
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

Significant Advances In NLP I

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



- Interactive Dialogue Systems
 - Apple's Siri
 - (Microsoft Cortana)
 - (Amazon Echo)
 - (Google Now)



Significant Advances In NLP II

Automatic Machine Translation

Xinhua story (Chinese) → Google translate (2015)

新华网海牙 3 月 2 4 日电 (记者陈贇潘治) 第三届核安全峰会 2 4 日在荷兰海牙举行。国家主席习近平出席并发表重要讲话，介绍中国核安全措施和成就，阐述中国关于发展和安全并重、权利和义务并重、自主和协作并重、治标和治本并重的核安全观，呼吁国际社会携手合作，实现核能持久安全和发展。

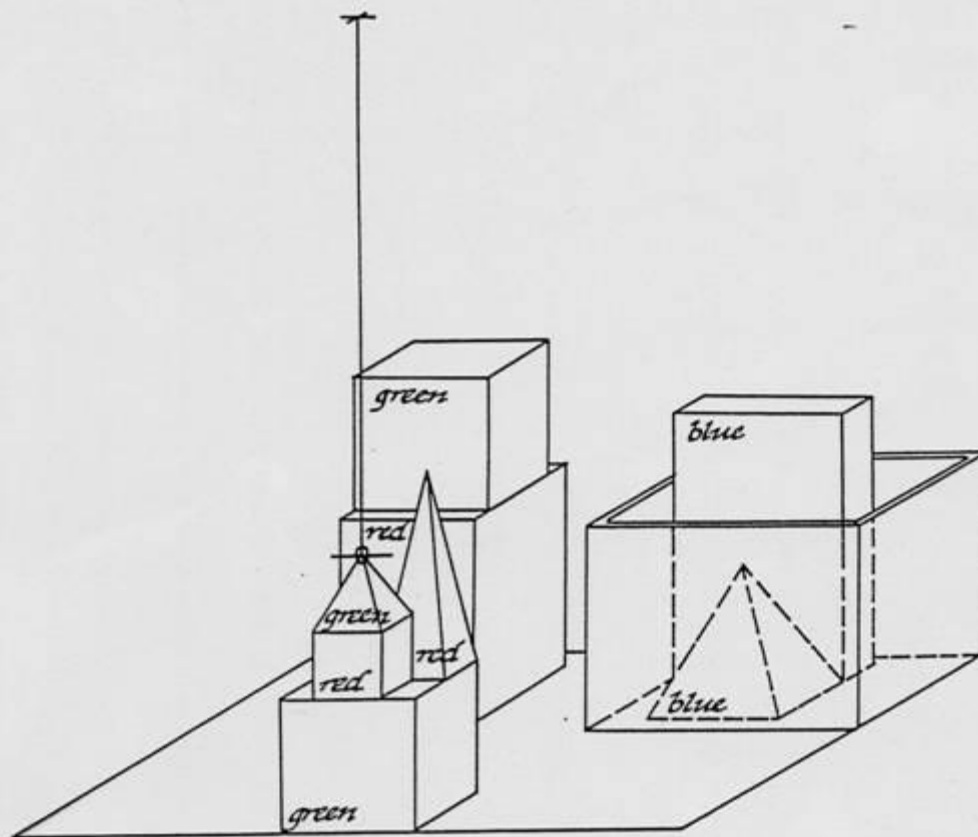
The Hague, *March 24* Xinhua (Reporter Chen Zhi Pan Governance) Third Nuclear Security Summit held in The Hague on the 24th . State President Xi Jinping attended and delivered an important speech on China's nuclear safety measures and achievements , elaborated China on the development and security of 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.

Early Successes: 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

Review: SHRDLU: A demonstration proof

Does the shortest thing the tallest pyramid's support supports anything green?



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**
 - (90% after “minor fixes”)

The Past: Crucial flaws in the paradigm

These and other later systems worked well, BUT

1. Person-years of work to port to new applications
2. Very limited coverage of English

Crucially, they 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 pre-existing text!

