Artificial Intelligence

Introduction to
Natural Language Processing

Dr. Mustafa Jarrar

Sina Institute, Birzeit University

mjarrar@birzeit.edu
www.jarrar.info

Mustafa Jarrar: Online Courses – Watch and Download Slides

Reference:
Mustafa Jarrar: Lecture Notes on Conceptual Analyses
University of Birzeit, Palestine, 2015

Online Course

Watch Do ad Slides

Mustafa Jarrar

Birzeit University, Palestine
mjarrar@birzeit.edu

Watch this lecture and download the slides from

http://jarrar-courses.blogspot.com/2011/11/artificial-intelligence-fall-2011.html

Outline

- NLP Applications
- NLP and Intelligence
- Linguistics Levels of ambiguity
- Language Models

Keywords: Natural Language Processing ,NLP, NLP Applications, NLP and Intelligence, Linguistics Levels of ambiguity, Language Models, Part of Speech Tagging,
المعالجة الآلية للغات الطبيعية , تطبيقات لغوية,الغموض اللغوي، التحليل اللغوي الآلي ,اللسانيات الحاسوبية

Motivation

Which NLP applications do you use every day?
(➔how much money these companies are making?)

- Google, Microsoft, Yahoo,
- Job Seeking
- Google translate Systran powers Babelfish
- Myspace, Facebook, Blogspot
- Tools for “business intelligence”
- …..

➔ Most ideas stem from Academia, but big guys have (several) strong NLP research labs (like Microsoft, Yahoo, AT&T, IBM, etc.)

Why Natural Language Processing?

- Huge amounts of data on the Internet, Intranets, desktops,
- We need applications for processing (understanding, retrieving, translating, summarizing, …) this large amounts of texts.
- Modern applications contain many NLP components. Imagine your address book without good NLP to smartly search your contacts!!!

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…
- 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

Natural Language? and Intelligence?

- **Artificial languages**, like C# and Java
- Automatic processing of computer languages is easy! why?

- **Natural Language**, that people speak, like English, Arabic, …
- Automatic processing (analyzing, understanding, generating,…) of natural languages is very difficult! why?

- Intelligence: Natural? and Artificial (AI).
- Computers are called intelligent if thy are able to process (analyze, understand, learn,…) natural languages as humans do.

- Modern NLP algorithms are based on machine learning, especially *statistical machine learning*.

NLP Current Motives

- Historically: peaks and valleys. Now is a peak, 20 years ago may have been a valley.
- Security agencies are typically interested in NLP.
- Most big companies nowadays are interested in NLP
- The internet and mobile devices are important driving forces in NLP research.

Computers Lack Knowledge!

Based on [1]

This is how computers “see” text in English.

kJfmmfj mmmvvv nnnffn333

Uj iheale eleee mnster vensi credur

Baboi oi cestnitze

Coovoel2^ ekk; ldsllk lkdf vnnjfj?

Fgmflmllk mlfm kfre xnnn!

- People have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
- Computers have
  - No common sense knowledge
  - No reasoning capacity

Linguistics Levels of Ambiguity/Analysis

Based on [1]

Speech

Written language

- Phonology: sounds / letters / pronunciation
  (*two, too*. سائد، صائد)

- Morphology: the structure of words
  (child – children, book - books; *كتاب-كتب، طفل-أطفال، أكل-يأكل*)

- Syntax: grammar, how these sequences are structured
  I saw the man with the telescope *رأيته بالنظارة*

- Semantics: meaning of the strings
  (table as data structure, table as furniture. *جدول-مصفوفة، جدول-نهر*)

➢ Dealing with all of these levels of ambiguity make NLP difficult

Issues in Syntax

Based on [1]

Syntax does not deal with the meaning of a sentence, but it may help?!

"the dog ate my homework"

Who ate? →dog

The important thing when we analyze a syntax is to identify the part of speech (POS): Dog = noun ; ate = verb ; homework = noun

There are programs that do this automatically, called: Part of Speech Taggers. (also called grammatical tagging)

Accuracy of English POS tagging: 95%.

