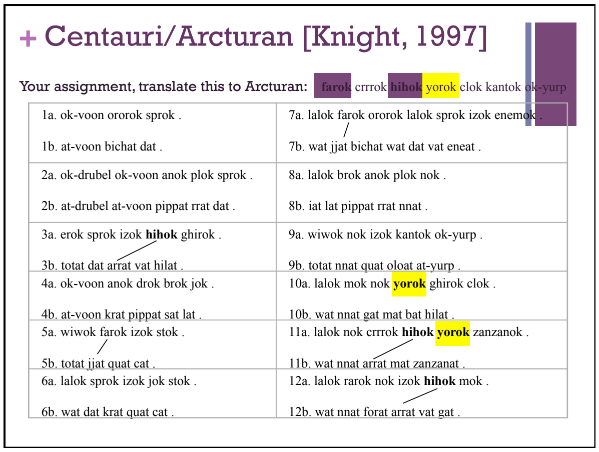
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok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



+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

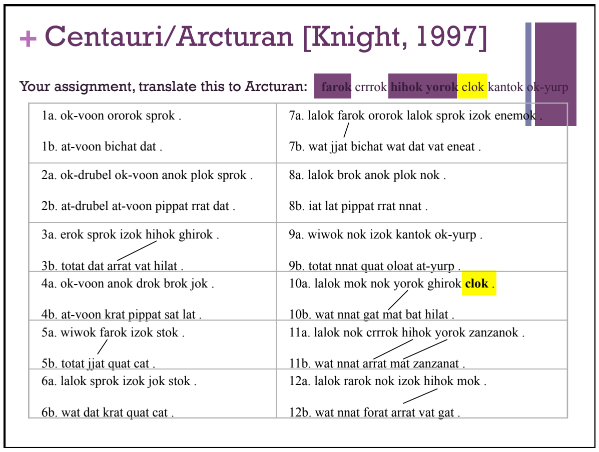
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok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



+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

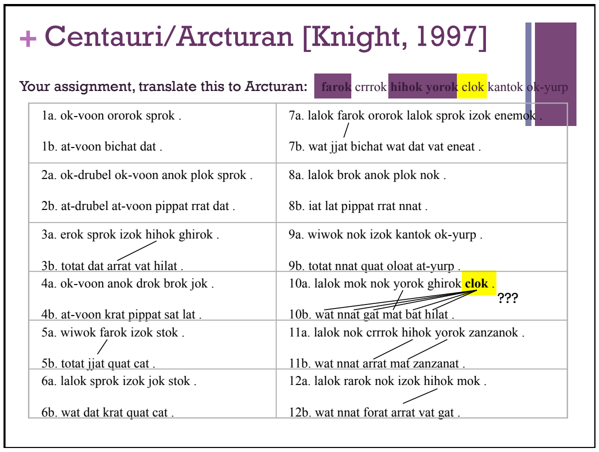
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok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



+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

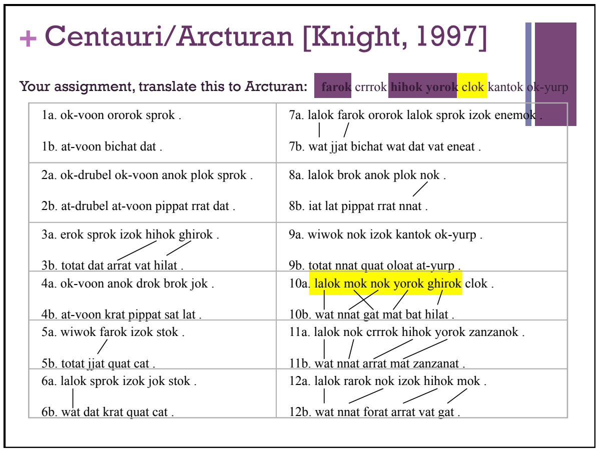
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok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



+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

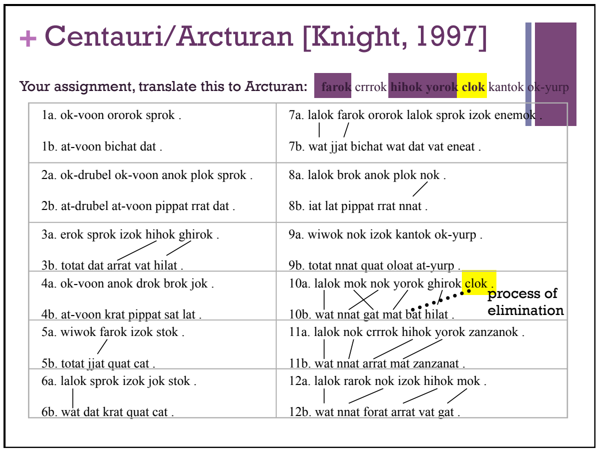
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok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