The State of NLP

NLP Past before 1995:

- Rich Representations

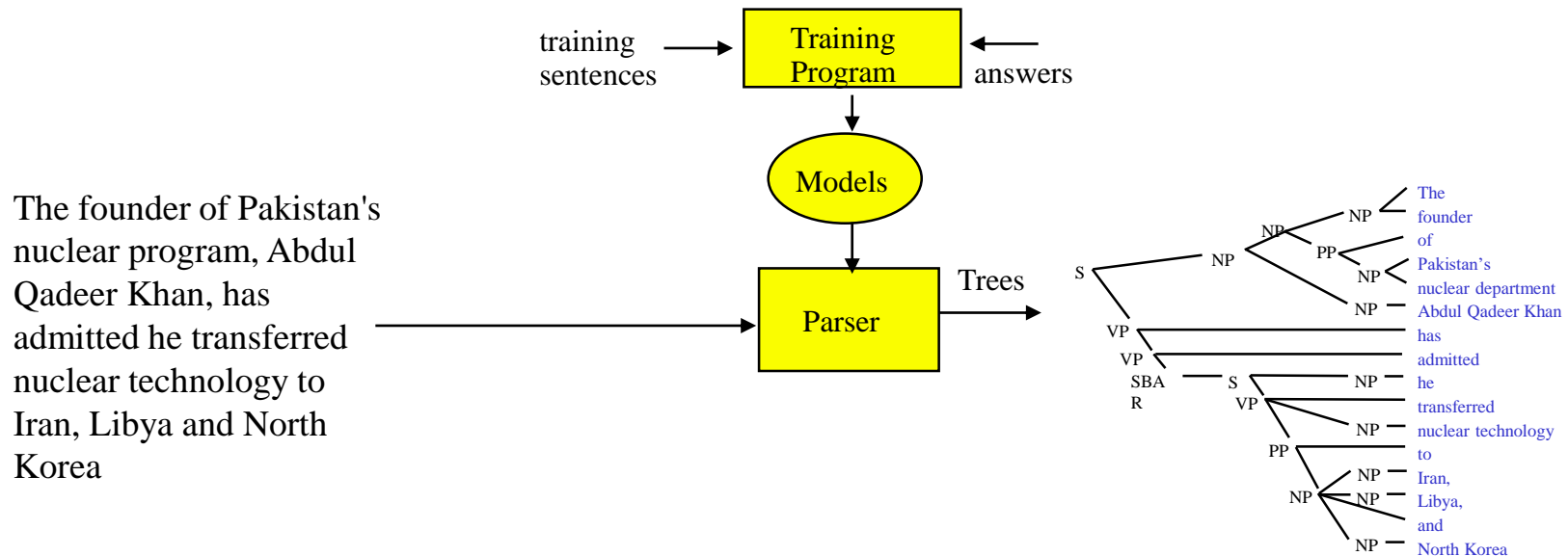
NLP Present:

- *Powerful Statistical Disambiguation*

1995: A breakthrough in parsing

10⁶ words of Treebank Annotation
+ Machine Learning = Robust Parsers

(Magerman '95)

