Ontology-Based Arabic Text Understanding

فهم النص العربى اعتماداً على الانطولوجيا

By

Ahmed Said

Computer Science Department

Agenda:

- What is understanding?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- **Objective**

Agenda:

- Motivation
- Problem definition
- Objective
- RELATED WORK
- References

Motivation

- The aim of this research is to study the role of ontologies and their contributions in the semantic representation phase.
- We hope that the generated semantic representation of meaning used by NLP systems for other multilingual applications such as cross lingual information retrieval, summarization, machine translation, and question answering.
- We expect that this approach redirect research to modern semantic construction technologies for developing an adequate model of semantic processing for Arabic.

Problem Definition

- There remains a need for an efficient model that can represent Arabic meaning from Arabic text
- Semantic processing of human language is a problematic issue of natural language processing.
- In this study we present semantic representation that reflects the meaning for Arabic sentences using ontology.

Objectives

The main goals of this thesis a to:

- To develop a model for semantic representation for Arabic text using Arabic ontology
- To establish Arabic concepts senses, as they should matched the correct meaning of the target word in the handled sentence.
- To improve the semantic representation of a sentence's information content by considering ontology attributes.
- To increase of machine understanding for Arabic text

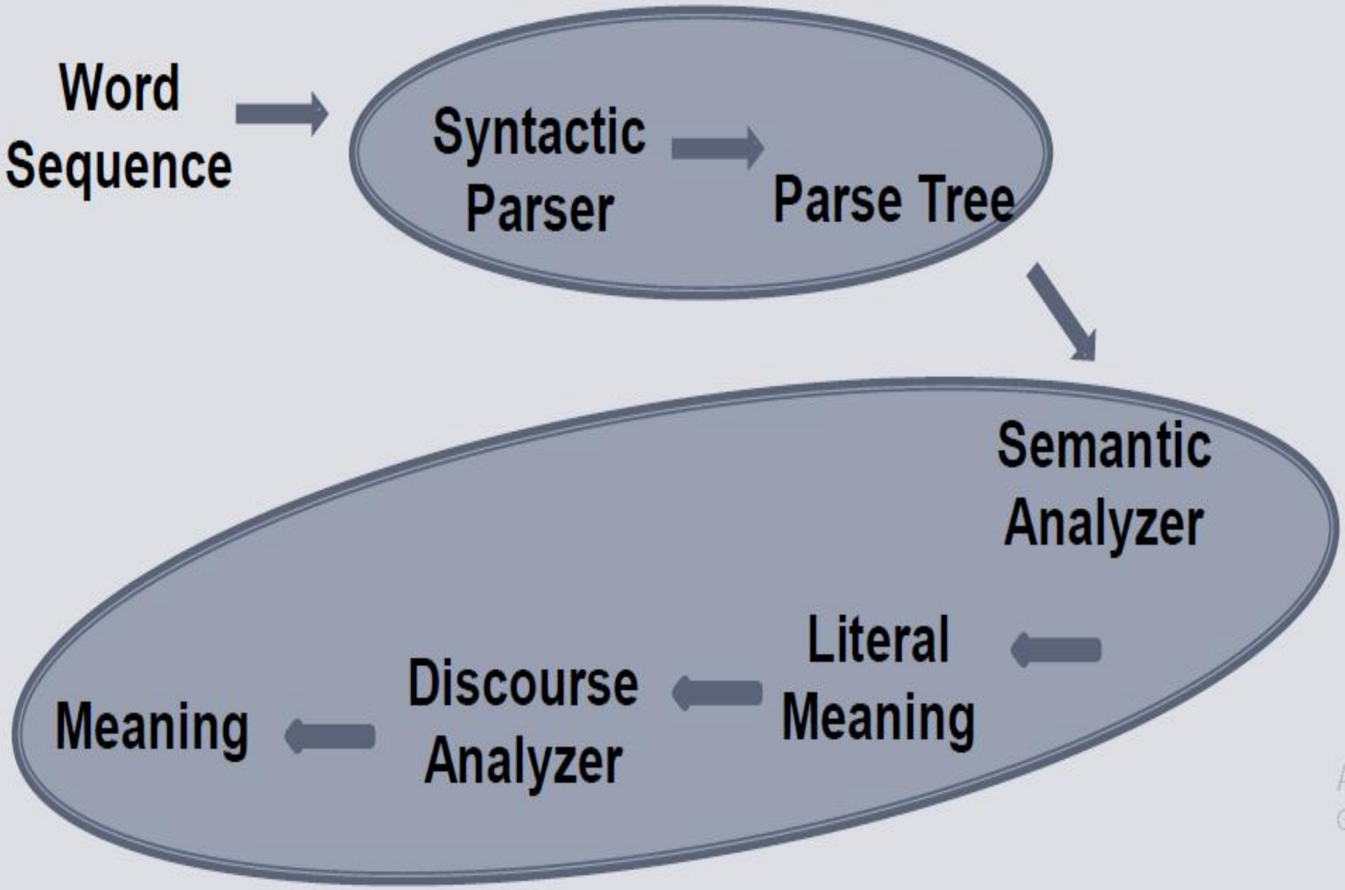
Agenda:

- What is understanding?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

What is Understanding?

- Computer understanding is a challenge problem in Artificial Intelligence
- Natural language processing needs an understanding system to make the machine understand human languages.
- Understanding is a transformation from one representation to another.

Understanding Natural Language



What is Text Analysis for Semantic Computing?

