http://www.youtube.com/watch?v=u_gBSWE_KYE

http://en.wikipedia.org/wiki/Vocaloid

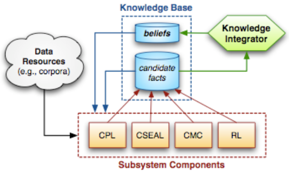
Natural Language Processing

CS151
David Kauchak

NELL

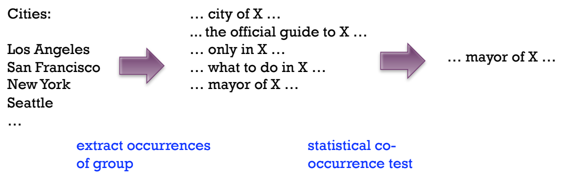
- NELL: Never-Ending Language Learning
  - http://rtw.ml.cmu.edu/rtw/
  - continuously crawls the web to grab new data
  - learns entities and relationships from this data
    - started with a seed set
    - uses learning techniques based on current KB to learn new information

NELL



- 4 different approaches to learning relationships
- Combine these in the knowledge integrator
  - idea: using different approaches will avoid overfitting
- Initially was wholly unsupervised, now some human supervision
  - cookies are food => internet cookies are food => files are food

+ An example learner: coupled pattern learner (CPL)



Cities:

Los Angeles
San Francisco
New York
Seattle
…

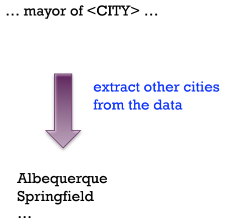
… city of X …
… the official guide to X …
… only in X …
… what to do in X …
… mayor of X …

… mayor of X …

extract occurrences of group

statistical co-occurrence test

+ CPL



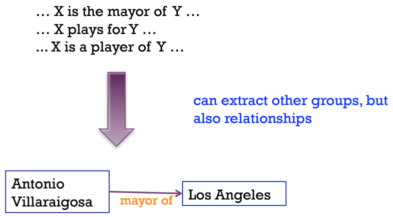
… mayor of <CITY> …

extract other cities from the data

Albequerque
Springfield
…

+ CPL

- Can also learn patterns with multiple groups



… X is the mayor of Y …
… X plays for Y …
… X is a player of Y …

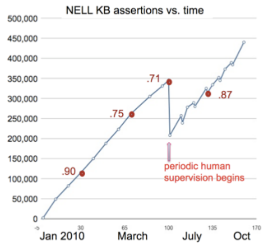
can extract other groups, but also relationships

Antonio Villaraigosa

mayor of

Los Angeles

+ NELL performance



estimated accuracy in red

For more details: http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf

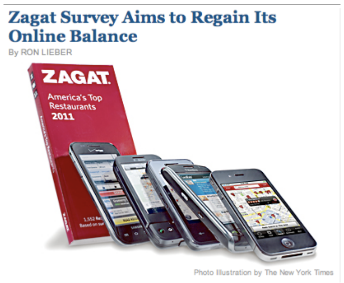
+ NELL

- The good:
  - Continuously learns
  - Uses the web (a huge data source)
  - Learns generic relationships
  - Combines multiple approaches for noise reduction
- The bad:
  - makes mistakes (overall accuracy still may be problematic for real world use)
  - does require some human intervention
  - still many general phenomena won't be captured

+ I say dear Watson…

- Question answering
- http://www.research.ibm.com/deepqa/

+ Why is NLP hard?



Zagat Survey Aims to Regain Its Online Balance

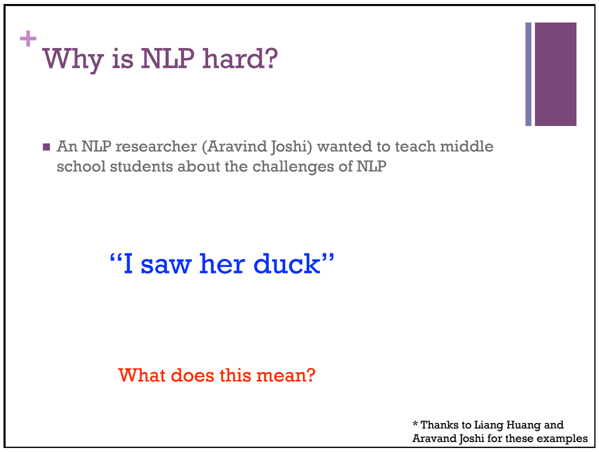
By RON LIEBER

Photo Illustration by The New York Times

What does this mean? How did you figure it out?

+ Why is NLP hard? "My lessons in life"

- What does: "… no one bats a hundred every time" mean?
- What happened to his hair?
- What does "… the lessons I have leant along the way." mean?
- "The parents were concerned."
  - parents of whom/what?
- "She would always complain about the breakfast."
  - who is "she"?



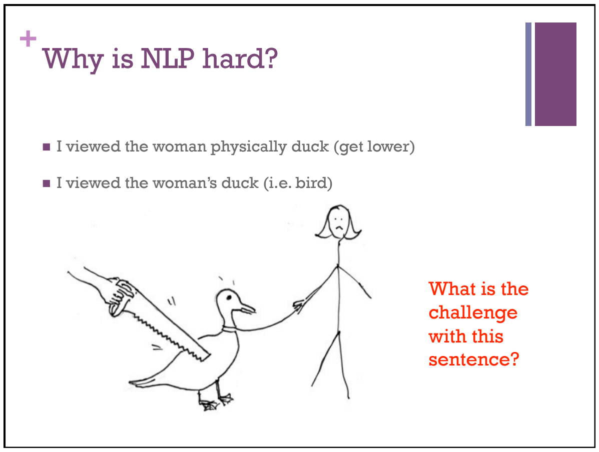
+ Why is NLP hard?

