NAMED ENTITY RECOGNITION (NER)

Gabriella Pasi
gabriella.pasi@unimib.it

Named Entity Recognition (NER)

A very important sub-task of Information Extraction, aimed at finding and classifying names in texts, for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

Named Entity Recognition

- *Named entity recognition*
  - identify words that refer to *names* of interest in a particular application
  - e.g., people, companies, locations, product names
- *Text is mapped to the identified names*
  - Task: what is the most likely mapping

Named Entity Recognition

- Example :

> Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

> Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

Organization, Location, Person

Linguistically difficult problem

- NER involves ***identification*** of *proper names* in texts, and ***classification*** into a set of predefined categories of interest.
- Three universally accepted categories: **person**, **location** and **organisation.**
- Other common tasks: recognition of **date/time expressions**, **measures** (percent, money, weight etc), email addresses etc.
- Other domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

Applications of NER

- Yellow pages with local search capabilities
- Monitoring trends and sentiment in textual social media
- Interactions between genes and cells in biology and genetics

Problems in NER task definition

- Category definitions are intuitively quite clear, but there are many grey areas.

- Many of these grey areas are caused by **metonymy**.
  - Organisation vs. Location : “**England** won the World Cup” vs. “The World Cup took place in **England**”.
  - Company vs. Artefact: “shares in **MTV**” vs. “watching **MTV**”
  - Location vs. Organisation: “she met him at **Heathrow**” vs. “the **Heathrow** authorities”

Named Entity Recognition

1) Rule-based

- Uses *lexicons* (lists of words and phrases) that categorize names
  - e.g., locations, peoples' names, organizations, etc.

- Rules also used to verify or find new entity names
  - e.g., "<number> <word> street" for addresses
  - "<street address>, <city>" or "in <city>" to verify city names
  - "<street address>, <city>, <state>" to find new cities
  - "<title> <name>" to find new names

- Rules either developed manually by trial and error or by using machine learning techniques

- Language dependent

Named Entity Recognition

- 2) *Statistical machine learning*
  - uses a probabilistic model of the words in and around an entity
  - probabilities estimated using *training data* (manually annotated text)
  - Hidden Markov Model (HMM) is one approach (other approaches Conditional Random Fields, Support Vector Machines…)

Hidden Markov model for NER

- Resolve ambiguity in a word using *context*
  - e.g., “marathon” is a location or a sporting event, “boston marathon” is a specific sporting event

- Model context using a *generative* model of the sequence of words

HMM for Extraction

- The HMM is based on augmenting the Markov chain

- *Markov Model (Markov Chain)* describes a process as a sequence of states (random variables) with transitions between them.
  - each transition has a probability associated with it
  - next state depends only on current state and transition probabilities

Markov Assumption: $P(q_i = a | q_1 \ldots q_{i-1}) = P(q_i = a | q_i - 1)$

HMM for Extraction

- The HMM is based on augmenting the Markov chain

- *Markov Model (Markov Chain)*

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{N} \pi_i = 1$

HMM for Extraction

- The HMM is based on augmenting the Markov chain

Markov chain useful to compute a probability for a sequence of ***observable events***.

Often events we are interested in are hidden: we do not observe them directly. For example, in a text we do not normally observe part-of-speech tags; we see words and must infer the tags from the word sequence.

We call the tags ***hidden*** because they are not observed.

HMM for Extraction

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

As with a first-order Markov chain, the probability of a particular state depends only on the previous state.

The probability of an output observation *oi* depends only on the state that produced the observation *qi* and not on any other states or any other observations.

HMM for Extraction

- To recognize named entities, find sequence of “labels” that give highest probability for the sentence
  - only the outputs (words) are visible or observed
  - states are “hidden”
  - E.g
    - Fred Smith, who lives at 10 Water Street, Springfield, MA, is a long time collector of tropical fish.
    - <start><name><not-an-entity><location><not-an-entity><end>

- *Viterbi* algorithm implements a Markov tagging process and can be used for recognition

Named Entity Recognition

- Accurate recognition requires about 1M words of training data (1,500 news stories)
  - may be more expensive than developing rules for some applications

- Both rule-based and statistical can achieve about 90% effectiveness for categories such as names, locations, organizations
  - others, such as product name, can be much worse

Open source sw for POS and NER:

https://nlp.stanford.edu/software/

CORENLP: https://corenlp.run/

What is NLP ?