- Finding more about what we already know
  - Ex. patterns that characterize known information
  - The search/browse OR *'finding a needle in a haystack'* paradigm

- Discovering what we did not know
  - Deriving new information from data
    - Ex. Relationships between known entities previously unknown
  - The *'extracting ore from rock'* paradigm

Levels of Text Analysis

- Information Extraction - those that operate directly on the text input
    - this includes entity, relationship and event detection
- Inferring new links and paths between key entities
    - sophisticated representations for information content, beyond the "bag-of-words" representations used by IR systems
- Scenario detection techniques
    - discover patterns of relationships between entities that signify some larger event, e.g. money laundering activities.

What do they all have in common?

- They all make use of knowledge of language (exploiting syntax and structure, different extents)
  - Named entities begin with capital letters
  - Morphology and meanings of words

- They all use some fundamental text analysis operations
  - Pre-processing, Parsing, chunking, part-of-speech, lemmatization, tokenization

- To some extent, they all deal with some language understanding challenges
  - Ambiguity, co-reference resolution, entity variations etc.

- Use of a core subset of theoretical models and algorithms
  - State machines, rule systems, probabilistic models, vector-space models, classifiers, EM etc.

What is NLP?

- Branch of AI
- 2 Goals
  - Science Goal: *Understand the way language operates*
  - Engineering Goal: *Build system that analyze and generate languages; reduce the man machine gap*
- Help in communication
  - With computers (ASR, TTS)
  - With other humans (MT)

WHAT IS UNDERSTANDING?

Awareness

Reading

Relating

Comprehending

Inference

Interpretation

Prediction

Creation

A CHANGE MUST COME

We can no longer cope with understanding unstructured data manually in a Big Data world.

We must tie technology that can scale horizontally to the function of understanding.

In short, Understanding must become *Automated*.

AUTOMATED UNDERSTANDING: IT'S ABOUT THE 80%

80%

- Awareness
- Reading
- Relating
- Comprehending

20%

- Inference
- Interpretation
- Prediction
- Creation

HOW DO YOU AUTOMATE UNDERSTANDING?

Inputs

Unstructured

Structured

Social

Multiple Languages

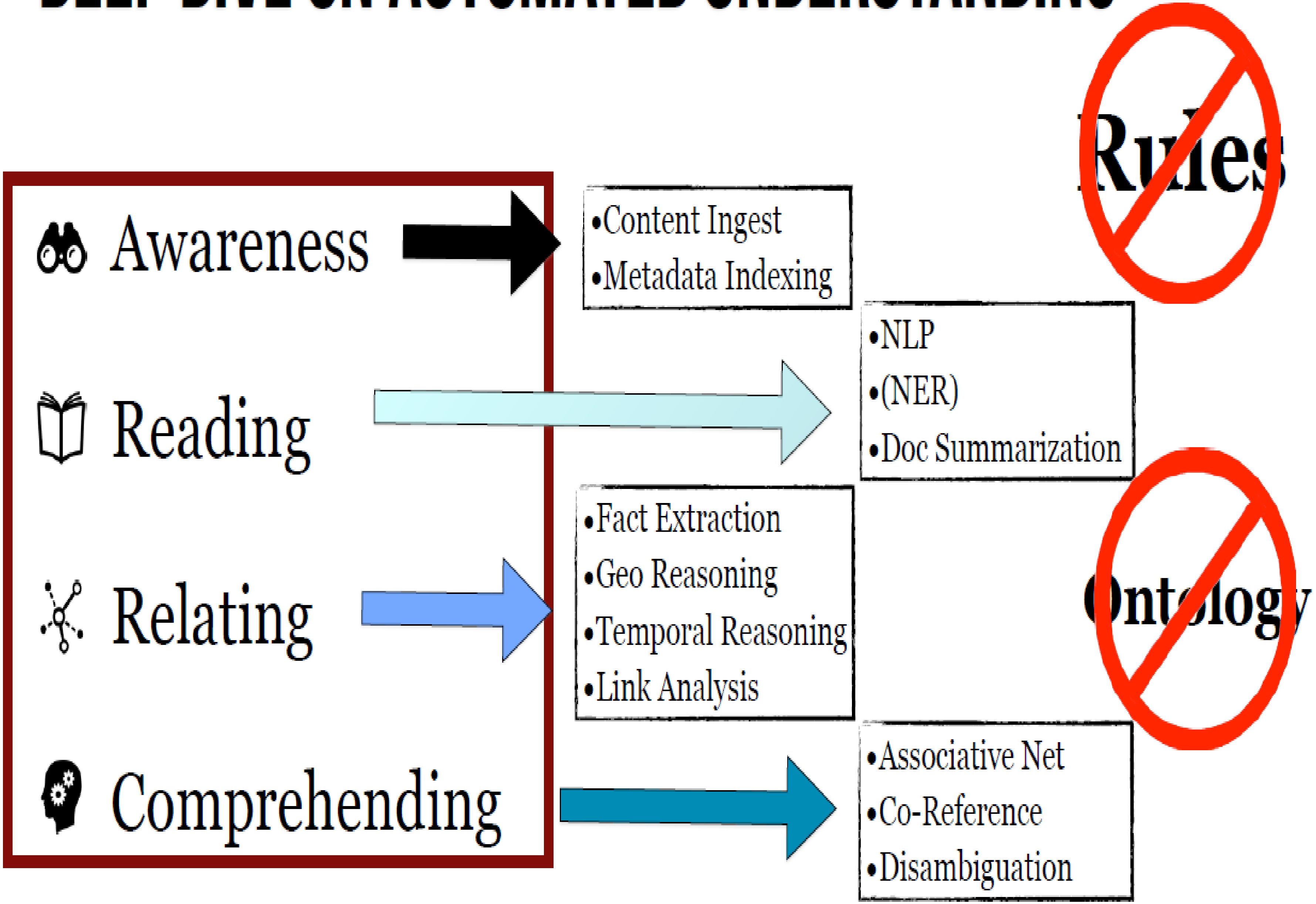
Integrated Functions

- Doc Summarization
- Associative Net
- Co-Reference
- Disambiguation
- Link Analysis
- Geo Reasoning
- Temporal Reasoning
- Fact Extraction
- NLP