+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

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

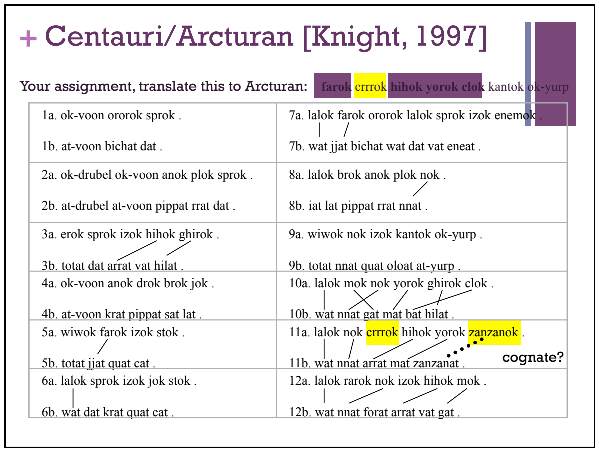


+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

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

process of elimination

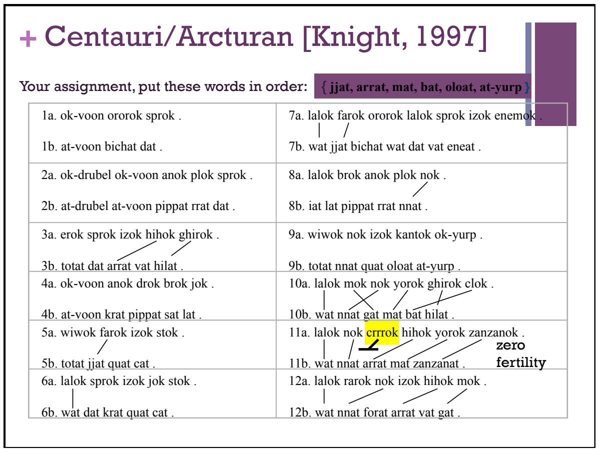


+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

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

cognate?



+ Centauri/Arcturan [Knight, 1997]

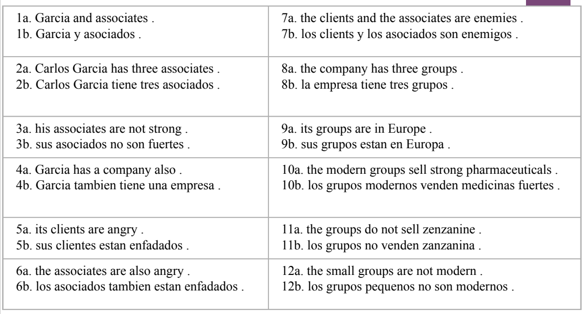
Your assignment, put these words in order: { jjat, arrat, mat, bat, oloat, at-yurp }

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

zero fertility

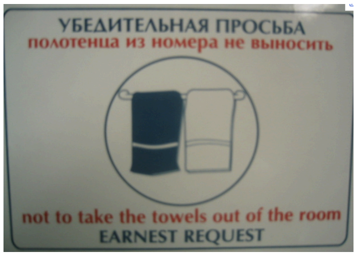
+ It’s Really Spanish/English

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa



1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

+


УБЕДИТЕЛЬНАЯ ПРОСЬБА
полотенца из номера не выносить
not to take the towels out of the room
EARNEST REQUEST

+ Warren Weaver (1947)

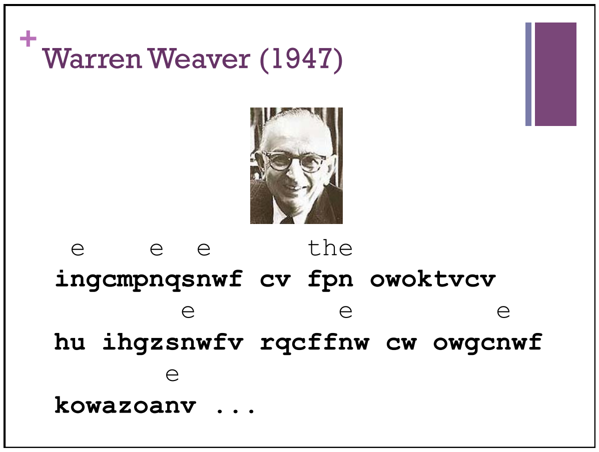
```
ingcmpnqsnwf cv fpn owoktvcv

hu ihgzsnwfv rqcffnw cw owgcnwf

kowazoanv ...
```