- An NLP researcher (Aravind Joshi) wanted to teach middle school students about the challenges of NLP

"I saw her duck"

What does this mean?

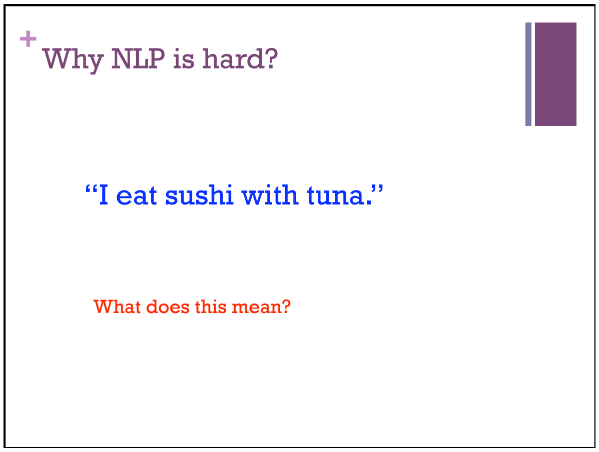
* Thanks to Liang Huang and Aravand Joshi for these examples



+ Why is NLP hard?

- I viewed the woman physically duck (get lower)
- I viewed the woman's duck (i.e. bird)

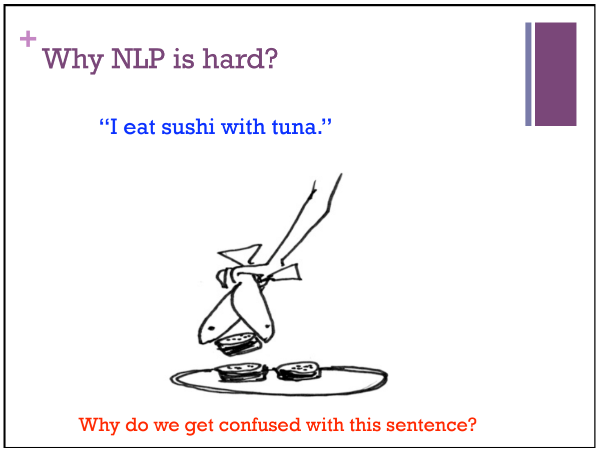
What is the challenge with this sentence?



+ Why NLP is hard?

"I eat sushi with tuna."

What does this mean?

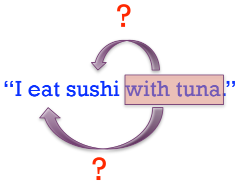


+ Why NLP is hard?

"I eat sushi with tuna."

Why do we get confused with this sentence?

Why NLP is hard?



"I eat sushi with tuna."

Structural ambiguity: what is "with tuna" associated with (PP attachment)

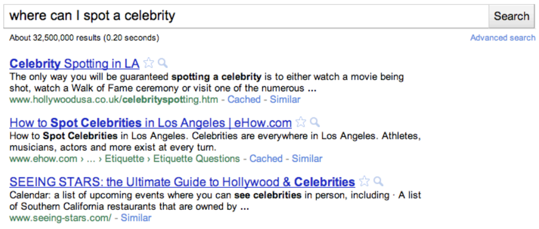
Why NLP is hard?

and some lexical ambiguity…

Maybe Google has it all figured out…

where can I spot a celebrity

Maybe Google has it all figured out…



where can I spot a celebrity

About 32,500,000 results (0.20 seconds)

Celebrity Spotting in LA

The only way you will be guaranteed **spotting a celebrity** is to either watch a movie being shot, watch a Walk of Fame ceremony or visit one of the numerous ...

www.hollywoodusa.co.uk/**celebrityspot**ting.htm - Cached - Similar

How to **Spot Celebrities** in Los Angeles | eHow.com

How to **Spot Celebrities** in Los Angeles. Celebrities are everywhere in Los Angeles. Athletes, musicians, actors and more exist at every turn.

www.ehow.com › ... › Etiquette › Etiquette Questions - Cached - Similar

SEEING STARS: the Ultimate Guide to Hollywood & **Celebrities**

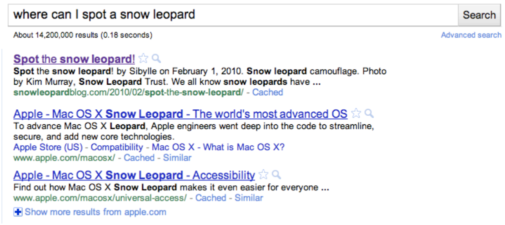
Calendar: a list of upcoming events where you can **see celebrities** in person, including · A list of Southern California restaurants that are owned by ...

www.seeing-stars.com/ - Similar

Maybe Google has it all figured out…

where can I spot a snow leopard

Maybe Google has it all figured out…



where can I spot a snow leopard

About 14,200,000 results (0.18 seconds)

Spot the **snow leopard!**

Spot the **snow leopard**! by Sibylle on February 1, 2010. **Snow leopard** camouflage. Photo by Kim Murray, **Snow Leopard** Trust. We all know **snow leopards** have **...**
snowleopardblog.com/2010/02/**spot**-the-**snow**-**leopard**/ - Cached

Apple - Mac OS X **Snow Leopard** - The world's most advanced OS

To advance Mac OS X **Leopard**, Apple engineers went deep into the code to streamline, secure, and add new core technologies.
Apple Store (US) - Compatibility - Mac OS X - What is Mac OS X?
www.apple.com/macosx/ - Cached - Similar

Apple - Mac OS X **Snow Leopard** - Accessibility

Find out how Mac OS X **Snow Leopard** makes it even easier for everyone **...**
www.apple.com/macosx/universal-access/ - Cached - Similar

