Natural Language Processing: An Introduction
NLP: The Ultimate Goal (1990)

The *Ultimate* Goal – For computers to use NL as effectively as humans do….

“Natural language, whether spoken, written, or typed, is the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce. As computers play a larger role in the preparation, acquisition, transmission, monitoring, storage, analysis, and transformation of information, endowing them with the ability to understand and generate information expressed in natural languages becomes more and more necessary.”
NLP: Grand Challenges (1990)

The *Ultimate* Goal – For computers to use NL as effectively as humans do....

Reading and writing text
- Abstracting
- Monitoring
- Extraction into Databases

Interactive Dialogue: Natural, effective access to computer systems
- Informal Speech Input and Output

Translation: Input and Output in Multiple Languages
Recent Significant Advances In NLP

- Web-scale information extraction & question answering
  - IBM’s Watson

- Interactive Dialogue Systems
  - Apple’s Siri
Recent Significant Advances In NLP

Automatic Machine Translation

Xinhua story (Chinese) → Google translate

The Hague, March 24 Xinhua (Reporter Chen Zhi Pan Governance) Third Nuclear Security Summit held in The Hague on the 24th. State President Xi Jinping attended and delivered an important speech on China's nuclear safety measures and achievements, elaborated China on the development and security of both the rights and obligations of both, both independently and in collaboration, both temporary and permanent nuclear security concept, called on the international community to work together nuclear achieve lasting security and development.
The Past: Human Machine Interfaces

- **SHRDLU (Winograd, 1969)**
  - A fragile demonstration of the fundamental vision

- **LUNAR (Woods, Webber, Kaplan 1971)**
  - Answering geologist’s questions about the Apollo 11 moon rocks

- **PUNDIT (Hirschman, Palmer, … 1989)**
  - Information Extraction from Real World texts
SHRDLU: A demonstration proof

Person: *PICK UP A BIG RED BLOCK.*
Computer: OK. (does it)
Person: *GRASP THE PYRAMID.*
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person: *FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.*
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK. (does it)
Person: *WHAT DOES THE BOX CONTAIN?*
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: *WHAT IS THE PYRAMID SUPPORTED BY?*
Computer: THE BOX.
SHRDLU: A demonstration proof
LUNAR – William Woods 1971

- NLP interface to database of analyses of Apollo 11 moon rocks

- Examples
  - What is the average concentration of aluminum in high alkali rocks?
  - How many breccias contain olivine?
  - Give me the modal analyses of those samples for all phases.

- Handled 78% of sentences typed by geologists at 1971 Lunar Rocks conference
  - (12% more with “minor fixes”)
The Past: Information Extraction

PUNDIT (Palmer et al. 1987)

Sample CASREP & automatic summary

FAILURE OF ONE OF TWO SACS. UNIT HAD LOW OUTPUT AIR PRESSURE. RESULTED IN SLOW GAS TURBINE START. TROUBLESHOOTING REVEALED NORMAL SAC LUBE OIL PRESSURE AND TEMPERATURE. EROSION OF IMPELLOR BLADE TIP EVIDENT. CAUSE OF EROSION OF IMPELLOR BLADE UNDETERMINED. NEW SAC RECEIVED.

<table>
<thead>
<tr>
<th>Status of Sac:</th>
<th>Part: sac</th>
<th>State: inoperative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding:</td>
<td>Part: air pressure</td>
<td>State: low</td>
</tr>
<tr>
<td>Finding:</td>
<td>Part: lube oil pressure</td>
<td>State: normal</td>
</tr>
<tr>
<td>Finding:</td>
<td>Part: lube oil temperature</td>
<td>State: normal</td>
</tr>
<tr>
<td>Damage:</td>
<td>Part: blade tip</td>
<td>State: eroded</td>
</tr>
<tr>
<td>Finding:</td>
<td>Agent: ship’s force</td>
<td>State: has new sac</td>
</tr>
</tbody>
</table>
The Past: Crucial flaws in the paradigm

These systems worked well, BUT

1. Usually, only for a small set of examples
2. Person-years of work to port to new applications and, often, to extend coverage on a single application
3. Very limited and inconsistent coverage of English
Interactive systems often worked well...

