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
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
    - 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-is.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

(Carlton 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 egypair 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

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

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1a.</td>
<td>ok-voon ororok sprok .</td>
<td>7a.</td>
</tr>
<tr>
<td>1b.</td>
<td>at-voon bichat dat .</td>
<td>7b.</td>
</tr>
<tr>
<td>2a.</td>
<td>ok-drabel ok-voon anok plok sprok .</td>
<td>8a.</td>
</tr>
<tr>
<td>2b.</td>
<td>at-drabel at-voon pippat rat dat .</td>
<td>8b.</td>
</tr>
<tr>
<td>3a.</td>
<td>erok sprok izok hihok ghirok .</td>
<td>9a.</td>
</tr>
<tr>
<td>3b.</td>
<td>totat dat arrat vat hilat .</td>
<td>9b.</td>
</tr>
<tr>
<td>4a.</td>
<td>ok-voon anok erok brok jok .</td>
<td>10a.</td>
</tr>
<tr>
<td>4b.</td>
<td>at-voon krat pippat sat lat .</td>
<td>10b.</td>
</tr>
<tr>
<td>5a.</td>
<td>wiwok farok izok stok .</td>
<td>11a.</td>
</tr>
<tr>
<td>5b.</td>
<td>totat jiat quat cat .</td>
<td>11b.</td>
</tr>
<tr>
<td>6a.</td>
<td>lalok sprok izok jok stok .</td>
<td>12a.</td>
</tr>
<tr>
<td>6b.</td>
<td>wat dat kratquat cat .</td>
<td>12b.</td>
</tr>
</tbody>
</table>
Partial solution:

Solution:
“jjat arrat mat bat oloat at-yurp”

---

Translate the sentence
“clients do not sell pharmaceuticals in Europe”

<table>
<thead>
<tr>
<th>1a. García and associates .</th>
<th>7a. the clients and the associates are enemies .</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b. García y asociados .</td>
<td>7b. los clientes y los asociados son enemigos .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2a. Carlos García has three associates .</th>
<th>8a. the company has three groups .</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b. Carlos García tiene tres asociados .</td>
<td>8b. la empresa tiene tres grupos .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3a. his associates are not strong .</th>
<th>9a. its groups are in Europe .</th>
</tr>
</thead>
<tbody>
<tr>
<td>3b. sus asociados no son fuertes .</td>
<td>9b. sus grupos están en Europa .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4a. García has a company also .</th>
<th>10a. the modern groups sell strong pharmaceuticals .</th>
</tr>
</thead>
<tbody>
<tr>
<td>4b. García tambien tiene una empresa .</td>
<td>10b. los grupos modernos venden medicinas fuertes .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5a. its clients are angry .</th>
<th>11a. the groups do not sell zentanina .</th>
</tr>
</thead>
<tbody>
<tr>
<td>5b. sus clientes están enfadados .</td>
<td>11b. los grupos no venden zentanina .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6a. the associates are also angry .</th>
<th>12a. the small groups are not modern .</th>
</tr>
</thead>
<tbody>
<tr>
<td>6b. los asociados también están enfadados .</td>
<td>12b. los grupos pequeños no son modernos .</td>
</tr>
</tbody>
</table>
The Tetun-English Experiment
Best Human: 4.55 on a scale of 1 to 5.

Training: 1102 sentence pairs

Registering To Vote.
Register Ba Voto.

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 huto parte iha consulta popular ida mai ne'e no atu hau' o i imi ni an ba Timor-LoroSu'e ni futuru, imi teoke tan olik namo atu voto.

Registration has not started yet.
Awalamou sei disk hulu.

Sample Test Sentence:
Sr Ian Martin, Representante Especial Secretaria Jeral nian ba Timor Lorosa’s explika, “Ami ba fatin tolu ne’e tanha fatin hirak ne’e maka fatins’ia 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
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

Spanish/English Bilingual Text

Statistical Analysis

Spanish

Que hambre tengo yo

Broken English

What hunger have I,
Hungry I am so,
I am so hungry,
Have I that hunger ...

English

English Text

Statistical Analysis

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
- Resources
- 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
  - ygrybby snrv arqempnt tr arqempnt lx ieyity utbty c igawphwn sqxcmn
  - dgnsh kilig mjjabl dyurmect kilig cweorx sqs vaysgqaaq nlgmfc qisgfmg wxxd ofijhqmb
  - ouwovnfs xtc ttecv drg dvapphdh igav olcxd utbty c igawphwn jipmnhrt
  - kdxyc jeju wjuk thpxqoabl ljsp thpxqoabl jikcv esqd nlgmfc qisgfmg wxxd bsgdnavar
  - flbk rjekicrcj arqempnt advkmencl olcxd emqt fyrkfoem fm sqxcmn
  - tdlqinu idsvr hghdckel thpxqoabl wjrggy wiscuad es ld pyxt qisgfmg vliif ofijhqmb
- Test
  - xqmrchbg frkfoem xtc eucmpvbb
  - kph sgu pfstfj uarnq esaicocv vmpva sbxojtbb wpfcbbv rpjprpf yntrya yustw fyrkfoem cpnb
  - ixaity fanyi vmpva kskwdnx irnago yntrya kxlfww uakl vmpva aqjau dvapphdh lynswnk
Thank you!