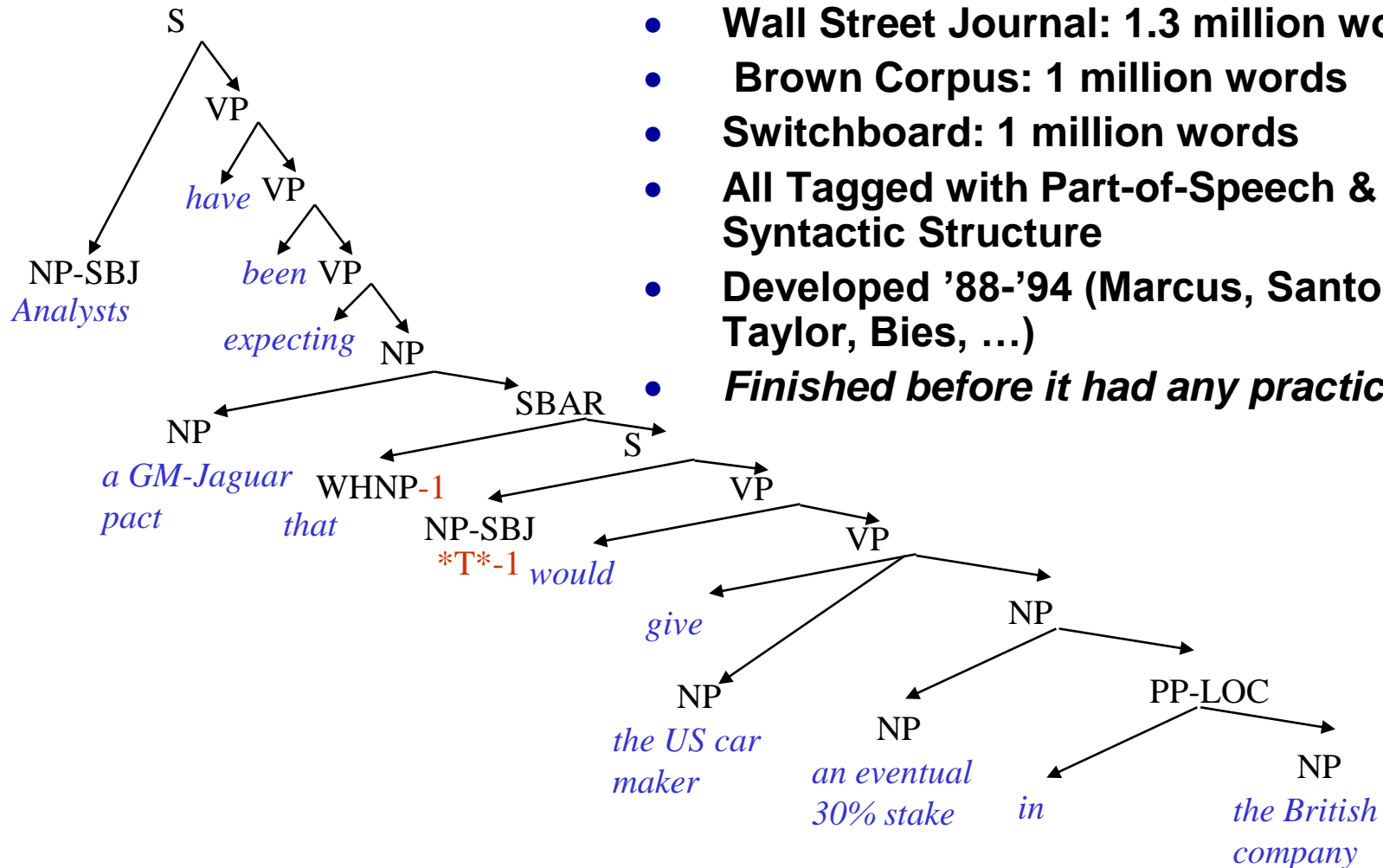


- 1990 Best hand-built parsers: ~40-60% accuracy (guess)

- 1995+ Statistical parsers: >90% accuracy

(both on short sentences)

The Penn Treebank: 1988-94



- Wall Street Journal: 1.3 million words
- Brown Corpus: 1 million words
- Switchboard: 1 million words
- All Tagged with Part-of-Speech & Syntactic Structure
- Developed '88-'94 (Marcus, Santorini, Taylor, Bies, ...)
- ***Finished before it had any practical use!***

Lexicalized parsing results

(Labeled Constituent Precision/Recall F1)