Show more results from apple.com

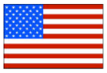
NLP applications

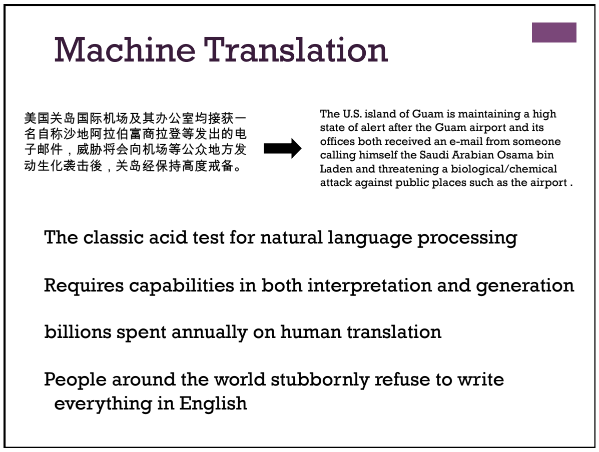
What are some places where you have seen NLP used?

What are NLP problems?

Language translation

Yo quiero
Taco Bell





Machine Translation

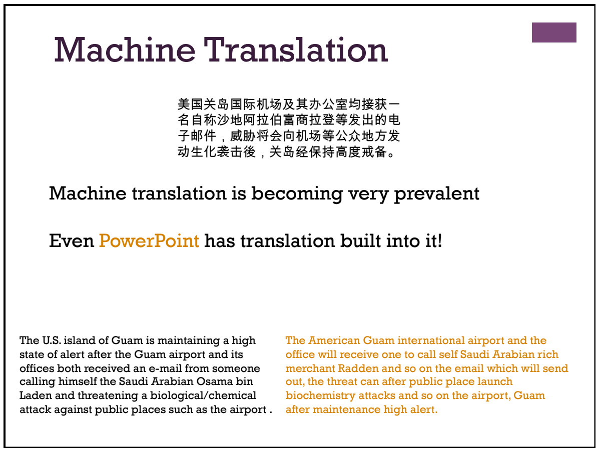
美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。 ➡ The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack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

The classic acid test for natural language processing

Requires capabilities in both interpretation and generation

billions spent annually on human translation

People around the world stubbornly refuse to write everything in English



Machine Translation

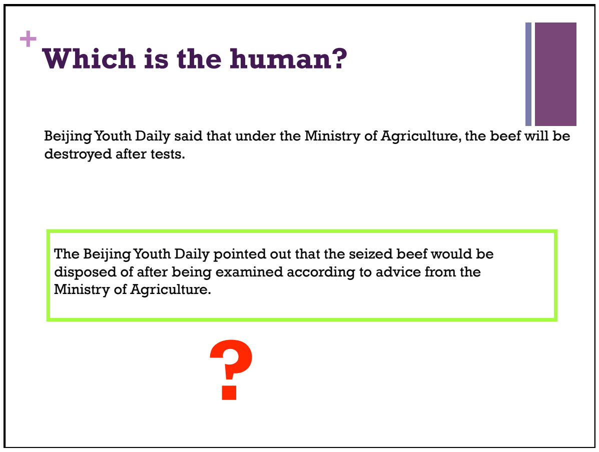
美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

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 .

The American Guam international airport and the office will receive one to call self Saudi Arabian rich merchant Radden and so on the email which will send out, the threat can after public place launch biochemistry attacks and so on the airport, Guam after maintenance high alert.

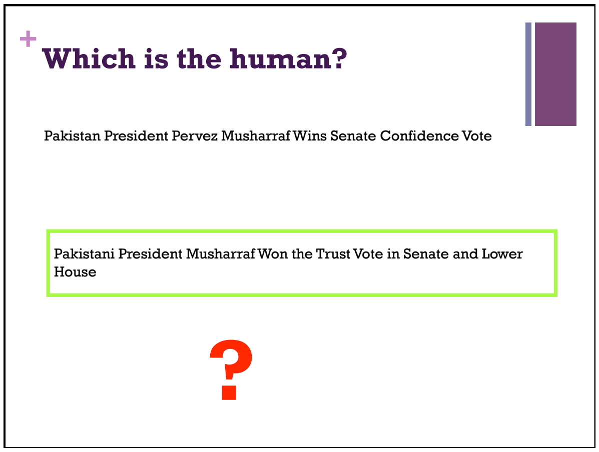


Which is the human?

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

?

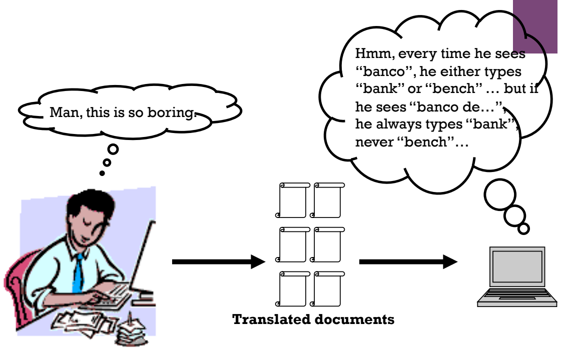
+ Which is the human?

There was not a single vote against him."

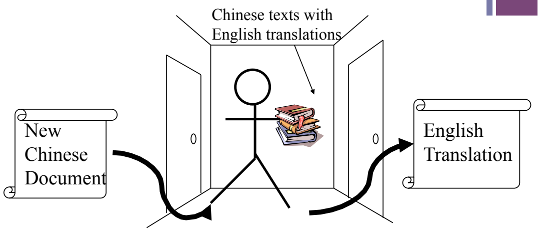
No members vote against him. "

?