...because of a magical fact:

*People automatically adapt and limit their language given a small set of exemplars if the underlying linguistic generalizations are HABITABLE*

This won’t handle non-interactive language
The State of NLP

NLP Past:
- Rich Representations

NLP Present:
- Powerful Statistical Disambiguation
An Early Robust Statistical NLP Application

• A Statistical Model For Etymology (Church ’85)

• Determining etymology is crucial for text-to-speech

<table>
<thead>
<tr>
<th>Italian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>AldriGHetti</td>
<td>lauGH, siGH</td>
</tr>
<tr>
<td>IannuCCi</td>
<td>aCCept</td>
</tr>
<tr>
<td>ItaliAno</td>
<td>hAte</td>
</tr>
</tbody>
</table>
An Early Robust Statistical NLP Application

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angeletti</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Iannucci</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Italiano</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Lombardino</td>
<td>58%</td>
<td>Italian</td>
</tr>
<tr>
<td>Asahara</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Fujimaki</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Umeda</td>
<td>96%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Anagnostopoulos</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Demetriadis</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Dukakis</td>
<td>99%</td>
<td>Russian</td>
</tr>
<tr>
<td>Annette</td>
<td>75%</td>
<td>French</td>
</tr>
<tr>
<td>Deneuve</td>
<td>54%</td>
<td>French</td>
</tr>
<tr>
<td>Baguenard</td>
<td>54%</td>
<td>Middle French</td>
</tr>
</tbody>
</table>

• Etymology can be determined with useful accuracy using a Naïve Bayes model based on letter sequences trigrams!
Information Extraction & Named Entity Recognition
Information Extraction

- Information extraction is the identification, in text, of specified classes of
  - names / entities
  - relations
  - events
- For relations and events, this includes finding the participants and modifiers (date, time, location, etc.).
- In other words, we build a data base with the information on a given relation or event:
  - people’s jobs
  - people’s whereabouts
  - merger and acquisition activity
  - disease outbreaks
  - genomics relation
Extraction Example

George Garrick, 40 years old, president of the London-based European Information Services Inc., was appointed chief executive officer of Nielsen Marketing Research, USA.

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>European Information Services, Inc.</td>
<td>London</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
<tr>
<td>CEO</td>
<td>Nielsen Marketing Research</td>
<td>USA</td>
<td>George Garrick</td>
<td>In</td>
</tr>
</tbody>
</table>
Named Entities

The who, where, when & how much in a sentence

The task: *identify atomic elements of information in text*

- Person names
- Company /organization names
- Locations (GPEs)
- Dates & times
- Percentages
- Monetary amounts
Won‘t simple lists solve the problem?

- too numerous to include in dictionaries
- changing constantly
- appear in many variant forms
- subsequent occurrences might be abbreviated

⇒ list search/matching doesn‘t perform well
Applications

- Information Extraction
- Summary generation
- Machine Translation
- Document organization/classification
- Automatic indexing of books
- .....
Yugoslav President Slobodan Milosevic received on Thursday the representatives of the Association of Yugoslav Banks, headed by its president Milos Milosavljevic, who is also the general director of JugoBanka.
Information Extraction from Propositions

Propositions are normalized connections from the parse trees. Entities and relations are extracted statistically from propositions.

Person: Slobodan Milosevic
Position: president
Organization: Yugoslavia

Person: Milos Milosevic
Position: president
Organization: Association of Yugoslav Banks

Person: Milos Milosevic
Position: general director
Organization: JugoBanka

CIS 391 - Intro to AI
Text Summarization

(For more on this topic, check out courses taught by Prof. Ani Nenkova)
MILAN, Italy, April 18. A small airplane crashed into a government building in heart of Milan, setting the top floors on fire, Italian police reported. There were no immediate reports on casualties as rescue workers attempted to clear the area in the city's financial district. Few details were available about it immediately set off fears that it might be a terrorist act akin to the Sept. 11 attacks in the United States. Those fears sent U.S. stocks tumbling to session lows in late morning trading.