Results

Method	Accuracy
PCFGs (Charniak 97)	73.0%
Conditional Models – Decision Trees (Magerman 95)	84.2%
Lexical Dependencies (Collins 96)	85.5%
Conditional Models – Logistic (Ratnaparkhi 97)	86.9%
Generative Lexicalized Model (Charniak 97)	86.7%
Generative Lexicalized Model (Collins 97)	88.2%
Logistic-inspired Model (Charniak 99)	89.6%
Boosting (Collins 2000)	89.8%

(Chris Manning, Stanford)

A Few Core Technologies

1. **Named Entity Recognition & Information Extraction**
2. **Machine Translation**
3. **Text Summarization**

Information Extraction & Named Entity Recognition

Information Extraction

- **Information extraction is the identification, in text, of specified classes of Named Entities +**
 - Relations
 - Events
- **For relations and events, this includes finding the participants and modifiers (date, time, location, etc.).**
- **Goal: fill out a data base with given relation or event types:**
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.

Position	Company	Location	Person	Status
President	European Information Services, Inc.	London	George Garrick	Out
CEO	Nielsen Marketing Research	USA	George Garrick	In

Named Entity Recognition

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

- Flag the who, where, when & how much in text
- **Person names**
- **Company /organization names**
- **Locations**
- **Dates & times**
- **Percentages**
- **Monetary amounts**

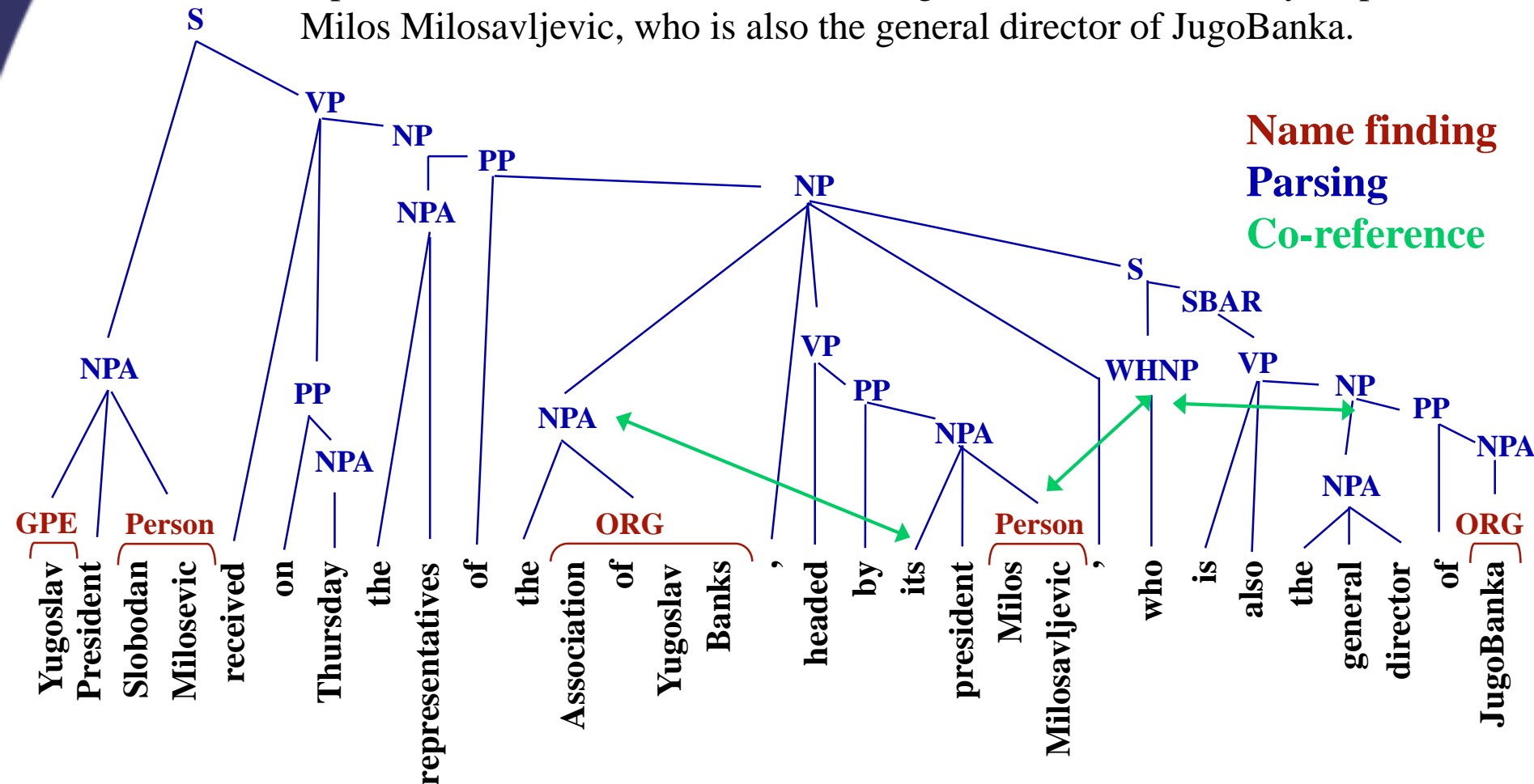
Won't simple lists solve the problem?