+ 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

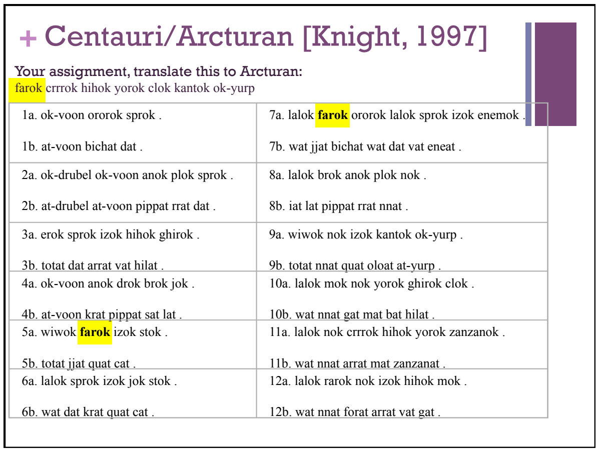


You can teach yourself to translate Chinese using *only* bilingual data (without grammar books, dictionaries, any people to answer your questions…)

+ 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 .

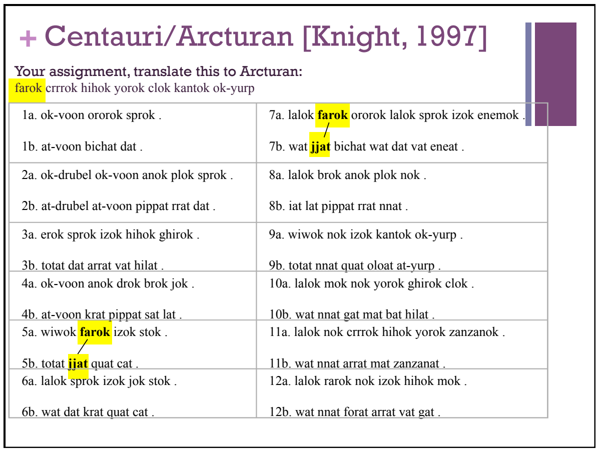


Centauri/Arcturan [Knight, 1997]

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 .

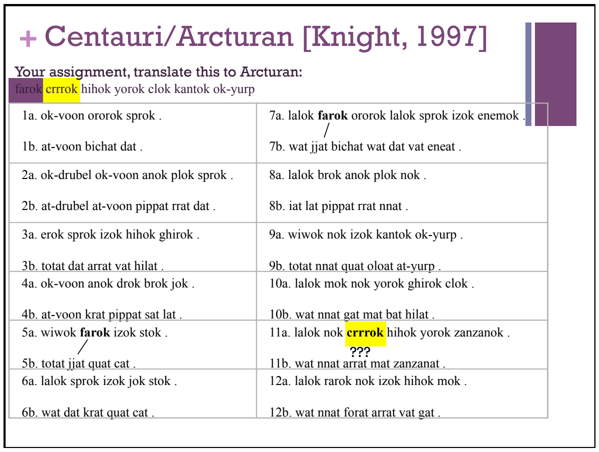


Centauri/Arcturan [Knight, 1997]

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 .

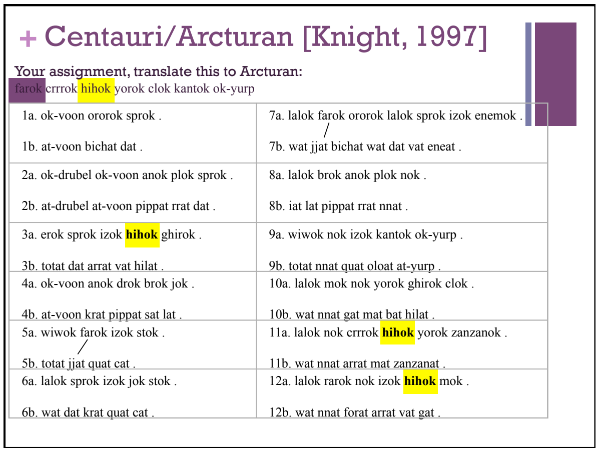


Centauri/Arcturan [Knight, 1997]

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 .

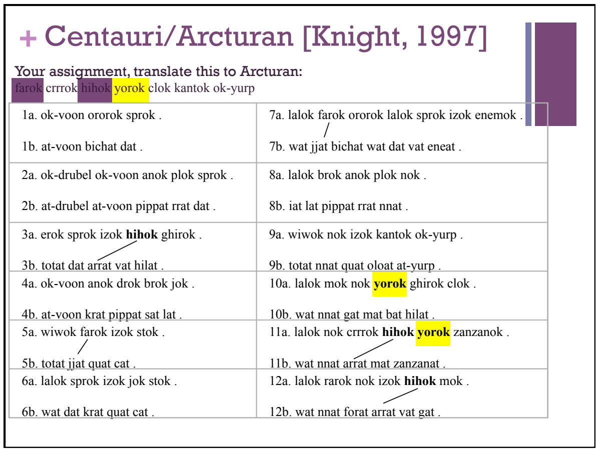


Centauri/Arcturan [Knight, 1997]

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 .

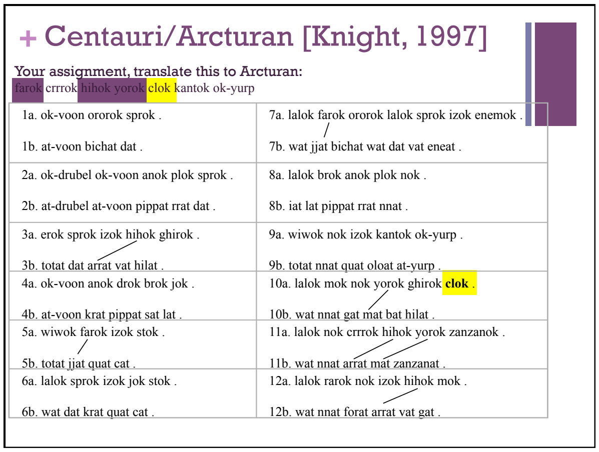


Centauri/Arcturan [Knight, 1997]

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 .