Outputs

- People understood in space and time
- Connections/relationships and their contexts
- Data fusion from multiple sources & types
- Links back to the source document if needed

DEEP DIVE ON AUTOMATED UNDERSTANDING



Rules

Awareness

- Content Ingest
- Metadata Indexing

Reading

- NLP
- (NER)
- Doc Summarization

Relating

- Fact Extraction
- Geo Reasoning
- Temporal Reasoning
- Link Analysis

Ontology

Comprehending

- Associative Net
- Co-Reference
- Disambiguation

A CHANGE IS HERE NOW

Automated Understanding is the next wave of Analytics. It deals with *your data (vs. your machine's data)* and how *you* make decisions.

It's here now, but we've only scratched the surface of the value it can create

It gives us *hope* to reclaim our lives from the abuse of attention and the constant worry of uncertainty.

A CHANGE IS HERE NOW

Understanding **secures**

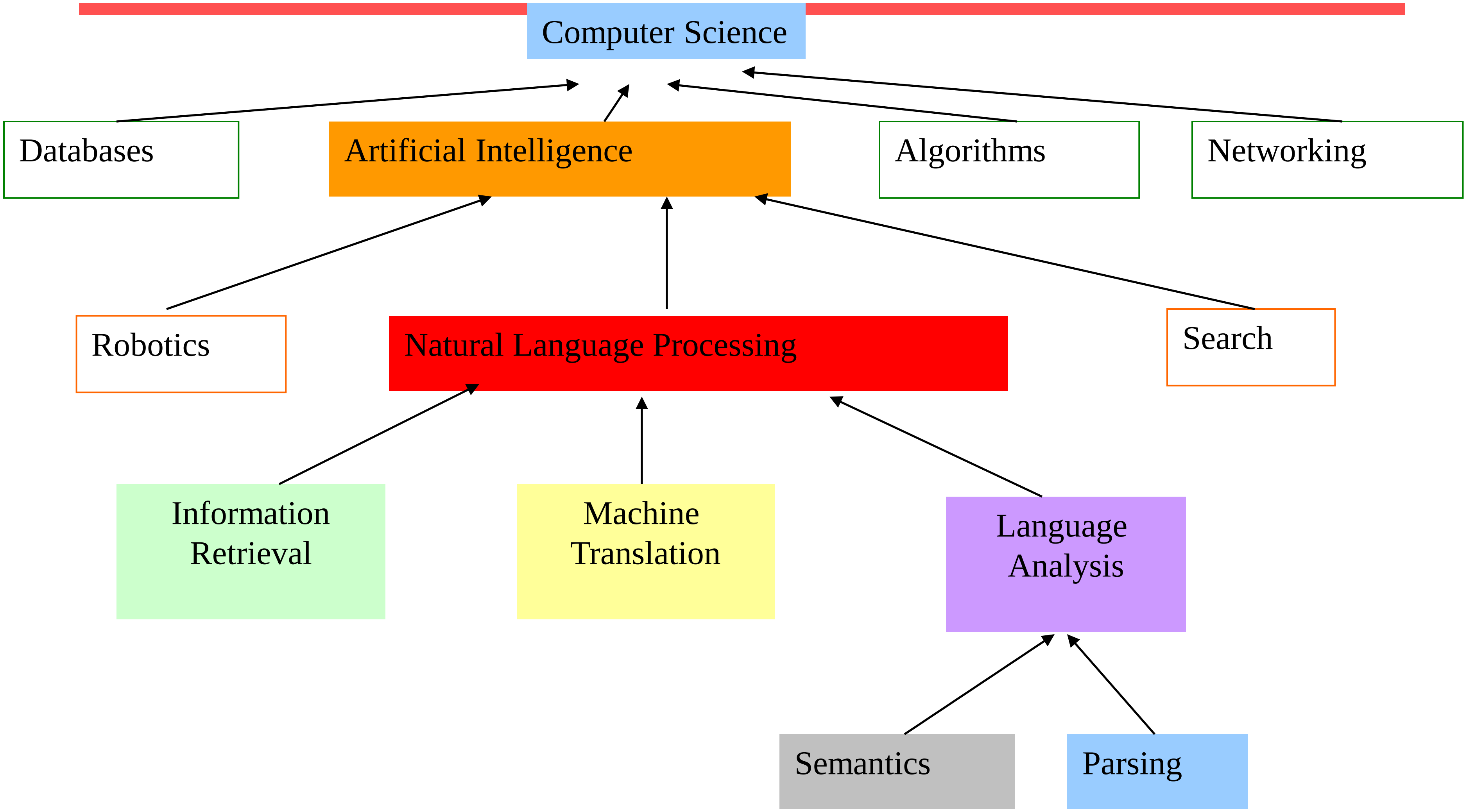
Understanding **empowers**

Understanding **creates Hope**

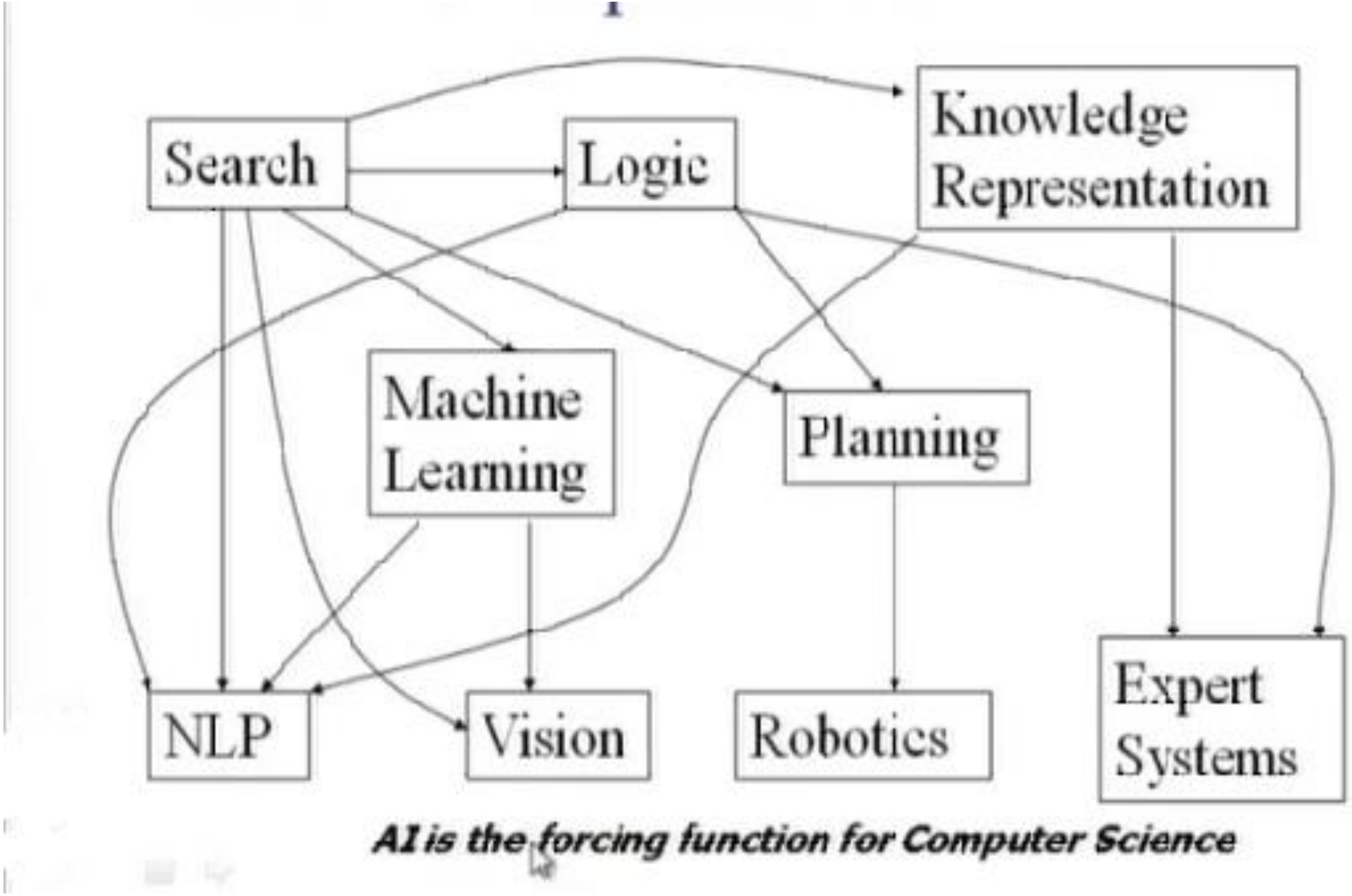
Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

Where does it fit in the CS taxonomy?