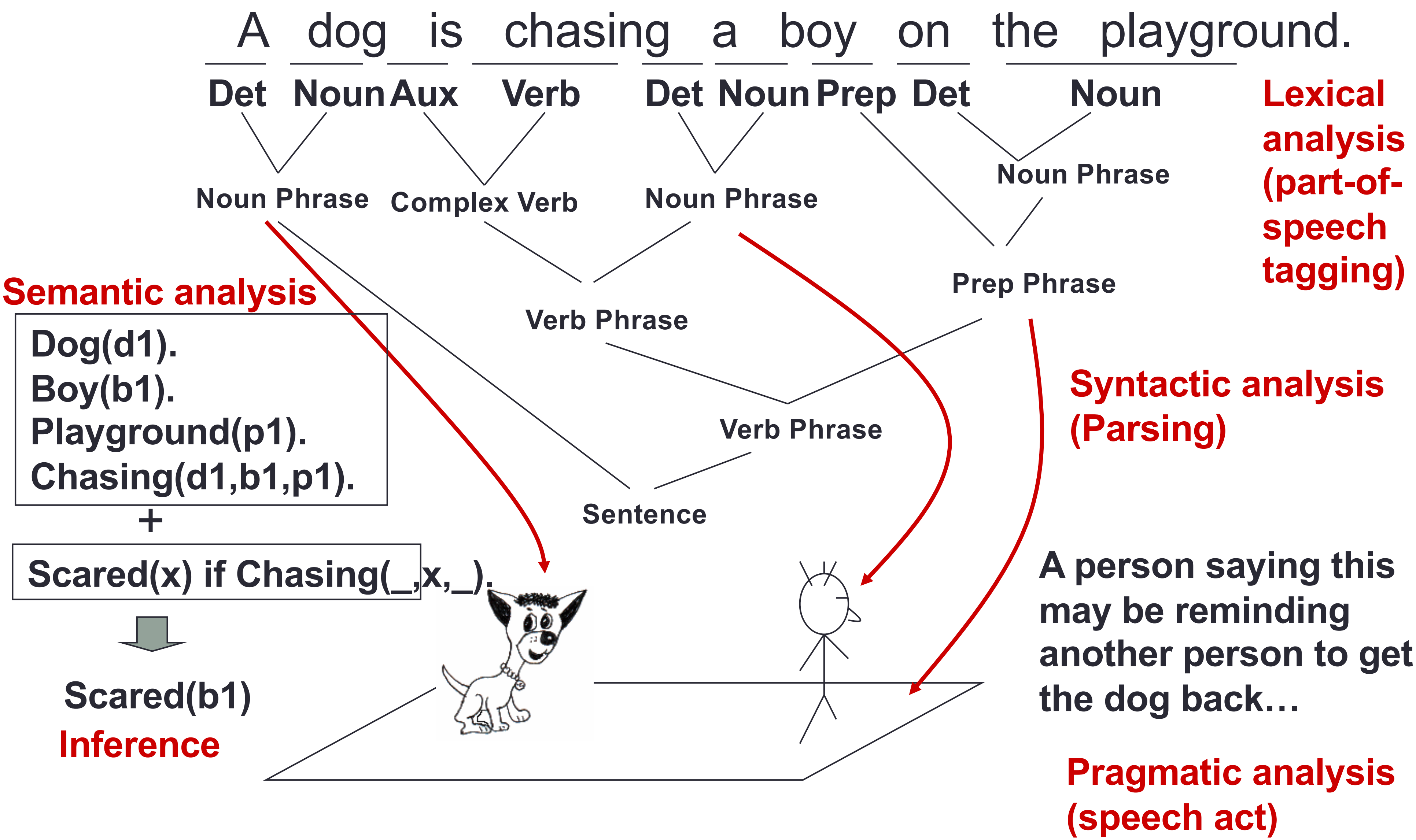
What is NLP?

Arabic text كلب هو مطاردة صبي في الملعب.