- 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

Levels of BBN Statistical Analysis (2005)

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.



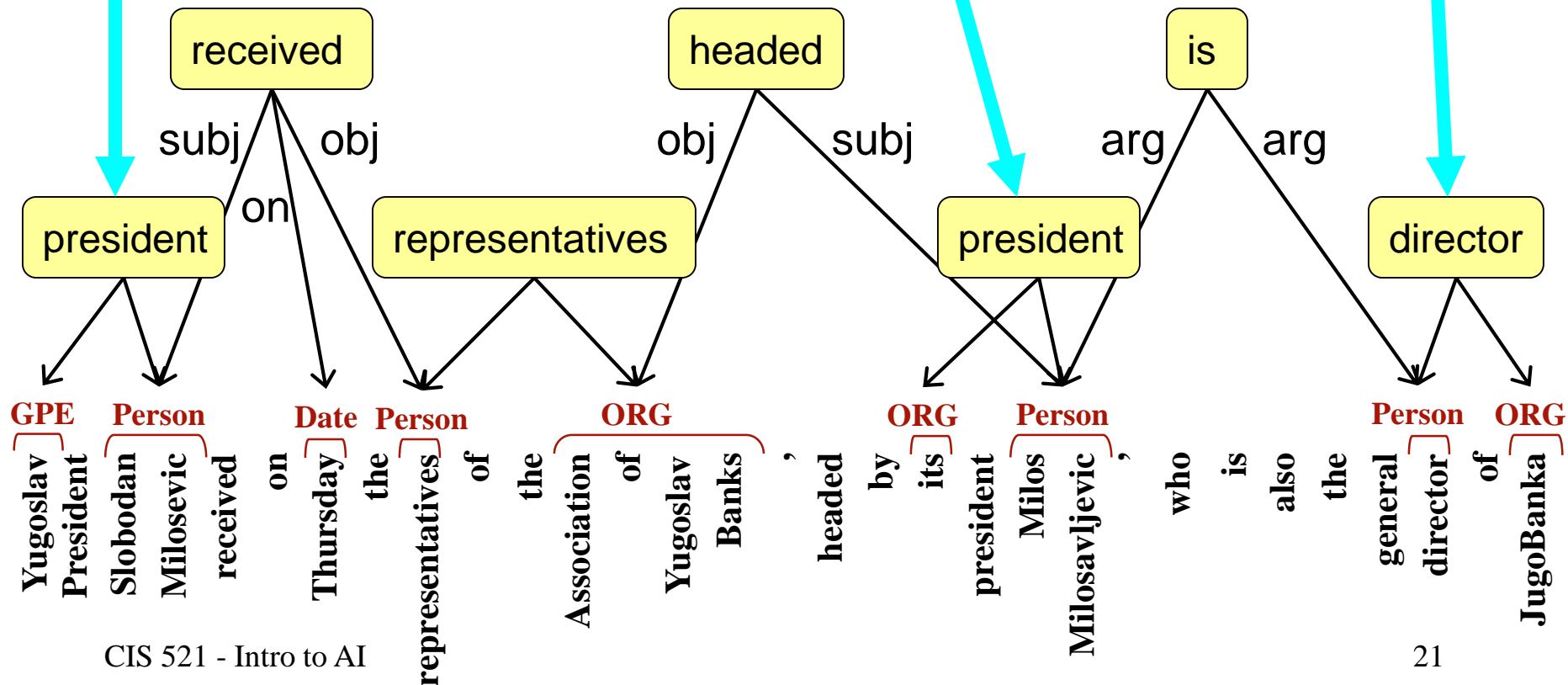
Information Extraction from Propositions