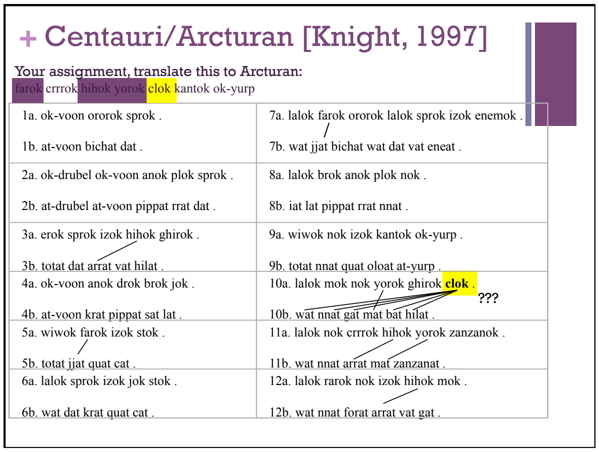


Centauri/Arcturan [Knight, 1997]

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 .

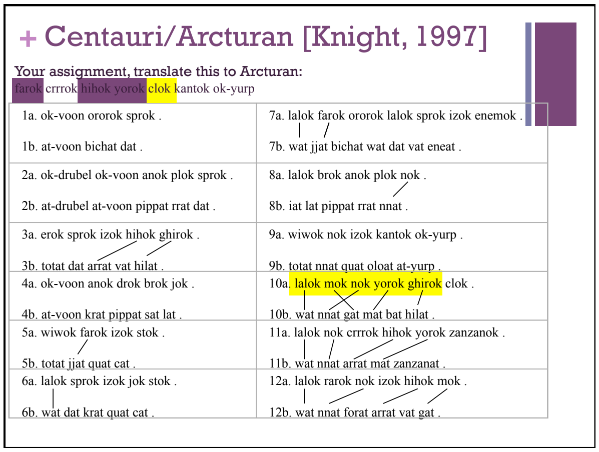


Centauri/Arcturan [Knight, 1997]

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 .

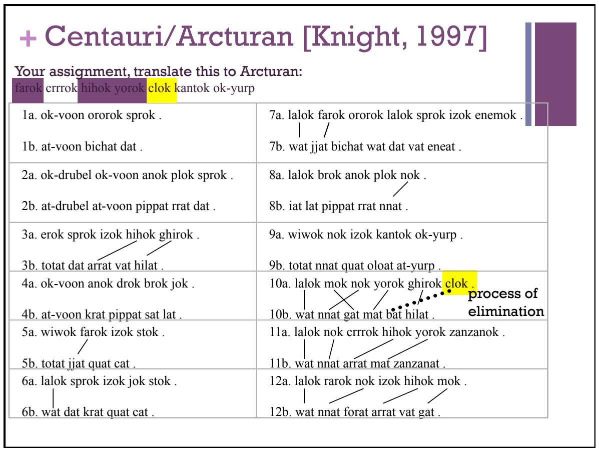


Centauri/Arcturan [Knight, 1997]

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 .

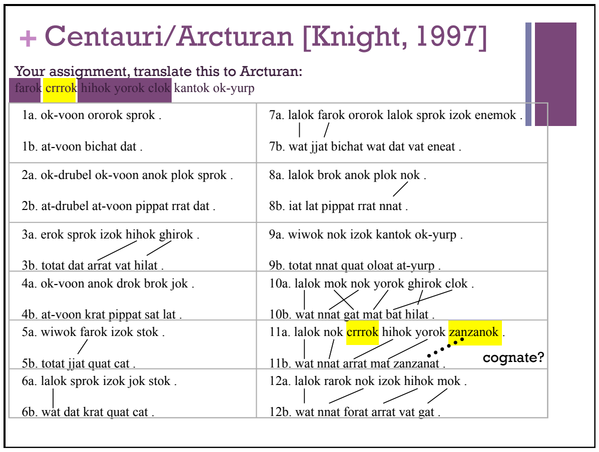


+ Centauri/Arcturan [Knight, 1997]

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

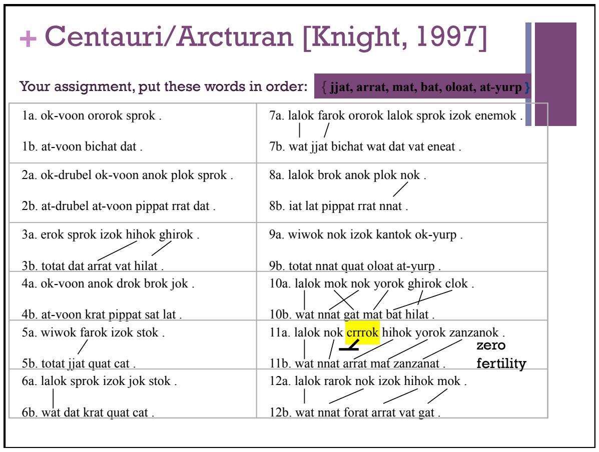


+ Centauri/Arcturan [Knight, 1997]

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?

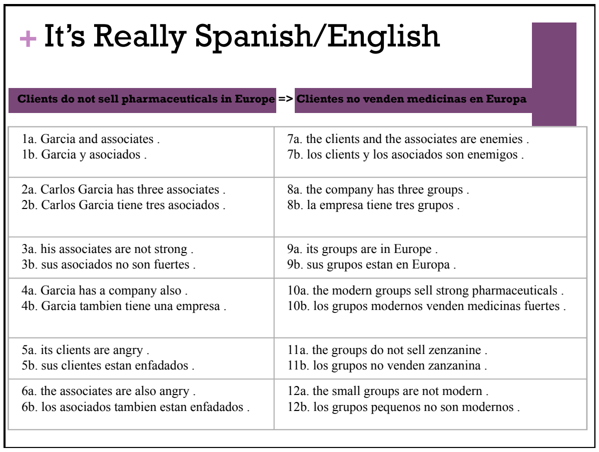


+ Centauri/Arcturan [Knight, 1997]

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 .