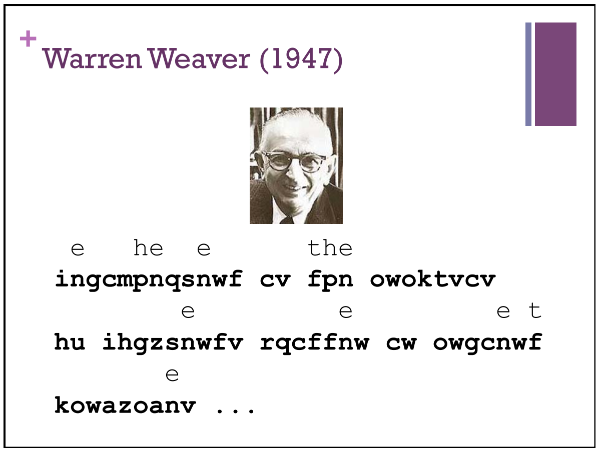
+ Warren Weaver (1947)

```
 e    e  e        e
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```



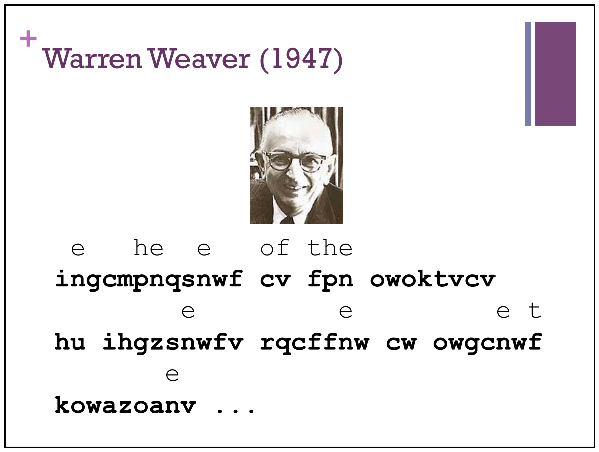
Warren Weaver (1947)

```
 e    e  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```



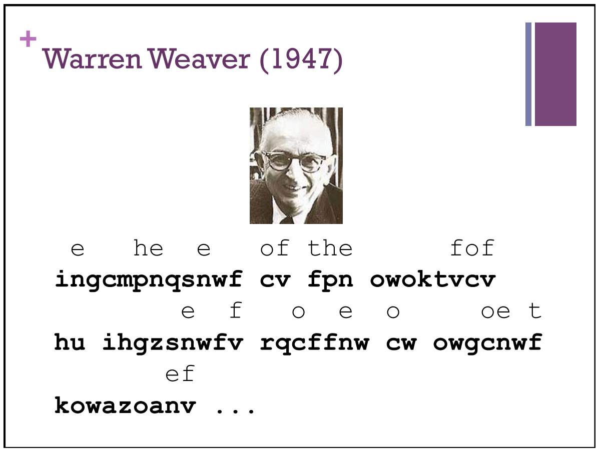
Warren Weaver (1947)

```
 e   he  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```



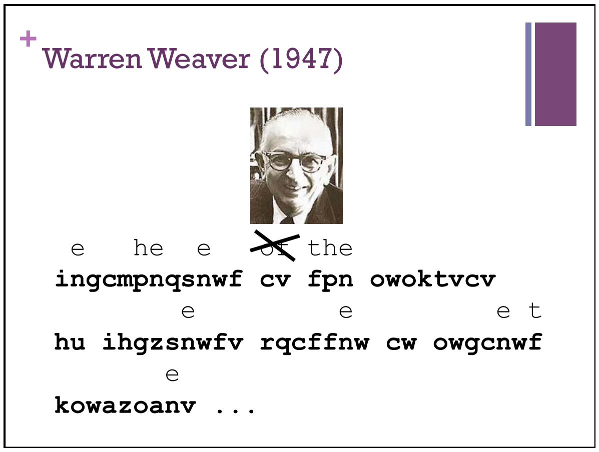
Warren Weaver (1947)