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

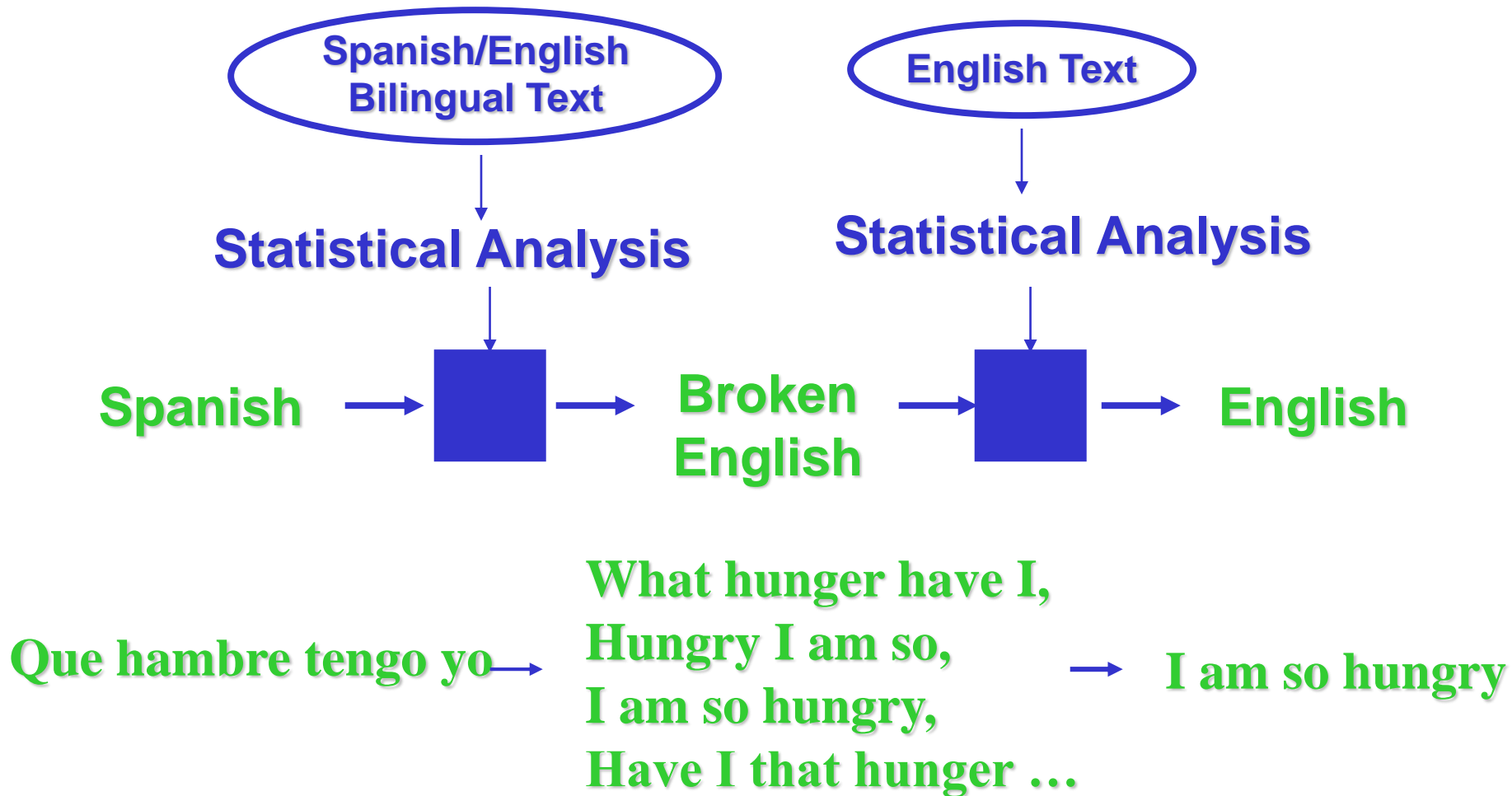


Statistical Machine Translation

(For more on this topic, check out courses taught by Prof. Chris Callison-Burch)

(Next several slides from Language Weaver)

Statistical Machine Translation Technology



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?	Are there better ways of doing it?
Viajaron alrededor del mundo.	They traveled around the world
Éstos son los mejores libros para leer.	These are the best books to read

TRANSLATION PARAMETERS

(Bilingual Data)

Las mejores compañías del mundo cuentan con traducciones de idioma para comunicarse con mercados mundiales.

Mejores	Best	19%
	Appropriate	7.3%
	Better	2.3%
	Ourperform	2.2%
	Finest	2%
Mundo	World	4.7%
	Globe	3.7%
	Globalizing	3.1%
	Worlds	2.5%

