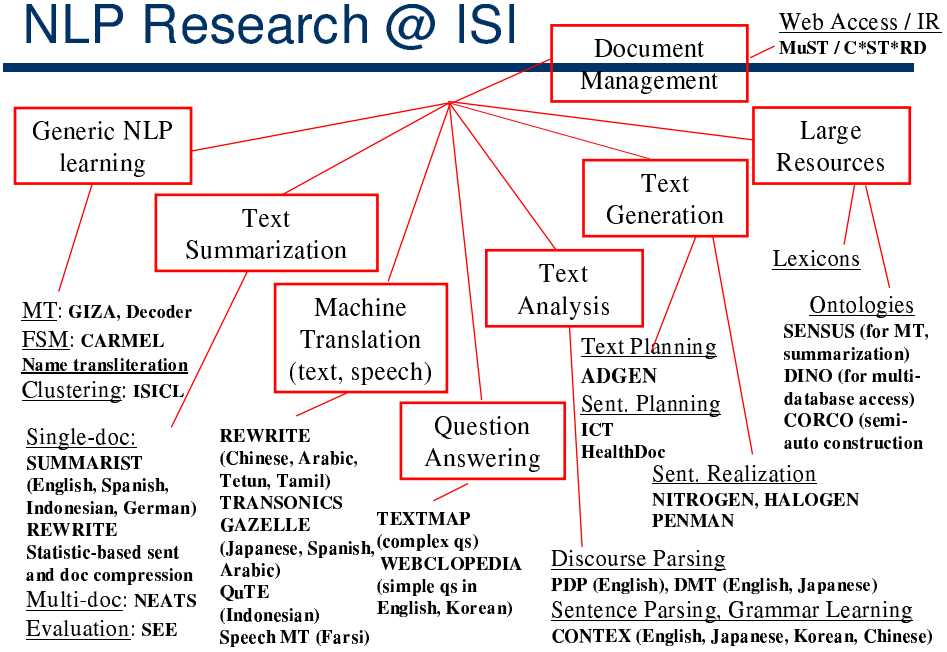
Research Challenges in Natural Language Processing

Daniel Marcu
Information Sciences Institute and Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu
http://www.isi.edu/~marcu/

What is language good for?

- Communication
  - Speech
  - Writing
- Storage of knowledge

NLP Research @ ISI



ISI Natural Language Group

- 9 PhD-level researchers, 3 full-time programmers, 15+ PhD students, postdocs, visitors, other students and helpers…
- Significant impact on Computational Linguistics worldwide:
  - Best Paper awards at AAAI-00, ACL-01, ACL-02
  - Good performance in (inter)national evaluations and demos
    - Best Chinese-English and Arabic-English MT systems (2003)
    - Hindi-English MT system in a month (June 2003)
  - Systems and tools available for licensing
  - Chairs and organizers of conferences, workshops, ACL, etc.
  - Courses:
    - CS 562 (Fall: Knight, Marcu);
    - CS 544 (Spring: Hobbs, Hovy, Marcu)

http://www.isi.edu/natural-language/nlp-at-isi.html

Research challenges

- Automatic knowledge acquisition from text
- Knowledge representation
  - Probability distributions; taxonomies; etc.
- Learning algorithms
  - Information theory; statistical estimation
- Scalability
  - Billion word corpora; cluster computer computations
- Problem definition

Sample applications

Question Answering

- Input: natural language questions
- Resource: 2 GB of text/the Web
- Output: short answer (1-5 words)

TextMap

- Who killed Lee Harvey Oswald?
  - Belli's clients have included Jack Ruby, who killed John F. Kennedy's assassin Lee Harvey Oswald, and Jim and Tammy Bakker.
- What river is called "China's Sorrow"?
  - The Xiaolongdi Dam Project will control formerly disastrous annual flooding on the Yellow River, known as the Sorrow of the Chinese Nation.
- What are the people who make fireworks called?
  - With an extravagant fireworks display dancing up the sides of the Washington Monument as the 21st century arrived in America, President Clinton called on the nation to fear not the future but to ``welcome it, embrace it, create it.''
- How many people live in Chile?
  - 2
- What is the distance between Saturn and the Sun?
  - more than 2 meters
- Progress:
  - 2002: 33% questions answered correctly
  - 2003: 48% questions answered correctly