Areas of AI and their inter-dependencies



Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

Why NLP is important?

- Fundamental transition from the Industrial Economy to the Knowledge Economy
- Knowledge is coded in Language
- Huge amounts of data on the Internet, Intranets, desktops
- We need applications for processing (understanding, retrieving , translating, summarizing, …) this large amounts of texts.

NLP challenging

Ambiguity

This is what makes NLP challenging: The Crux of the problem

لماذا يكون من الصعب معالجة اللغة العربية؟

- اللغة العربية عادة ما تكتب مع التشكيل الاختياري
- ليس هناك مفهوم لحروف تدل علي أسماء الأعلام

بعض اشكال الغموض (1)

- اللفظة متجانسة
  - أسم / فعل: [كتب] ؛ [ذهب]
- تقسيم الكلمه
  - أسم / حرف+أسم : [بعقوبه] / [ب][عقوبه] ؛ [وجد] / [و+جد]
- معني الكلمة
  - يزور / يُزوِّر ؛ أرض / أرض
- نحوي
  - خطيب المسجد الجديد.   (من/ما هو الجديد؟)
  - أحب أمي أكثر من أبي. (من يحب من أكثر؟)

بعض اشكال الغموض (2)

- عطف:
  - رأيت الأبنية والكباري تحت التشييد. (كلاهما أم الكباري فقط؟)
- علاقات الأشاره للكلمات بضمائر
  - قابل الصحفي الوزير الذي انتقده. (من انتقد من؟)
- حذف الضمير
  - جاء الأستاذ وانصرف [هو].