```
 e   he  e   of the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```



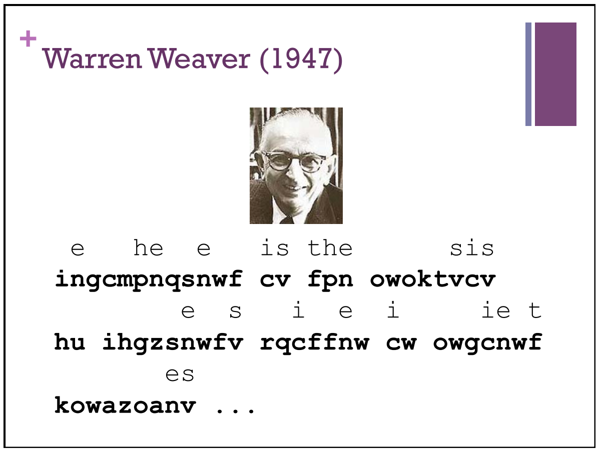
Warren Weaver (1947)

```
 e   he  e   of the      fof
ingcmpnqsnwf cv fpn owoktvcv
        e  f   o  e  o     oe t
hu ihgzsnwfv rqcffnw cw owgcnwf
       ef
kowazoanv ...
```



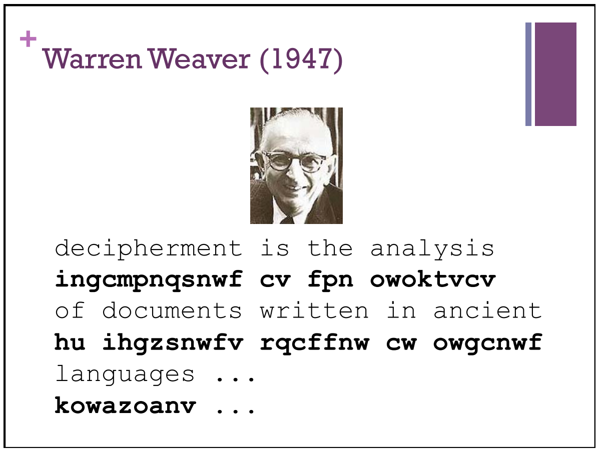
+ Warren Weaver (1947)

```
 e   he  e   ~~of~~ the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```



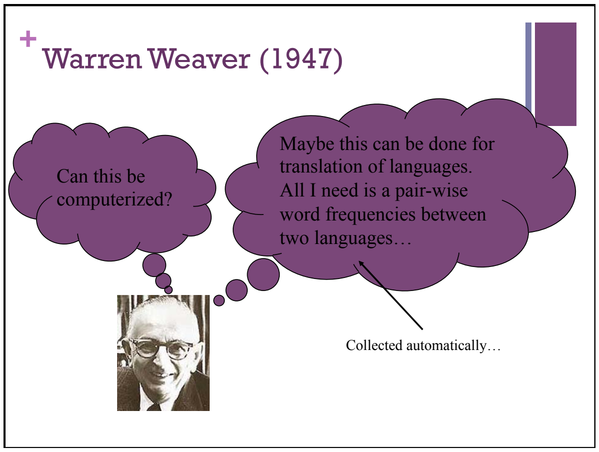
+ Warren Weaver (1947)

```
 e   he  e   is the      sis
ingcmpnqsnwf cv fpn owoktvcv
        e  s   i  e  i     ie t
hu ihgzsnwfv rqcffnw cw owgcnwf
       es
kowazoanv ...
```



+ Warren Weaver (1947)