Machine Translation

- Input:
  - Chinese/French/Arabic text
- Output:
  - Corresponding English translation
- Resource:
  - Lots of bilingual text
    - 100M words Arabic-English data
    - 140M words Chinese-English data

Progress

(Charles Wayne: DARPA)

2002

insistent Wednesday may recurred her trips to Libya tomorrow for flying

Cairo 6-4 (AFP) - an official announced today in the Egyptian lines company for flying Tuesday is a company " insistent for flying " may resumed a consideration of a day Wednesday tomorrow her trips to Libya of Security Council decision trace international the imposed ban comment .

And said the official " the institution sent a speech to Ministry of Foreign Affairs of lifting on Libya air , a situation her receiving replying are so a trip will pull to Libya a morning Wednesday " .

2003

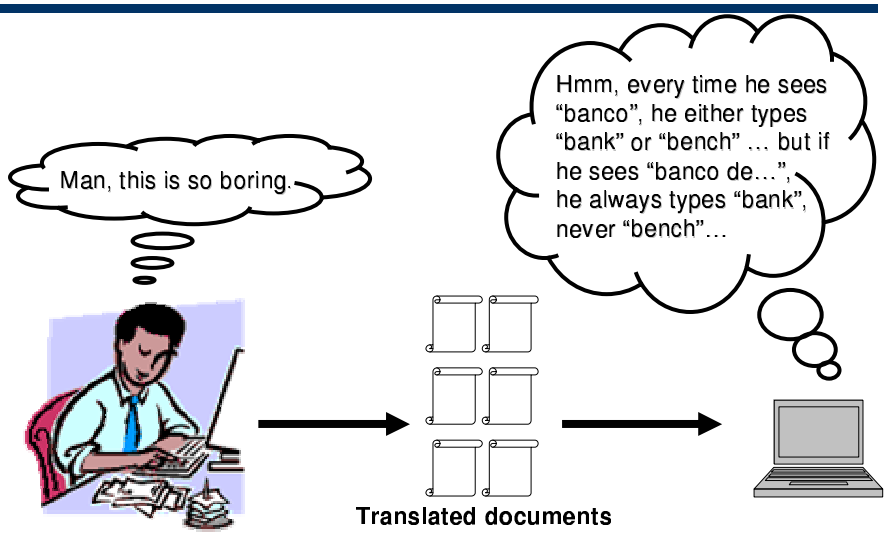
Egyptair Has Tomorrow to Resume Its Flights to Libya

Cairo 4-6 (AFP) - said an official at the Egyptian Aviation Company today that the company egyptair may resume as of tomorrow, Wednesday its flights to Libya after the International Security Council resolution to the suspension of the embargo imposed on Libya.

" The official said that the company had sent a letter to the Ministry of Foreign Affairs, information on the lifting of the air embargo on Libya, where it had received a response, the first take off a trip to Libya on Wednesday morning ".

How does Statistical Machine Translation work?

Data-Driven Machine Translation



Translate the sentence [Knight, 1997] “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 .