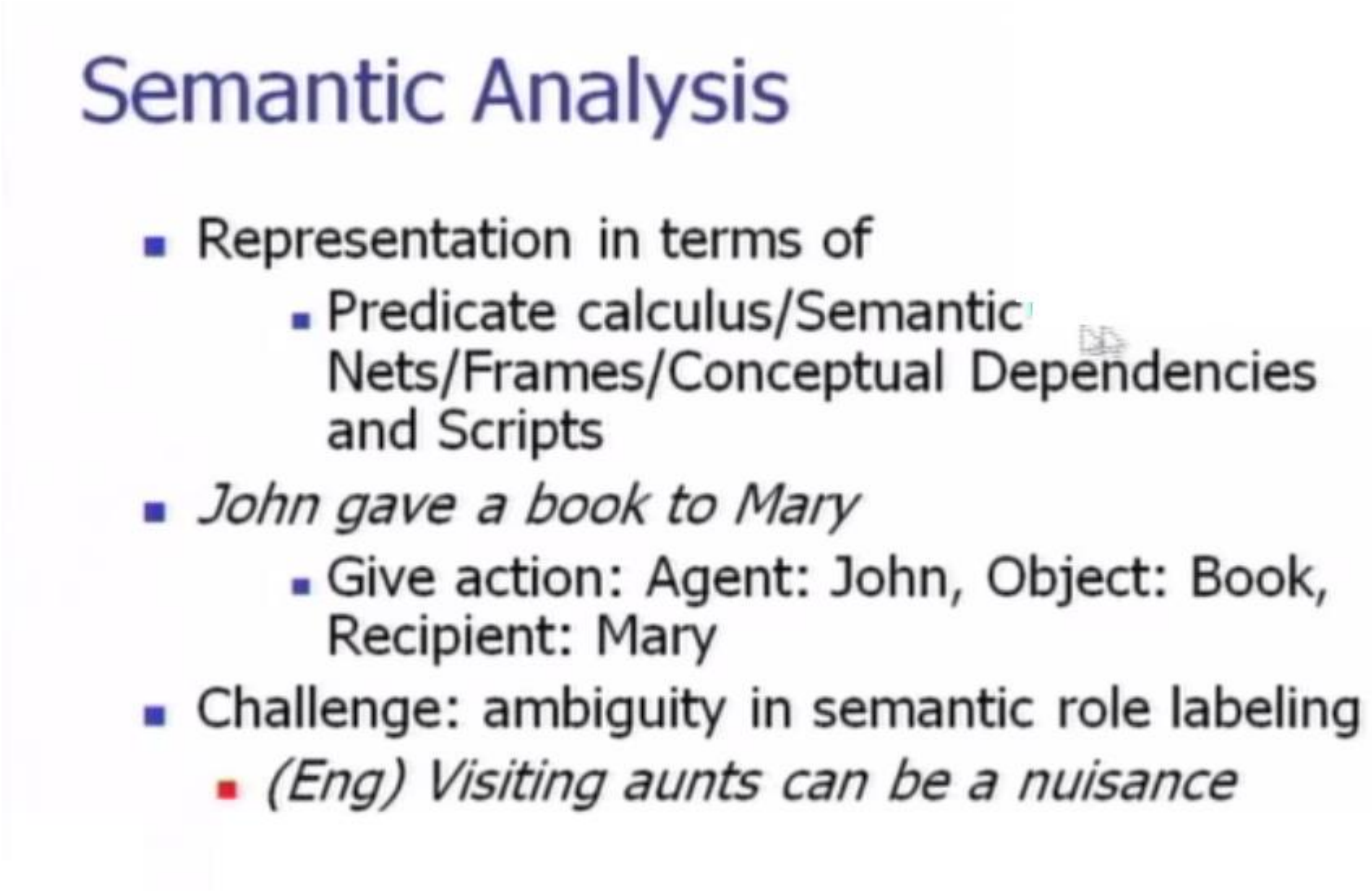
Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

Stages of language processing

- Phonetics and Phonology علم الأصوات الكلامية
- Morphology علم الصرف
- Lexical Analysis
- Syntactic Analysis بناء الجملة ؛ النحو
- Semantic Analysis دلالات ؛ المعاني
- Pragmatics البراغماتية
- Discourse السياق

Stages of language processing



Semantic Analysis

- Representation in terms of
  - Predicate calculus/Semantic Nets/Frames/Conceptual Dependencies and Scripts
- *John gave a book to Mary*
  - Give action: Agent: John, Object: Book, Recipient: Mary
- Challenge: ambiguity in semantic role labeling
  - *(Eng) Visiting aunts can be a nuisance*

Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

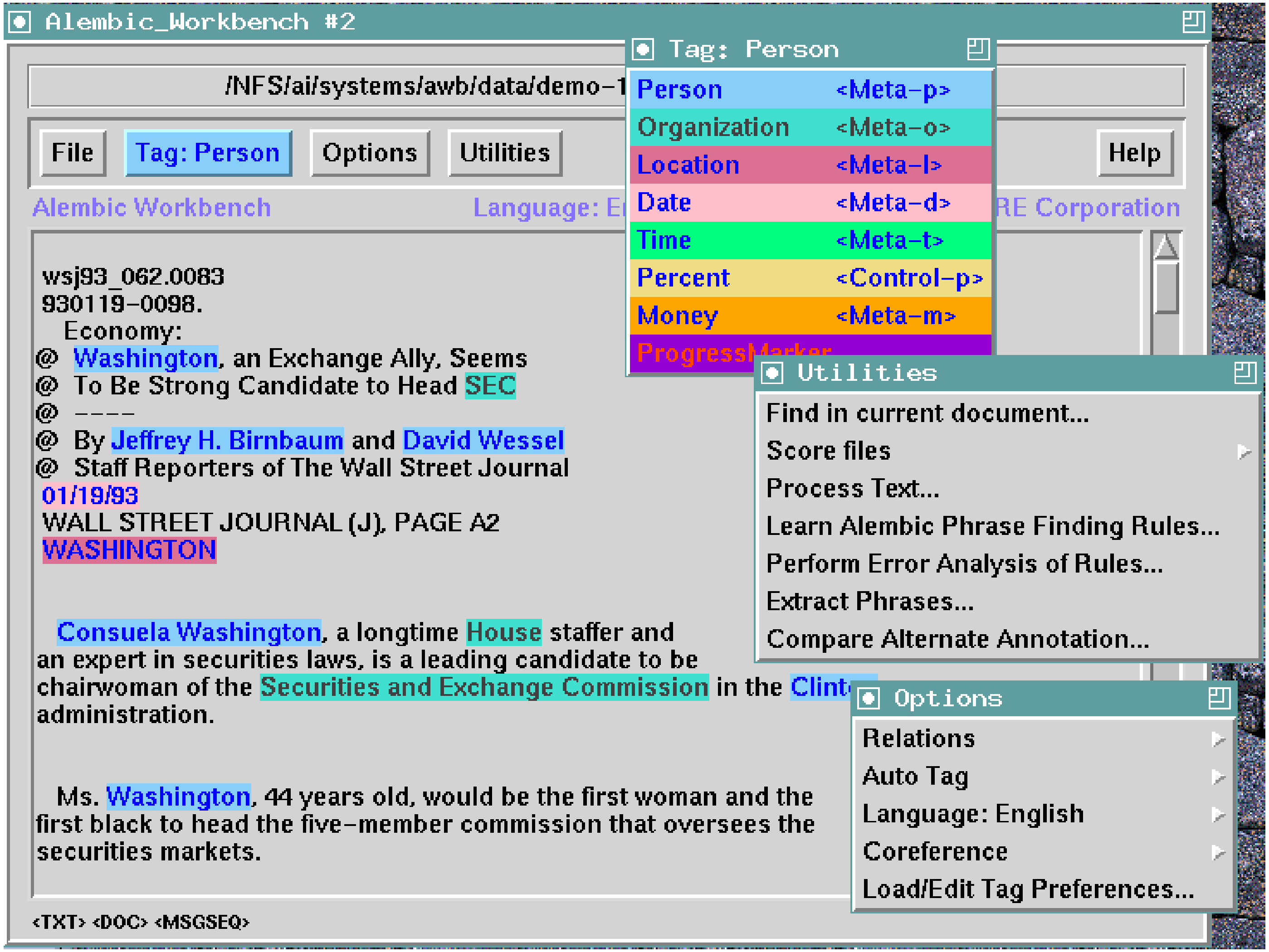
NLP Approches

- NLP can more or less follow theoretical insights
  - Rule-based: model system with linguistic rules
  - Statistical: model system with probabilities of what normally happens
- Hybrid models combine the two

The Annotation of Data

- If we want to learn linguistic properties from data, we need to annotate the data
  - Train on annotated data
  - Test methods on other annotated data
- Through the annotation of corpora, we encode linguistic information in a computer-usable way.