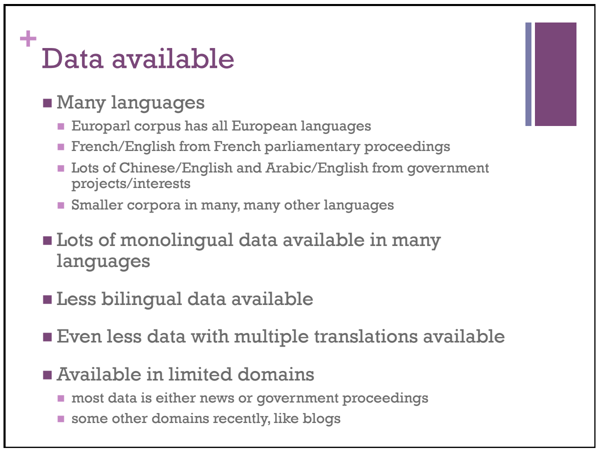
zero fertility



+ It’s Really Spanish/English

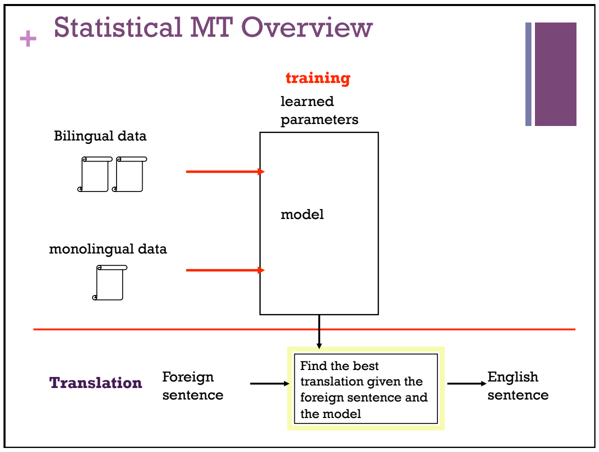
Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

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 .



Data available

- Many languages
  - Europarl corpus has all European languages
  - French/English from French parliamentary proceedings
  - Lots of Chinese/English and Arabic/English from government projects/interests
  - Smaller corpora in many, many other languages
- Lots of monolingual data available in many languages
- Less bilingual data available
- 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

training

learned parameters

Bilingual data

model

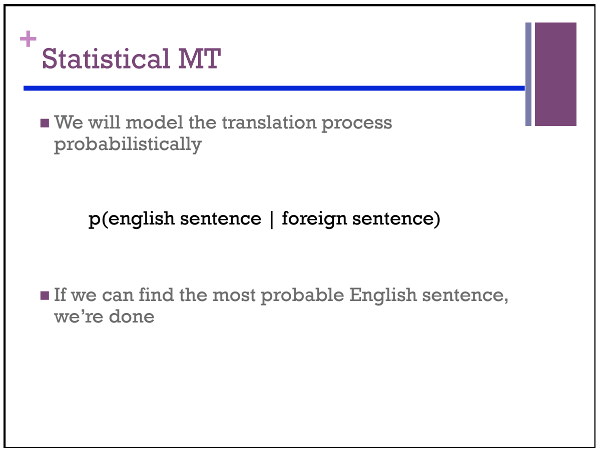
monolingual data

Translation

Foreign sentence

Find the best translation given the foreign sentence and the model

English sentence

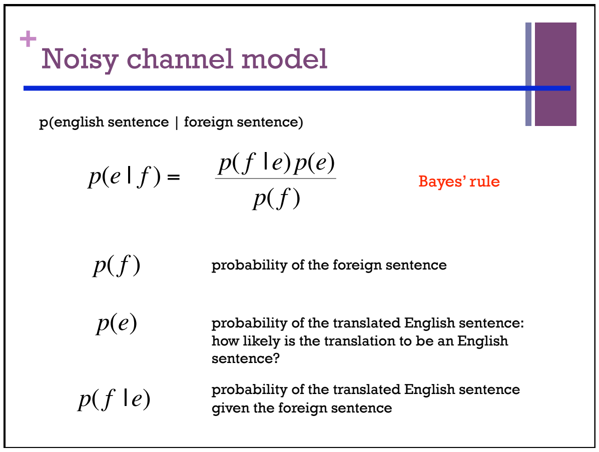


Statistical MT

- We will model the translation process probabilistically

p(english sentence | foreign sentence)

- If we can find the most probable English sentence, we're done



Noisy channel model

p(english sentence | foreign sentence)

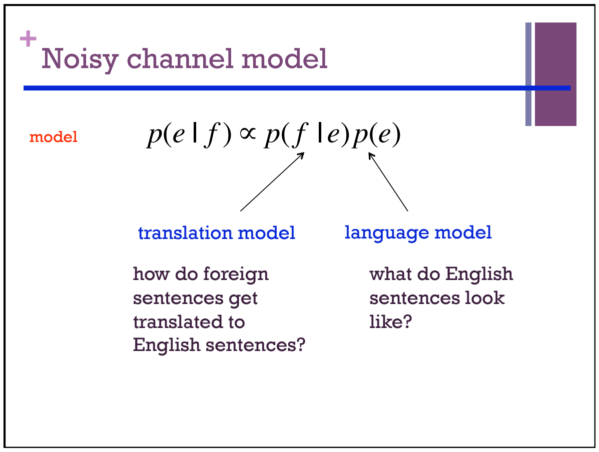
$$p(e \mid f) = \frac{p(f \mid e)p(e)}{p(f)}$$

Bayes' rule

$p(f)$ probability of the foreign sentence

$p(e)$ probability of the translated English sentence: how likely is the translation to be an English sentence?