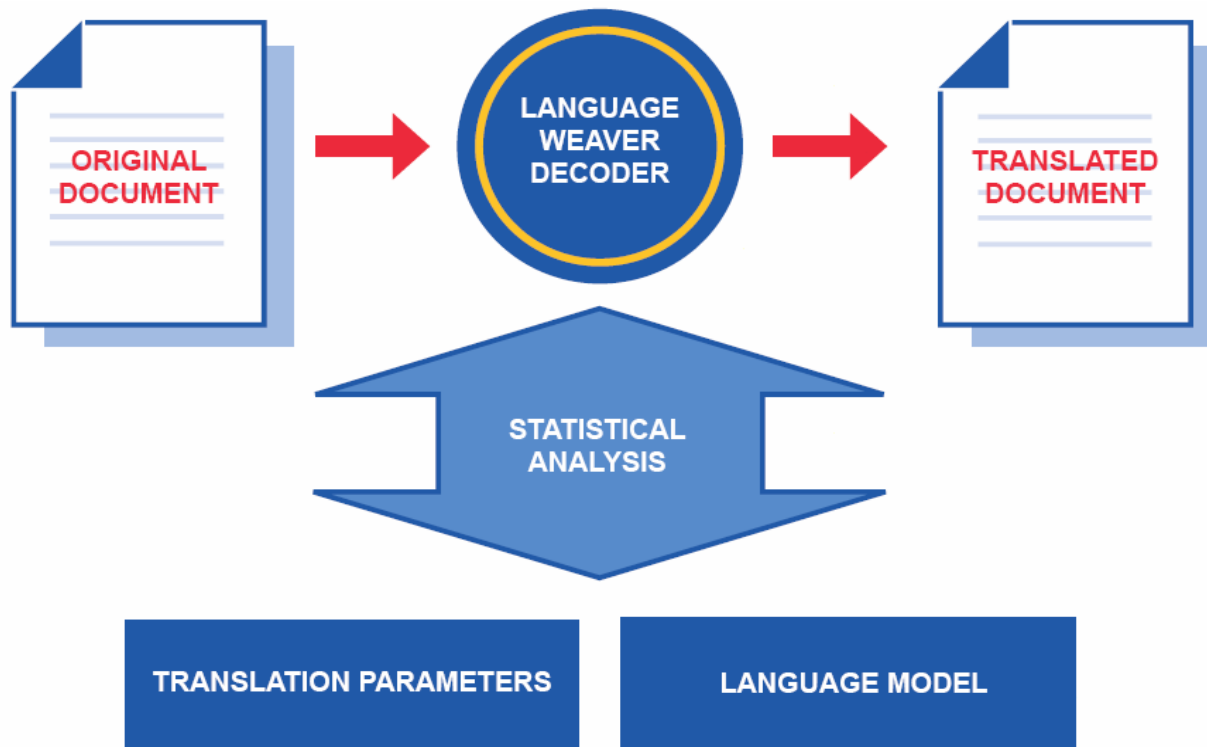


LANGUAGE MODEL

(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



Broadcast Monitoring

BBN MAPS & Language Weaver MT

BBN Broadcast Monitoring System - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address http://vms.bbn.com/web_ui/app/root_menu.aspx Go

Search: Go

Arabic Keyboard since midnight in English all channels Results Overview Intro



Channel: CCTV
Time: 28 Jun 05 15:33:11 GMT



Channel: Al Arabiya
Time: 28 Jun 05 15:33:45 GMT



Channel: Al Jazeera
Time: 28 Jun 05 15:34:17 GMT

size of economic activity is expected to be.

الاقتصادي المنتظر.

ale 34

he Palestinians nearly 10 percent of the entire population of Lebanon and the estimated unemployment rates between the Lebanese about twenty-eight percent, according to figures from non-official, unemployment among refugees ranging between five and fifty and five sixty-six percent and refers to statistics that between the years two ninety thousand of the last century was not exceeding the number of permits for the work abber five Palestinians out of some 0,000 a licence source of foreign workers.

ويشكل الفلسطينيون قرابة العشرة في المئة من مجمل سكان لبنان وفيما تقدر معدلات البطالة بين اللبنانيين بحوالي ثمانية وعشرين في المئة وفق أرقام غير رسمية فإن البطالة بين اللاجئين تتراوح بين خمسة وخمسين وخمسة وستين في المئة وتشير الإحصاءات إلى أنه بين الأعوام اثنين وتسعين وألفين من القرن الماضي لم يتجاوز عدد رخص العمل المطاط الفلسطينيين الخمسة من أصل حوالي خمسين ألف ترخيص مصدر للعمال الأجانب.