```
decipherment is the analysis
ingcmpnqsnwf cv fpn owoktvcv
of documents written in ancient
hu ihgzsnwfv rqcffnw cw owgcnwf
languages ...
kowazoanv ...
```

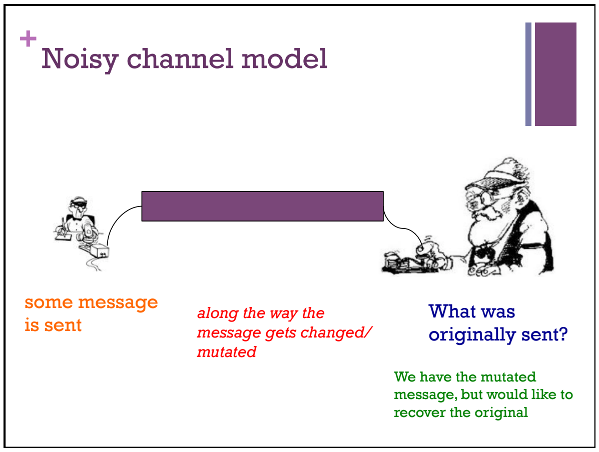


+ Warren Weaver (1947)

Can this be computerized?

Maybe this can be done for translation of languages. All I need is a pair-wise word frequencies between two languages…

Collected automatically…



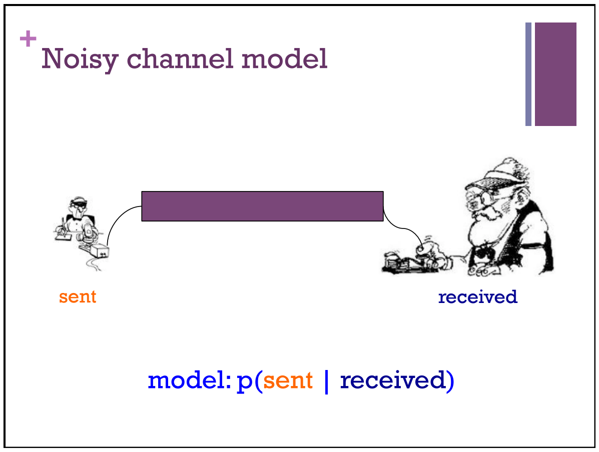
+ Noisy channel model

some message is sent

along the way the message gets changed/ mutated

What was originally sent?

We have the mutated message, but would like to recover the original

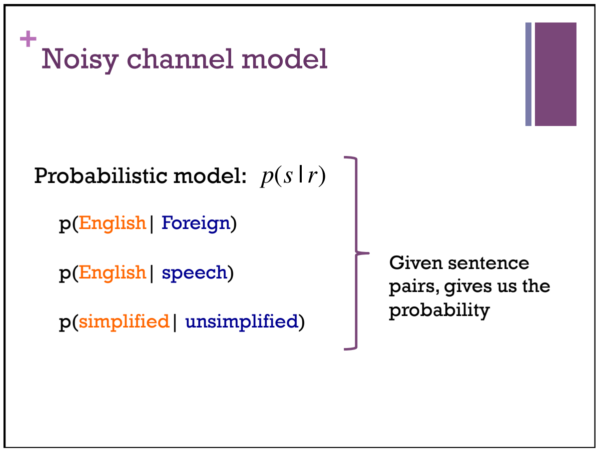


+ Noisy channel model

sent

received

model: p(sent | received)



+ Noisy channel model

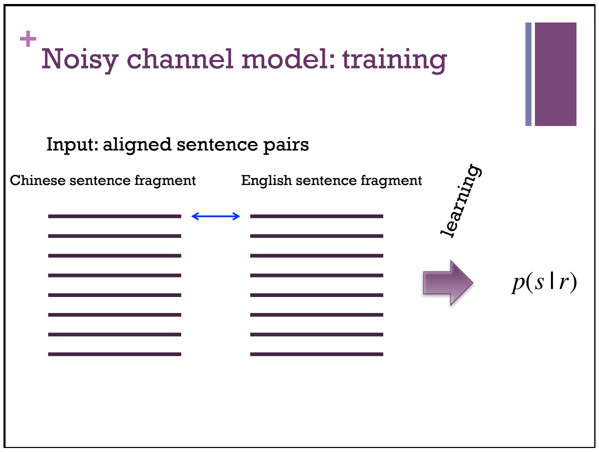
Probabilistic model: $p(s \mid r)$

p(English | Foreign)

p(English | speech)

p(simplified | unsimplified)

Given sentence pairs, gives us the probability



+ Noisy channel model: training

Input: aligned sentence pairs

+ Noisy channel model: applying

Conditioned on the input, what is the most-likely output

$$\underset{s}{\arg\max}\ p(s \mid r)$$

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

+ Noisy channel model

$p(s \mid r) =$

Bayes' rule

+ Noisy channel model

$$p(s \mid r) = \frac{p(r \mid s)\,p(s)}{p(r)}$$

Bayes' rule

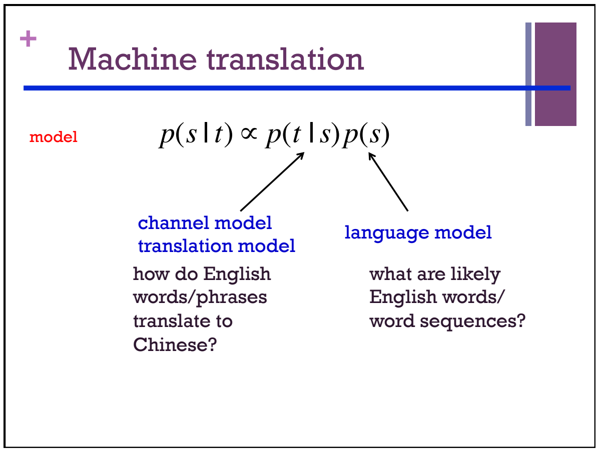
$p(r)$	probability of the received message
$p(s)$	language model: what are likely word sequences?
$p(r \mid s)$	translation model: how does the mutation/translation process happen? what operations are valid?

+ Noisy channel model

$$p(s \mid r) \propto p(r \mid s)\,p(s)$$

why?

~~$p(r)$~~	~~probability of the received~~ message
$p(s)$	language model: what are likely word sequences?
$p(r \mid s)$	translation model: how does the translation process happen? what operations are valid?



Machine translation

model

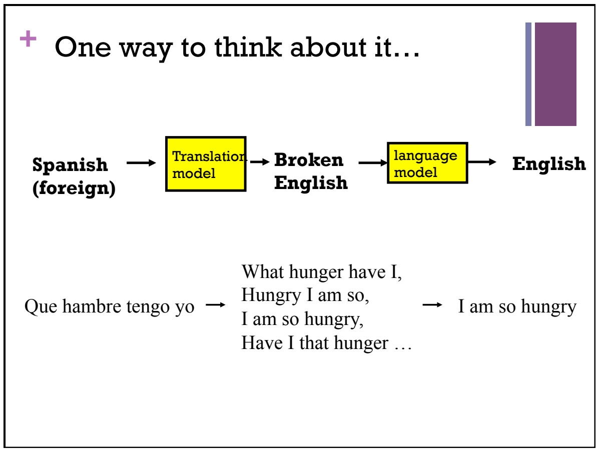