$p(f \mid e)$ probability of the translated English sentence given the foreign sentence



Noisy channel model

model

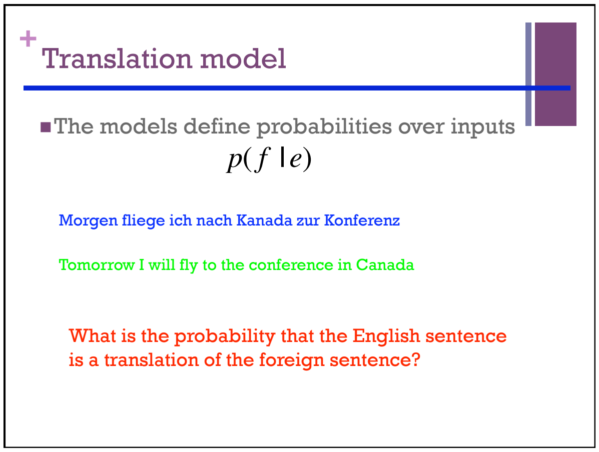
$$p(e \mid f) \propto p(f \mid e)p(e)$$

translation model

how do foreign sentences get translated to English sentences?

language model

what do English sentences look like?



Translation model

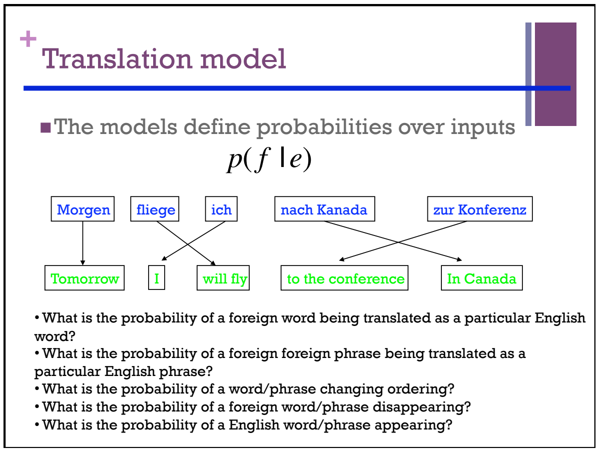
- The models define probabilities over inputs

$$p(f \mid e)$$

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

What is the probability that the English sentence is a translation of the foreign sentence?

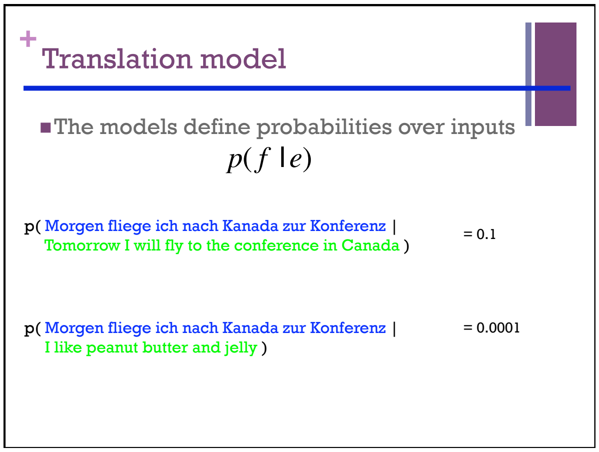


Translation model

- The models define probabilities over inputs

$$p(f \mid e)$$

- What is the probability of a foreign word being translated as a particular English word?
- What is the probability of a foreign foreign phrase being translated as a particular English phrase?
- What is the probability of a word/phrase changing ordering?
- What is the probability of a foreign word/phrase disappearing?
- What is the probability of a English word/phrase appearing?



Translation model

- The models define probabilities over inputs

$$p(f \mid e)$$

p(Morgen fliege ich nach Kanada zur Konferenz | Tomorrow I will fly to the conference in Canada) = 0.1

p(Morgen fliege ich nach Kanada zur Konferenz | I like peanut butter and jelly) = 0.0001

+ Language model

- The models define probabilities over inputs

$p(e)$

Tomorrow I will fly to the conference in Canada

- What is the probability that an English sentence starts with "Tomorrow"?
- What is the probability of seeing the word "fly"?
- What is the probability of seeing the word "fly" given that the previous two words are "I will"?

+ Language model

- The models define probabilities over inputs

$p(e)$

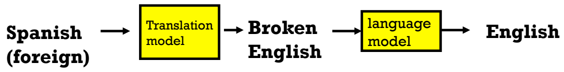
p(Tomorrow I will fly to the conference in Canada) = .001

p(fly conference the Canada Tomorrow will I to) = .0000001

+ What is a probability distribution?