An Annotation Tool

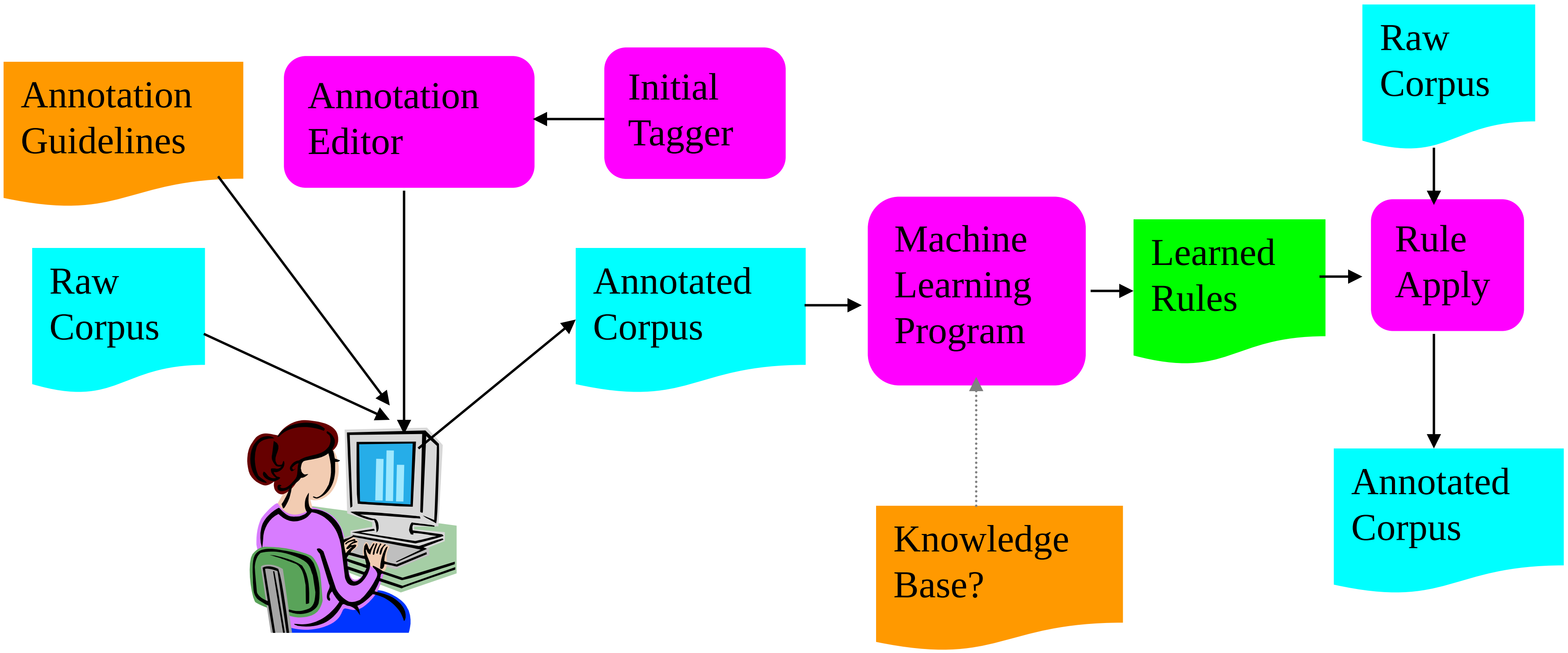


wsj93_062.0083
930119-0098.
Economy:
@ Washington, an Exchange Ally, Seems
@ To Be Strong Candidate to Head SEC
@ ----
@ By Jeffrey H. Birnbaum and David Wessel
@ Staff Reporters of The Wall Street Journal
01/19/93
WALL STREET JOURNAL (J), PAGE A2
WASHINGTON

Consuela Washington, a longtime House staffer and an expert in securities laws, is a leading candidate to be chairwoman of the Securities and Exchange Commission in the Clint administration.

Ms. Washington, 44 years old, would be the first woman and the first black to head the five-member commission that oversees the securities markets.

Knowledge Discovery Methodology



Annotation Guidelines

Annotation Editor

Initial Tagger

Raw Corpus

Annotated Corpus

Machine Learning Program

Learned Rules

Raw Corpus

Rule Apply

Annotated Corpus

Knowledge Base?

Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

NLP Applications

- Classifiers: *classify a set of document into categories, (as spam filters)*
- Information Retrieval: *find relevant documents to a given query.*
- Information Extraction: *Extract useful information from resumes; discover names of people and events they participate in, from a document.*
- Machine Translation: *translate text from one human language into another*
- Question Answering: *find answers to natural language questions in a text collection or database…*

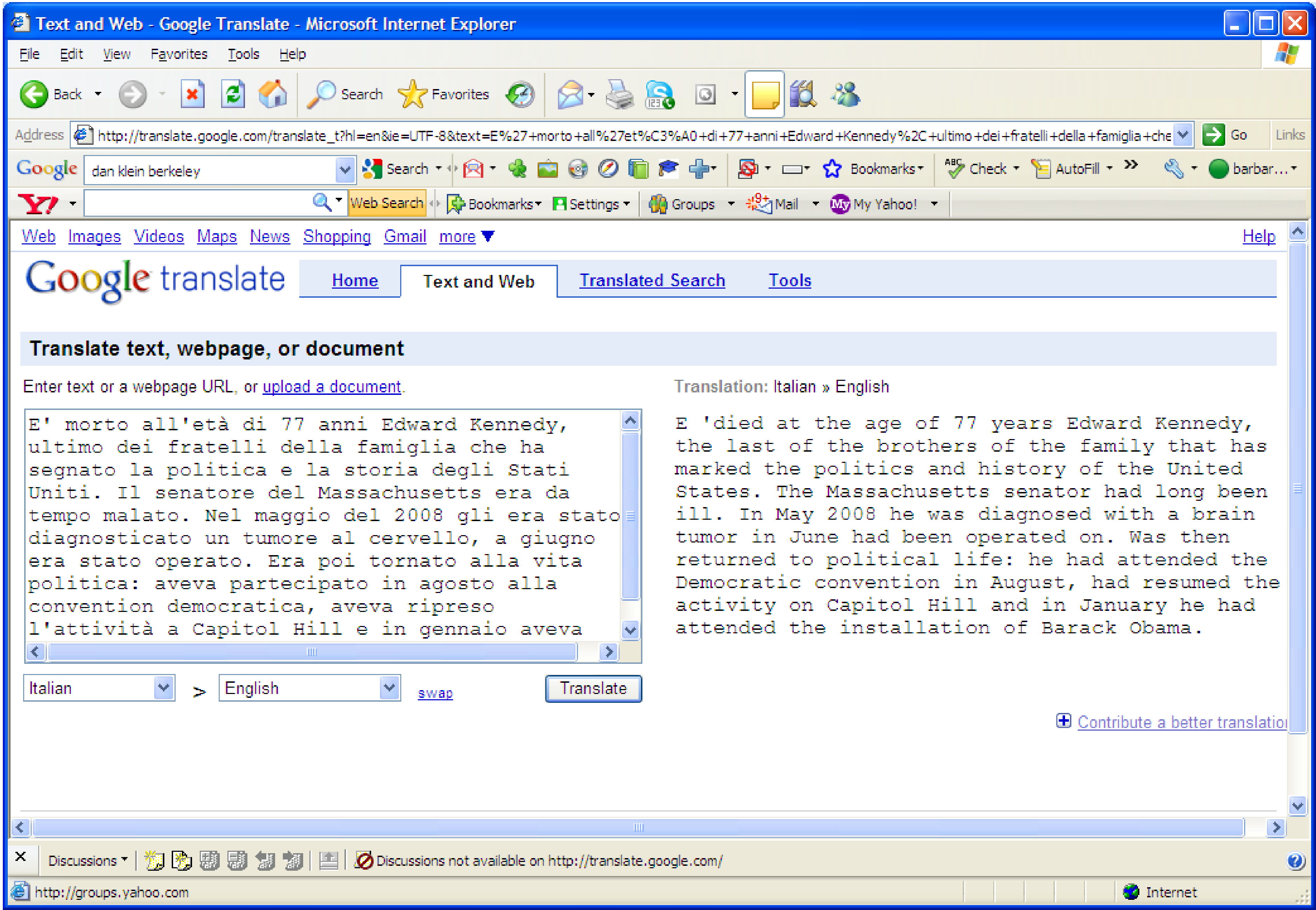
NLP Applications(cont.)