$$p(s \mid t) \propto p(t \mid s)\, p(s)$$

channel model
translation model

how do English words/phrases translate to Chinese?

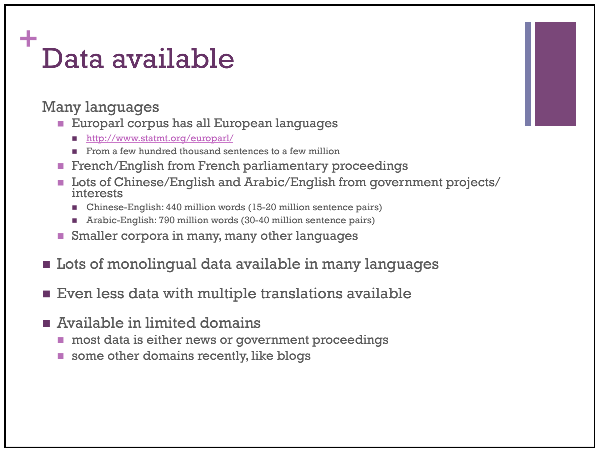
language model

what are likely English words/ word sequences?



One way to think about it…

Que hambre tengo yo → What hunger have I, Hungry I am so, I am so hungry, Have I that hunger … → I am so hungry

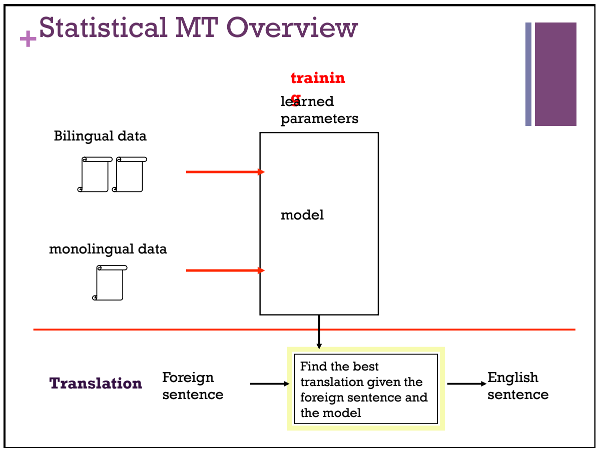


Data available

Many languages

- Europarl corpus has all European languages
  - http://www.statmt.org/europarl/
  - From a few hundred thousand sentences to a few million
- French/English from French parliamentary proceedings
- Lots of Chinese/English and Arabic/English from government projects/ interests
  - Chinese-English: 440 million words (15-20 million sentence pairs)
  - Arabic-English: 790 million words (30-40 million sentence pairs)
- Smaller corpora in many, many other languages

- Lots of monolingual data available in many languages
- Even less data with multiple translations available
- Available in limited domains
  - most data is either news or government proceedings
  - some other domains recently, like blogs

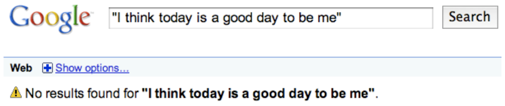


Statistical MT Overview

Language modeling

Answers the question of how likely a sentence is to be an English sentence

I think today is a good day to be me

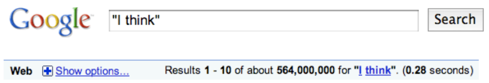

"I think today is a good day to be me"

Web Show options...

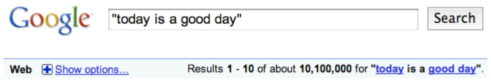
No results found for **"I think today is a good day to be me"**.

Language modeling

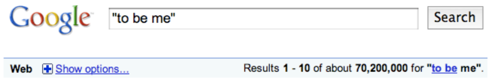
I think today is a good day to be me


"I think"

Web Show options... Results **1** - **10** of about **564,000,000** for "**I think**". (**0.28** seconds)


"today is a good day"

Web Show options... Results **1** - **10** of about **10,100,000** for "**today is a good day**".


"to be me"