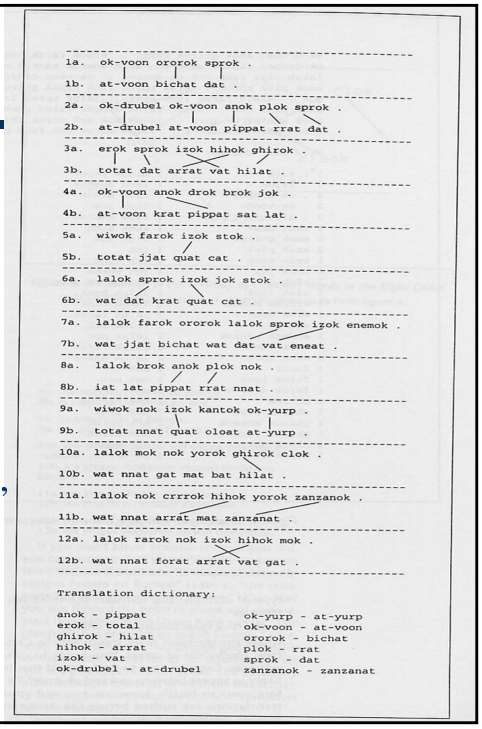
Translate the sentence [Knight, 1997] “farok crrrok hihok yorok clok kantok ok-yurp”

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 .

Partial solution:

Solution:

"jjat arrat mat bat oloat at-yurp"



1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

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

Translation dictionary:

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

Translate the sentence

"clients do not sell pharmaceuticals in Europe"

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clientes 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 .

The Tetun-English Experiment
Best Human: 4.55 on a scale of 1 to 5.

Training: 1102 sentence pairs

Registering To Vote .
Regista Ba Vota .

To be a part of the upcoming popular consultation and to have your say in the future of East Timor , you must first register to vote .
Atu halo parte iha consulta popular ida mai ne'e no atu hato'o imi nia lian ba Timor-Loro-Sa'e nia futuru , imi tenke tau uluk naran atu vota .

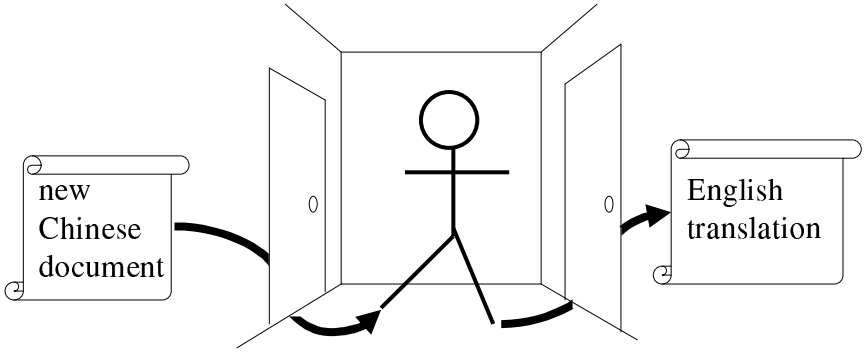
Registration has not started yet .
Arolamentu sei dauk hahu .

Sample Test Sentence:
Sr Ian Martin, Reprezentante Espesial Sekretariu Jeral nian ba Timor Lorosa's esplika, "Ami ba fatin tolu ne'e tanba fatin hirak ne'e maka fatin sira be ami iha preokupasaun boot liu".

Translation produced by non-Tetun speaker:
Mr. Ian Martin, the [UN] Secretary General's Special Envoy (= representative) for East Timor, explained: "We go to the three locations, because these locations are locations that we are very worried about".

Hidden Reference:
"We went to those three places because they are places that are of the most concern to us", Mr. Ian Martin, the Special Representative of the Secretary-General for East Timor, explained.