- A probability distribution defines the probability over a space of possible inputs
- For the language model, what is the space of possible inputs?
  - A language model describes the probability over ALL possible combinations of English words
- For the translation model, what is the space of possible inputs?
  - ALL possible combinations of foreign words with ALL possible combinations of English words

+ 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

Translation (aka decoding)

$$p(e \mid f) \propto p(f \mid e) p(e)$$

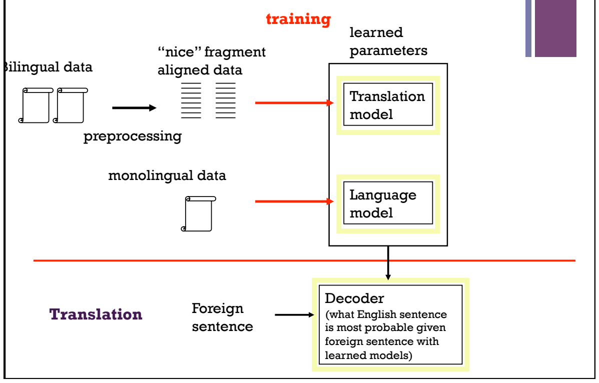
- Let's assume we have a translation model and a language model
- Given a foreign sentence, what question do we want to ask to translate that sentence into English?

$$\arg_e \max p(e \mid f) \propto p(f \mid e) p(e)$$

Training a system

- What type of data would be useful for training a translation model: p(f|e)?
  - Bilingual data: documents/sentences that are equivalent, but in two languages
- What type of data would be useful for training a language model: p(e)?
  - Monolingual data: lots of text in English

Statistical MT Overview



training

learned parameters

3ilingual data

"nice" fragment aligned data

preprocessing

Translation model

monolingual data

Language model

Translation

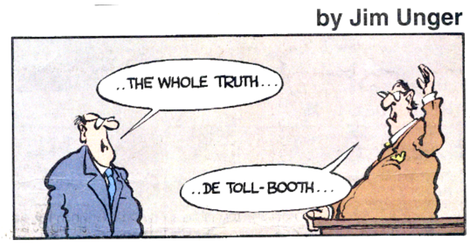
Foreign sentence

Decoder
(what English sentence is most probable given foreign sentence with learned models)

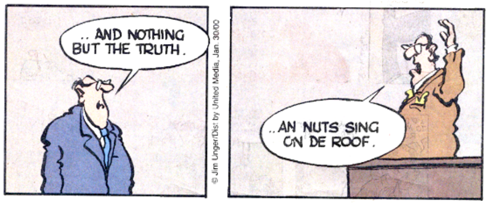
A bad language model

HERMAN
REPEAT AFTER ME... I SWEAR TO TELL THE TRUTH...
I SWERVE TO SMELL DE SOUP...

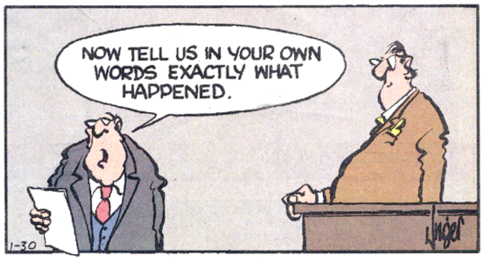
A bad language model



A bad language model

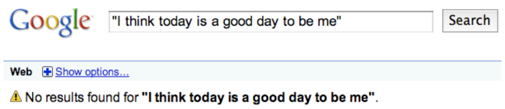

.. AND NOTHING BUT THE TRUTH.
© Jim Unger/Dist by United Media, Jan. 30/00
..AN NUTS SING ON DE ROOF.

A bad language model



Language modeling

I think today is a good day to be me



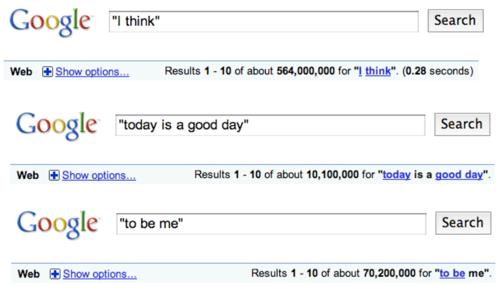
"I think today is a good day to be me"

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

Language modeling is about dealing with data sparsity!

Language modeling

I think today is a good day to be me



"I think"

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

"today is a good day"

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

"to be me"

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

Our friend the chain rule

Step 1: decompose the probability

$P(\text{I think today is a good day to be me}) =$

$P(\text{I}) \times$

$P(\text{think} \mid \text{I}) \times$

$P(\text{today} \mid \text{I think}) \times$

$P(\text{is} \mid \text{I think today}) \times$

$P(\text{a} \mid \text{I think today is}) \times$

$P(\text{good} \mid \text{I think today is a}) \times$

…

How can we simplify these?

The n-gram Approximation

Assume each word depends only on the previous n-1 words (e.g. trigram: three words total)

$P(\text{is} \mid \text{I think today}) \approx P(\text{is} \mid \text{think today})$

$P(\text{a} \mid \text{I think today is}) \approx P(\text{a} \mid \text{today is})$

$P(\text{good} \mid \text{I think today is a}) \approx P(\text{good} \mid \text{is a})$

Estimating probabilities

- How do we find probabilities? $P(\text{is} \mid \text{think today})$
- Get real text, and start counting!

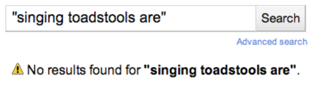
$$P(\text{is} \mid \text{think today}) = \frac{\text{count}(\text{think today is})}{\text{count}(\text{think today})}$$

Smoothing

Is this sentence reasonable?

The Singing Toadstools are a new band.

$$P(\text{are} \mid \text{Singing Toadstools}) = \frac{\text{count(Singing Toadtools are)}}{\text{count(Singing Toadstools)}}$$



"singing toadstools are"

⚠ No results found for **"singing toadstools are"**.

Smoothing

P(I think today is a good day to be me) =

P(I) x

P(think | I) x

P(today | I think) x

P(is | think today) x

P(a | today is) x

P(good | is a) x

…

If any of these has never been seen before, prob = 0!

Ideas? What did you do for NB?

Smoothing: Add One (Laplacian)

- Add one smoothing: $P(c \mid ab) \approx \frac{C(abc)+1}{C(ab)+V}$
- Works very poorly. DO NOT DO THIS
- Add delta smoothing: $P(c \mid ab) \approx \frac{C(abc)+\delta}{C(ab)+\delta V}$
- Still very bad. DO NOT DO THIS