Web Show options... Results **1** - **10** of about **70,200,000** for "**to be me**".

Language modeling

I think today is a good day to be me

Must also worry about the ordering of the words

How likely is "I think" to be preceded by "today is a good day"?

The main challenge with language modeling is dealing with data sparsity

+

Language modeling

Most common: n-gram language models

p(word | previous words)

More data the better (Google n-grams)

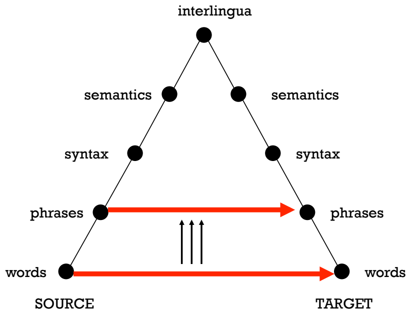
Domain is important!

+

DON'T
DROPPINGS!
Who owns this PET?
You must teachs your dog manners.
Then we are dog lovers each other.
thank you.

+

Translation models: MT Pyramid

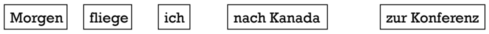


+

Phrase-Based Statistical MT

Morgen fliege ich nach Kanada zue Konferenze

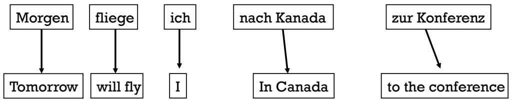
Phrase-Based Statistical MT


Morgen
fliege
ich
nach Kanada
zur Konferenz

Foreign input segmented in to phrases

- "phrase" is any sequence of words

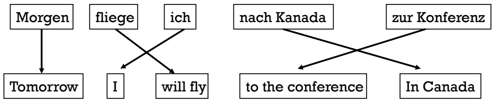
Phrase-Based Statistical MT


Morgen
fliege
ich
nach Kanada
zur Konferenz
Tomorrow
will fly
I
In Canada
to the conference

Each phrase is probabilistically translated into English

- P(to the conference | zur Konferenz)
- P(into the meeting | zur Konferenz)

Phrase-Based Statistical MT


Morgen
fliege
ich
nach Kanada
zur Konferenz
Tomorrow
I
will fly
to the conference
In Canada

Phrases are probabilistically re-ordered

See [Koehn et al, 2003] for an intro.

Advantages of Phrase-Based

Many-to-many mappings can handle non-compositional phrases

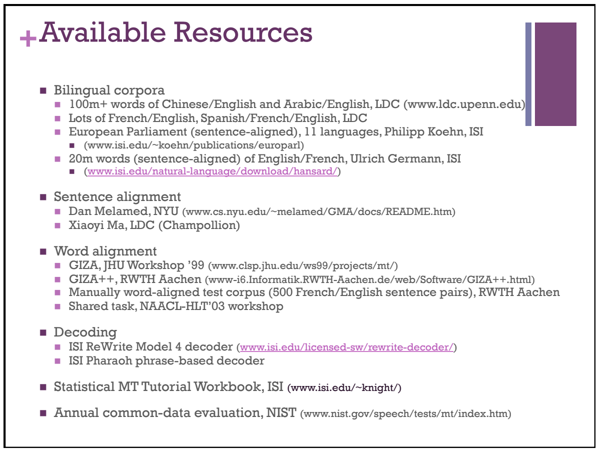
Easy to understand

Local context is very useful for disambiguating

- "Interest rate" → …
- "Interest in" → …

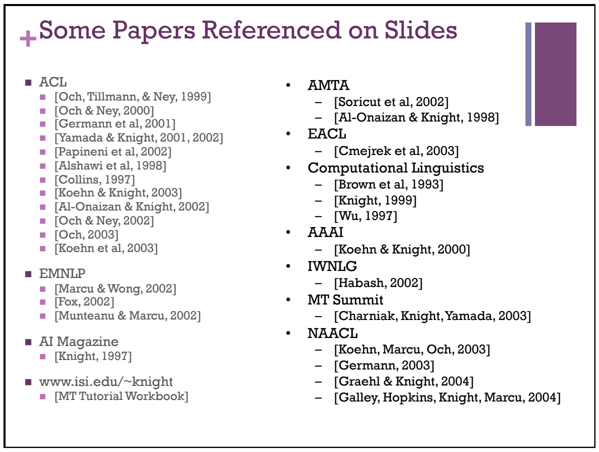
The more data, the longer the learned phrases

- Sometimes whole sentences



Available Resources

- Bilingual corpora
  - 100m+ words of Chinese/English and Arabic/English, LDC (www.ldc.upenn.edu)