Chinese Room Experiment



Chinese texts with
English translations

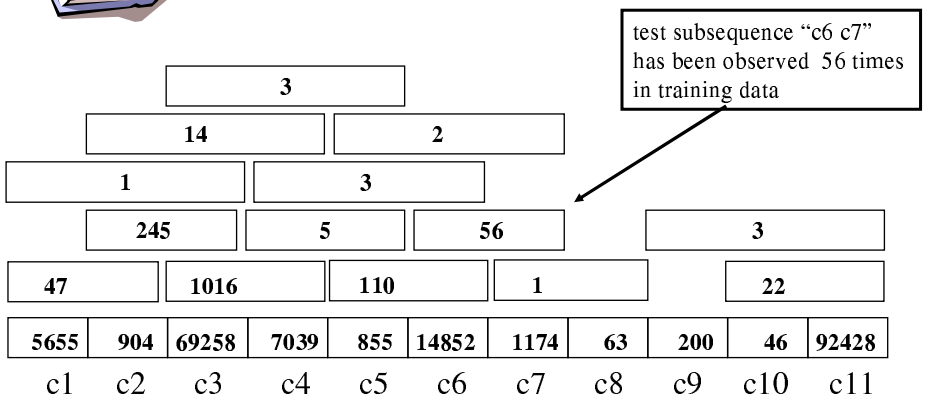
HIGH ACCURACY

DOES THIS PERSON KNOW CHINESE?

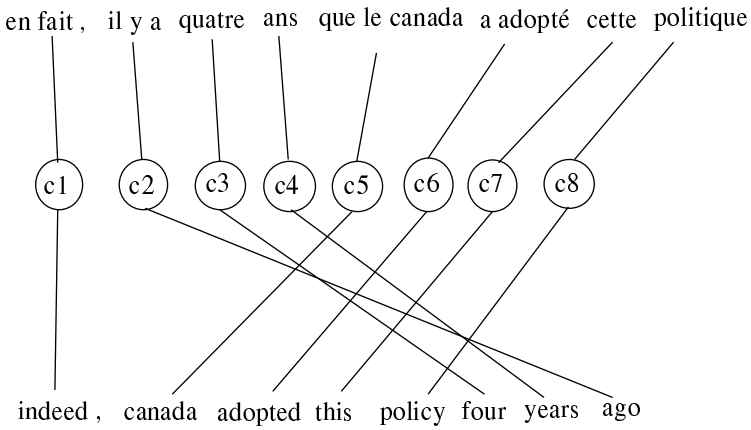
Chinese word or phrase => sentence pairs containing it

Chinese Room Experiment

170k sentence pairs of bilingual training data
(3.5m words translated)



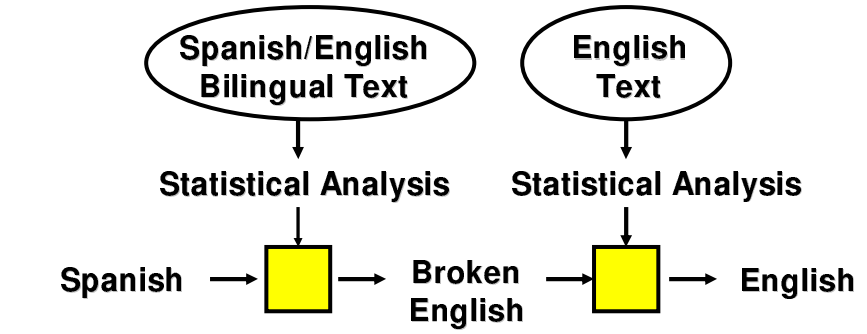
Example of generative story: [Marcu and Wong, 2002]