ale 38

social conditions for Palestine refugees in Lebanon and began the

الأوضاع الاجتماعية للاجئين الفلسطينيين في لبنان. ما بدأ بشكا

In Point:

Not set



Out Point:

Not set

العربية

دنيا المال

الاقتصاد اللبناني

السماح لبعض الفلسطينيين بالعمل في مهن محصورة باللبنانيين

779 00

الشركة تفتح

العربية

الجنسيات

Paused

Play

Snapshot

Video

Channel:

Al Arabiya

Time:

28 Jun 05 16:31:49 GMT

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冀。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，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，同比增长22%，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.

Text Summarization

(For more on this topic, check out courses taught by Prof. Ani Nenkova)

Search for:

[U.S.](#)
[World](#)
[Finance](#)
[Sci/Tech](#)
[Entertainment](#)
[Sports](#)

[View Today's Images](#)

[View Archive](#)

[About Newsblaster](#)

[About today's run](#)

[Newsblaster in Press](#)

[Academic Papers](#)

Article Sources:

[washingtonpost.com](#)
(215 articles)
[news.bbc.co.uk](#)
(190 articles)
[baltimoresun.com](#)
(146 articles)
[boston.com](#)
(104 articles)
[seattletimes](#)
[nwsource.com](#)
(63 articles)
[timesonline.co.uk](#)
(258 articles)

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](#)) Harmeet 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, 293 words)



(Includes old work by Prof. Nenkova)