  - Lots of French/English, Spanish/French/English, LDC
  - European Parliament (sentence-aligned), 11 languages, Philipp Koehn, ISI
    - (www.isi.edu/~koehn/publications/europarl)
  - 20m words (sentence-aligned) of English/French, Ulrich Germann, ISI
    - (www.isi.edu/natural-language/download/hansard/)
- Sentence alignment
  - Dan Melamed, NYU (www.cs.nyu.edu/~melamed/GMA/docs/README.htm)
  - Xiaoyi Ma, LDC (Champollion)
- Word alignment
  - GIZA, JHU Workshop '99 (www.clsp.jhu.edu/ws99/projects/mt/)
  - GIZA++, RWTH Aachen (www-i6.Informatik.RWTH-Aachen.de/web/Software/GIZA++.html)
  - Manually word-aligned test corpus (500 French/English sentence pairs), RWTH Aachen
  - Shared task, NAACL-HLT'03 workshop
- Decoding
  - ISI ReWrite Model 4 decoder (www.isi.edu/licensed-sw/rewrite-decoder/)
  - ISI Pharaoh phrase-based decoder
- Statistical MT Tutorial Workbook, ISI (www.isi.edu/~knight/)
- Annual common-data evaluation, NIST (www.nist.gov/speech/tests/mt/index.htm)



Some Papers Referenced on Slides

- ACL
  - [Och, Tillmann, & Ney, 1999]
  - [Och & Ney, 2000]
  - [Germann et al, 2001]
  - [Yamada & Knight, 2001, 2002]
  - [Papineni et al, 2002]
  - [Alshawi et al, 1998]
  - [Collins, 1997]
  - [Koehn & Knight, 2003]
  - [Al-Onaizan & Knight, 2002]
  - [Och & Ney, 2002]
  - [Och, 2003]
  - [Koehn et al, 2003]
- EMNLP
  - [Marcu & Wong, 2002]
  - [Fox, 2002]
  - [Munteanu & Marcu, 2002]
- AI Magazine
  - [Knight, 1997]
- www.isi.edu/~knight
  - [MT Tutorial Workbook]

- AMTA
  - [Soricut et al, 2002]
  - [Al-Onaizan & Knight, 1998]
- EACL
  - [Cmejrek et al, 2003]
- Computational Linguistics
  - [Brown et al, 1993]
  - [Knight, 1999]
  - [Wu, 1997]
- AAAI
  - [Koehn & Knight, 2000]
- IWNLG
  - [Habash, 2002]
- MT Summit
  - [Charniak, Knight, Yamada, 2003]
- NAACL
  - [Koehn, Marcu, Och, 2003]
  - [Germann, 2003]
  - [Graehl & Knight, 2004]
  - [Galley, Hopkins, Knight, Marcu, 2004]