Dictionary learned automatically

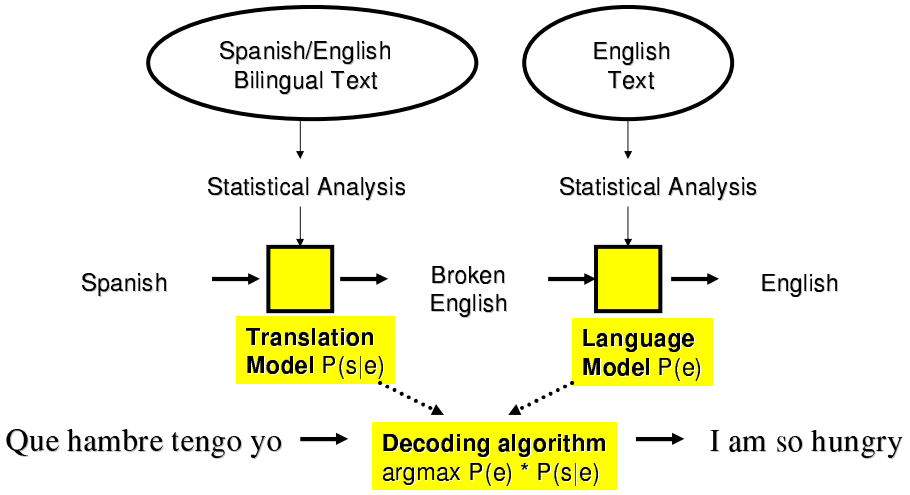
- possessives:
  - (la réaction de : 's response) :5.30995e-06
  - (la réponse de : 's response) :5.29405e-06
- complex nominals:
  - (monsieur le président : mr. speaker) :4.07544e-05
- negation:
  - (ne est pas : is not) :0.000173962
  - (est inadmissible : is not good enough) :2.31636e-06
  - (ne est pas ici : is not here) :3.57434e-06
- adj-noun order:
  - (fièvre typhoïde : paratyphoid fever) :1.77204e-06
- paraphrase:
  - (fonction de les plus importantes : most important responsibility) :1.77e-06

Statistical MT Systems



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

Statistical MT Systems



Discussion

- Bilingual texts encode lots of knowledge about how sentences should be translated.
- The Tetun/Chinese experiments were too easy
  - Participants utilized their own knowledge of English
- Yet, experiments helped us learn a lot about how we may go about building better SMT systems

Assignment

- /auto/home-scf-22/csci544/Coding on aludra has two files
  - training: 1.77M parallel sentences of Venusian-Martian (~37M words on each side)
  - Test: 15 sentences in Venusian.
- Translate the test sentences into Martian and email to marcu@isi.edu two files as distinct attachments
  - A file called "translations-yourname", one translation per line (total of 15 lines).
  - A file called "discussion-yourname" that explains what you did to produce the translations and the lessons you have learned.
- Due date
  - April 5, 2004, 11:59am.
- Resources
  - Tetun experiment paper: http://www.isi.edu/~marcu/papers/mt-scarce-2002.pdf
- Suggestions and things that worked in the past
  - Create a version of *training* that contains only sentences that have words that occur in the test corpus.
  - Write a program that for every Venusian test phrase displays 10-20 sentences (with the corresponding Martian translations) in which that phrase was used in *training*.
  - Don't underestimate the utility of Unix tools (grep "phrase" –A 2 < filename)
  - There is no connection between the length of the words in training and their length in the original language.
- I will post the answer key and the winner on April 12, 2004, on my web site at (http://www.isi.edu/~marcu/DecodingExperiment.html) .

Example

- Training
  - ygrybthy snrv arqempnt rr arqempnt rx ieyiyt utbty e igawphwm sqkxcmnj
  - dqpsn kilugn mjhahl dymdrcn kilugn ewroeq suaf vaysgqaax nlgmfc qsfmg wxdd ojfjhmqb
  - ouwofvnfs xte ptcev drg dvapphadt iguv okcxd utbty e igawphwm jsdpnhrt
  - ktdxyc jdju wjuk thpxqoabl lfjm thpxqoabl jkcv esld nlgmfc qsfmg wxdd bgpdnwar
  - flkb rgdciccjr arqempnt edvkmdcl okcxd cmql frykfoen fm sqkxcmnj
  - tdkqinu idvsr hghdxel thpxqoabl wjrggy wiscuhad esld pyxnt qsfmg vlifj ojfjhmqb
- Test
  - xqmrcbigr frykfoen xte etcumpvbv
  - kph vgu pifsfj uarns meruq esiccocv vnnpva sbxojibtb wpfcbvv rpjprpf ytnrtya uystw frykfoen cpnb
  - ixaiytk fanyi vnnpva lxskwdnx irnago ytnrtya kxljfwuw uaklk vnnpva aqjau dvapphadt lmynwk

Thank you!