- Summarization: *Produce a readable summary, e.g., news about oil today.*
- Sentiment Analysis: *identify people opinion on a subjective.*
- Speech Processing: *book a hotel over the phone, TTS (for the blind)*
- OCR: *both print and handwritten.*
- Spelling checkers, grammar checkers, auto-filling, ….. and more

Application #1: Machine Translation

- Using different techniques for linguistic analysis, we can:
  - Parse the contents of one language
  - Generate another language consisting of the same content

Example: Machine Translation



http://translate.google.com/translate_t?hl=en&ie=UTF-8&text=E%27+morto+all%27et%C3%A0+di+77+anni+Edward+Kennedy%2C+ultimo+dei+fratelli+della+famiglia+che

dan klein berkeley

Translate text, webpage, or document

Enter text or a webpage URL, or upload a document.

E' morto all'età di 77 anni Edward Kennedy, ultimo dei fratelli della famiglia che ha segnato la politica e la storia degli Stati Uniti. Il senatore del Massachusetts era da tempo malato. Nel maggio del 2008 gli era stato diagnosticato un tumore al cervello, a giugno era stato operato. Era poi tornato alla vita politica: aveva partecipato in agosto alla convention democratica, aveva ripreso l'attività a Capitol Hill e in gennaio aveva

Italian > English

Translation: Italian » English

E 'died at the age of 77 years Edward Kennedy, the last of the brothers of the family that has marked the politics and history of the United States. The Massachusetts senator had long been ill. In May 2008 he was diagnosed with a brain tumor in June had been operated on. Was then returned to political life: he had attended the Democratic convention in August, had resumed the activity on Capitol Hill and in January he had attended the installation of Barack Obama.

Discussions not available on http://translate.google.com/

http://groups.yahoo.com

Application #2: Question Answering

Yesterday Holly was running a marathon when she twisted her ankle. David had pushed her.

1. When did the running occur?

2. When did the twisting occur?

3. Did the pushing occur before the twisting?

4. Did Holly keep running after twisting her ankle?

Application #3: Information Extraction

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

Company$_{NG}$ Set-UP$_{VG}$ Joint-Venture$_{NG}$ with Company$_{NG}$

Produce$_{VG}$ Product$_{NG}$

The joint venture, Bridgestone Sports Taiwan Cp., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

KEY:

Named Entity tagging

Chunk parsing: NGs, VGs, preps, conjunctions

Information Extraction: Filling Templates

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

Activity:

Type: PRODUCTION

Company:

Product: *golf clubs*

Start-date:

The joint venture, Bridgestone Sports Taiwan Cp., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Activity:

Type: PRODUCTION

Company: Bridgestone Sports Taiwan Co

Product: iron and "metal wood" clubs

Start-date: DURING 1990

NLP applications

- Text Categorization
  - Classify documents by topics, language, author, spam filtering, information retrieval (relevant, not relevant), sentiment classification (positive, negative)
- Spelling & Grammar Corrections
- Speech Recognition
- Summarization
- Dialog Systems
  - Language generation

Corpus-based statistical approaches to tackle NLP problem

- How can a can a machine understand these differences?
  - Decorate the cake with the frosting
  - Decorate the cake with the kids
- Rules based approaches, i.e. hand coded syntactic constraints and preference rules:
  - The verb *decorate* require an animate being as agent
  - The object *cake* is formed by any of the following, inanimate entities (cream, dough, frosting…..)
- Such approaches have been showed to be time consuming to build, do not scale up well and are very brittle to new, unusual, metaphorical use of language
  - *To swallow* requires an animate being as agent/subject and a physical object as object
    - I swallowed his story
    - The supernova swallowed the planet

Corpus-based statistical approaches to tackle NLP problem

- A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from text collections (corpora)
- Statistical models are robust, generalize well and behave gracefully in the presence of errors and new data.
- So:
  - Get large text collections
  - Compute statistics over those collections
  - (The bigger the collections, the better the statistics)

Corpus-based statistical approaches to tackle NLP problem

- Topic categorization: classify the document into semantics topics

Document 1

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie.

Topic = sport

Document 2

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine.

Corpus-based statistical approaches to tackle NLP problem

- Topic categorization: classify the document into semantics topics

Document 1 (sport)

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan …

Document 2 (disasters)

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as….

- From (labeled) corpora we can learn that:

  #(sport documents containing word *Cup*) > #(disaster documents containing word *Cup*) **-- feature**

- We then need a statistical model for the topic assignment

Corpus-based statistical approaches to tackle NLP problem

- **Feature extractions (usually linguistics motivated)**
- **Statistical models**

- **Data (corpora, labels, linguistic resources)**

Agenda:

- What is NLP?
- Where does it fit in the CS taxonomy?
- Why NLP is important?
- Stages of language processing
- NLP Approaches
- Applications
- Objective

Objective

Objective

- Generate semantic representation based on ontology.

Acknowledgement

- Some of the slides in this presentation are based on the following resources, but with many additions and revision.