Witnesses reported hearing a loud explosion from the 30-story office building, which houses the administrative offices of the local Lombardy region and sits next to the city's central train station. Italian state television said the crash put a hole in the 25th floor of the Pirelli building. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. No further details were immediately available.
CBC News: Canadian hostages relax after long ordeal in Iraq

Summary from multiple countries, from articles in English

The 74-year-old peace activist from Pinner, northwest London, who was kidnapped four months ago, had been rescued with two Canadian hostages by an international team of special forces soldiers led by SAS men. (article 9) Harmmeet Singh Sooden celebrated his 33rd birthday in Baghdad on Friday, one day after an international force rescued the Canadian man and two other hostages following four months of captivity in Iraq. (article 21) Freed UK hostage. Full statement var clickExpire * 04/8/2006 LONDON, England (Reuters) Freed peace campaigner Norman Kember flew home to Britain on Saturday following his rescue by special forces soldiers after being held hostage in Iraq for nearly four months. (article 16) The operation was conducted after coalition forces detained two people the night before, Lynch said (Watch what led to the rescue operation 2:41) One of the detainees knew where the hostages were. (article 2) Kember and his two Canadian colleagues owe their freedom to a rift among their Iraqi kidnappers, a western security source close to the rescue operation said yesterday. (article 11) Loved ones of the freed Kember have joined politicians in speaking of their joy at the release of the 74-year-old peace activist who had been held in Iraq for nearly four months. (article 14)

Other summaries about this story:
- Summary from the United Kingdom, from articles in English (18 articles) [compare]
- Summary from United States, from articles in English (7 articles) [compare]
- Summary from Canada, from articles in English (2 articles) [compare]

Event tracking:
- Track this story's development in time

Story keywords
Kember, IRAQ, Norman, hostages, Baghdad

Source articles
1. Freed UK hostage thanks rescuers amid criticism (CNN, 03/25/2006, 859 words)
2. Troops act fast on intelligence to free hostages (CNN, 03/24/2006, 997 words)

(Includes work by Prof. Nenkova)
Machine Translation

(Several slides from Language Weaver)
Why use computers in translation?

- Too much translation for humans
- Technical materials too boring for humans
- Greater consistency required
- Need results more quickly
- Not everything needs to be top quality
- Reduce costs

- Any one of these may justify machine translation or computer aids
Statistical Machine Translation Technology

Spanish/English Bilingual Text

Statistical Analysis

English Text

Statistical Analysis

Spanish

Broken English

English

Qué hambre tengo yo

What hunger have I,

Hungry I am so,

I am so hungry,

Have I that hunger ...

I am so hungry
How A Statistical MT System Learns

**Sources of Bilingual Parallel Data**
- Translation memories
- Translated Archives
- Dictionaries/glossaries
- Internet
- Human Translations

**Pre-processing**
- Format Filtering
- Transcription
- Document Alignment
- Segment Alignment

**Parallel Corpus**
- ¿Hay mejores formas de hacerlo?
- Viajaron alrededor del mundo.
- Estos son los mejores libros para leer.
- Are there better ways of doing it?
- They traveled around the world.
- These are the best books to read.

**Translation Parameters**
(Bilingual Data)
- Mejores: 19%
- Best: 10%
- Appropriate: 7.3%
- Better: 2.3%
- Ourperform: 2.2%
- Finest: 2%

- Mundo: 4.7%
- World: 4.7%
- Globe: 3.7%
- Globalizing: 3.1%
- Worlds: 2.9%

**Language Model**
(Monolingual Data)
- The best companies in the world: 27%
- Best Companies in the globe: 17%
- World’s best companies: 11%
- Better companies in the world: 5.1%
- The world companies are best: 4.8%
Translating a New Document

ORIGINAL DOCUMENT → LANGUAGE WEAVER DECODER → TRANSLATED DOCUMENT

STATISTICAL ANALYSIS

TRANSLATION PARAMETERS

LANGUAGE MODEL
Broadcast Monitoring
BBN MAPS & Language Weaver MT
In Point

Out Point

CIS 391 - Intro to AI

Social conditions for Palestine refugees in Lebanon and began the 10,000 license source of foreign workers.