Identify collocations

mother in law, hot dog

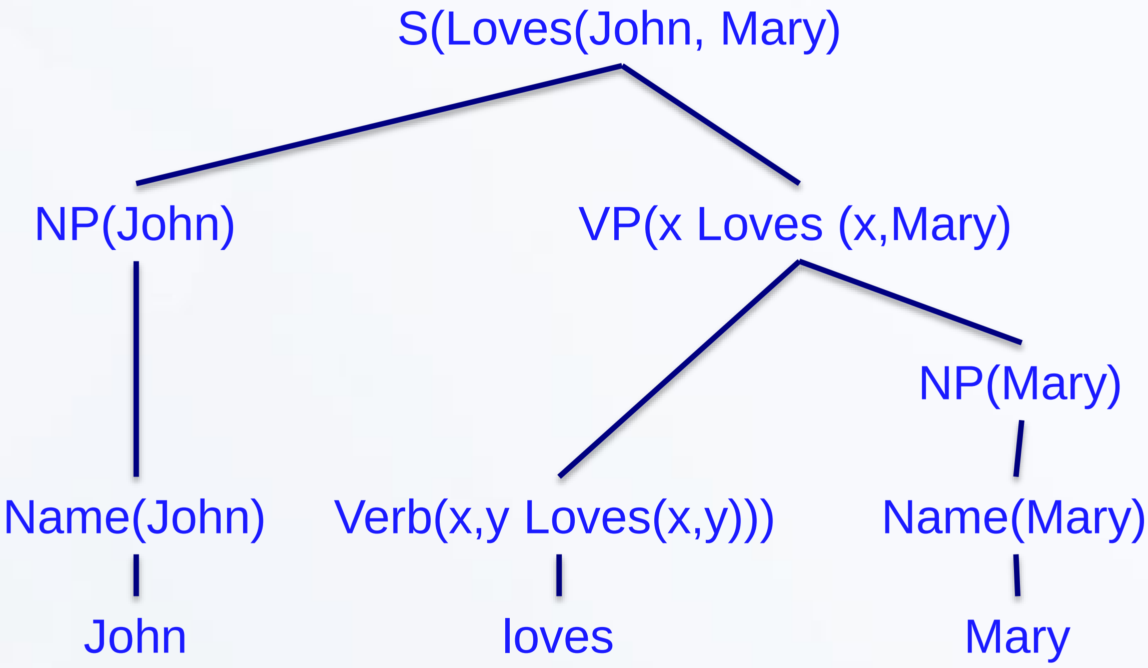
Compositional versus non-compositional collocates

Issues in Syntax (Part of Speech Tagging)

Based on [1]

Assume input sentence **S** in natural language **L**. Assume you have rules (*grammar* **G**) that describe syntactic regularities (patterns or structures). Given **S** & **G**, find syntactic structure of **S**. Such a structure is called a Parse Tree

Pars tree: John loves Mary



Helps a computer to automatically answer questions like -Who did what and when?

Issues in Syntax

Based on [1]

Shallow Parsing:

An analysis of a sentence which identifies the constituents (noun groups,verbs, verb groups, etc.), but does not specify their internal structure, nor their role in the main sentence.

Example:

“John Loves Mary”

“John” “Loves Mary”

subject predicate

Identify basic structures as:

NP-[John] VP-[Loves Mary]

More Issues in Syntax

Based on [1]

Anaphora Resolution: resolving what a pronoun, or a noun phrase refers to. *“The dog entered my room. It scared me”*

Preposition Attachment

I saw the man in the park with a telescope

رأيت الرجل الجالس بالنظارة

The son asked the father to drive him home

طلبت الأم من البنت تصفيف شعرها

Issues in Semantics

How to understand the meaning, specially that words are ambiguous and **polysemous** (may have multiple meanings)

Buy this table? serve that table? sort the table?

هل رأيت هذه الطاولة. هل خدمت هذه الطاولة.

How to learn the meaning of words?

- From available dictionaries? WordNet?
- Applying statistical methods on annotated examples?

How to learn the meaning (word-sense disambiguation)?

Assume a (large) amount of annotated data = training
Assume a new text not annotated = test

Learn from previous experience (training) to classify new data (test)
Decision trees, memory based learning, neural networks

Language Models

Three approaches to Natural Language Processing (Language Models)

- Rule-based: using a predefined set of rules (knowledge)
- Statistical: using probabilities of what normally people write or say
- Hybrid models combine the two

Acknowledgement

Some of the slides in this lecture are based on the following resources , but with many additions and revision:

[1] Rada Mihalcea: Natural Language Processing, 2008
www.cs.odu.edu/~mukka/cs480f09/Lecturenotes/.../Intro1.ppt

[2] Markus Dickinson: Introduction to Natural Language Processing (NLP), Linguistics 362 course, 2006
http://www9.georgetown.edu/faculty/mad87/06/362/syllabus.html