A number of permits for the work, according to figures that between the Lebanese about twenty-eight percent, according to these two million workers ranging between five and fifty-five percent of the population of Lebanon and only about ten percent of the Palestinian population.

The rate of unemployment among non-officials is between five and sixty-six percent. The rate was not exceeding the duration of two years two million workers out of some quarter of a million people.

Statistics that between the last two years budgeting, a number of permits for the work, according to figures that between the Lebanese about twenty-eight percent, according to these two million workers ranging between five and fifty-five percent of the population of Lebanon and only about ten percent of the Palestinian population.

The rate of unemployment among non-officials is between five and sixty-six percent. The rate was not exceeding the duration of two years two million workers out of some quarter of a million people.

The rate of unemployment among non-officials is between five and sixty-six percent. The rate was not exceeding the duration of two years two million workers out of some quarter of a million people.

The rate of unemployment among non-officials is between five and sixty-six percent. The rate was not exceeding the duration of two years two million workers out of some quarter of a million people.
Language Weaver  Hybrid Translation Technology

- Chinese Source Text
  Sample 1:
  
  车展，一向是衡量一个国家汽车消费现状和市场潜力的“晴雨表”。本届北京国际车展有24个国家的1200余家厂商参展，8天接待40余万名参观者，创下了中国车展的新纪录，让人深切地感受到汽车市场启动的信号。

  "中国是世界最后一个最大的汽车市场”。多年来，这句话更多地包含着汽车商人的一种希翼。然而如今，越来越多的事实预示着它正在变为现实。

  来自本届车展的一组数据很有说服力。《北京青年报》的一份现场调查显示，35岁以下参观者约占35%；62.1%的被访者表示，参观车展主要是为近期买车搜集信息，甚至在展会上就有可能购买或预订合适的产品；76%的被访者表示最近两年会购买私家车。

  今年以来，国内轿车市场的强劲增长让厂家喜上眉梢。据国家统计局公布的数字，在前4个月，全国共生产轿车26.79万辆，增长27.6%；特别是4月份，生产轿车9万辆，同比增长50.5%，创造了十几年来轿车月产增长的最高纪录。从销售看，一季度，全国轿车生产企业共销售轿车18.8万辆，产销率达105%；轿车库存比年初下降1.1万辆，下降幅度近25%。

Language Weaver Experimental Syntax MT
Sample 1:

The motor show, has always been the' barometer' of a national car consumption and market potential. The Beijing International Auto Show has more than 1,200 exhibitors from 24 countries and 8 days of receiving more than 40 million visitors, setting a new record in China's auto show, are deeply aware of the automobile market signals.

"China is one of the largest automobile market in the world. Over the years, this phrase implies more auto businessmen. But now, more and more facts indicates that it is to become a reality.

Data from the Motor Show is very convincing. The Beijing Qingnian Bao Report on-the-spot investigation showed that about 35 percent of 35-year-old visitors, 62.1 percent of the respondents said that the truck was mainly to buy a car in the near future to collect information, even at the exhibition may purchase or suitable products; 76% of respondents indicated in the past two years to buy private cars.

Since the beginning of this year, the strong growth of the domestic car market. According to the figures released by the National Bureau of Statistics, in the first four months, the country produced 267,900 vehicles, up 27.6 percent; in particular, in April, the production of 90,000 vehicles, an increase of 50.5% over the same period last year, setting a record high for the monthly output growth over the past 10-odd years. In terms of sales in the first quarter, manufacturing enterprises in the country sold 188,000 cars, up 22 percent over the same period of last year, up 10.5 percent; 11,000 vehicles, dropping by nearly 25 percent lower than the beginning of the year.
MT Challenges: Ambiguity

- **Syntactic Ambiguity**
  I saw the man on the hill with the telescope

- **Lexical Ambiguity**
  - E: book
  - S: libro, reservar

- **Semantic Ambiguity**
  - Homography:
    ball(E) = pelota, baile(S)
  - Polysemy:
    kill(E), matar, acabar (S)
  - Semantic granularity
    esperar(S) = wait, expect, hope (E)
    be(E) = ser, estar(S)
    fish(E) = pez, pescado(S)