How can a computer make **sense** out of this **string?**

Morphology	- What are the basic units of meaning (words)? - What is the meaning of each word?
Syntax	- How are words related with each other?
Semantics	- What is the “combined meaning” of words?
Pragmatics	- What is the “meta-meaning”? (speech act)
Discourse	- Handling a large chunk of text
Inference	- Making sense of everything

An example of NLP



Pragmatics

“Pragmatics is the study of the relationships between language and context, which are fundamental to explain the understanding of the language itself. "

Giovanni wanted to buy Carlo a present for his birthday so he went to get his pig; he shook it, but heard no noise; he should have given Bill a gift with his own hands. " To understand this story you need to know many facts: gifts are usually bought with money, piggy banks can be in the shape of a pig, pigs are usually made of materials such as plastic or metal, money in a container made of such materials generally make a metallic noise, etc ...

To understand this story you need to know many facts: gifts are usually bought with money, piggy banks can be in the shape of a pig, pigs are usually made of materials such as plastic or metal, money in a container made of such materials generally make a metallic noise, etc ...

The notion of Context

"*The situation in which the communicative act takes place, the set of knowledge, beliefs and the like shared by both the issuer and the recipient and such as to guide the understanding of the communicative act*."

a) knowledge of the role and status of the interlocutors.
b) knowledge of the spatial and temporal location.
c) knowledge of the level of formality.
d) knowledge of the tool.
e) knowledge of the appropriate topic.
f) knowledge of the domain that determines the usage of a language.

The notion of Context

The notion of context is not a simple one.

The context includes at a minimum the beliefs and assumptions of the language users, relating to:

- previous, simultaneous and future actions (both verbal and non-verbal).
- temporal, spatial and social situations.
- the state of knowledge and attention of those who participate to social interactions

The notion of context

"Let's go here, please.»

This sentence shows how the context of a dialogue is important:

It is anomalous if the two interlocutors are in the same place designated as here.
On the other hand, it makes sense if the two interlocutors are consulting a geographical map and refer to a place on that map.

If we can do all the above levels of analysis for all the sentences in all languages, then we could …

- *Automatically answer our emails*
- *Translate languages accurately*
- *Help us manage, summarize, and aggregate information*
- *Use speech as a UI (when needed)*
- *Talk to us / listen to us*

BAD NEWS:

- **Unfortunately, we cannot right now.**
- **General NLP = "Complete AI"**

NLP is difficult!!!!!!!

- Natural language is designed to make human communication efficient. Therefore,
  - We omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
  - We keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve
- This makes EVERY step in NLP hard
  - **Ambiguity is a “killer”!**
  - Common sense reasoning is pre-required

An example of ambiguity

- Get the cat with the gloves.

Examples of challenges

- Word-level ambiguity
  - “design” can be a noun or a verb (Ambiguous POS)
  - “root” has multiple meanings (Ambiguous sense)
- Syntactic ambiguity
  - “natural language processing” (Modification)
  - “A man saw a boy *<u>with a telescope</u>*.” (PP Attachment)
- Anaphora resolution
  - “John persuaded Bill to buy a TV for himself.” (himself = John or Bill?)
- Presupposition
  - “He has quit smoking.” implies that he smoked before.

Despite all the challenges, research in NLP has also made a lot of progress…

A brief history of NLP

- Early enthusiasm (1950's): Machine Translation
  - Too ambitious
  - Bar-Hillel report (1960) concluded that fully-automatic high-quality translation could not be accomplished without knowledge (Dictionary + Encyclopedia)
- Less ambitious applications (late 1960's & early 1970's): Limited success, failed to scale up
  - Speech recognition
  - Dialogue (Eliza)
  - Inference and domain knowledge (SHRDLU="block world")
- Real world evaluation (late 1970's – now)
  - Story understanding (late 1970's & early 1980's)
  - Large scale evaluation of speech recognition, text retrieval, information extraction (1980 – now)
  - Statistical approaches enjoy more success (first in speech recognition & retrieval, later others)
- Current trend:
  - Boundary between statistical and symbolic approaches is disappearing.
  - We need to use all the available knowledge
  - Application-driven NLP research (bioinformatics, Web, Question answering…)

Deep understanding in limited domain

Shallow understanding

Knowledge representation

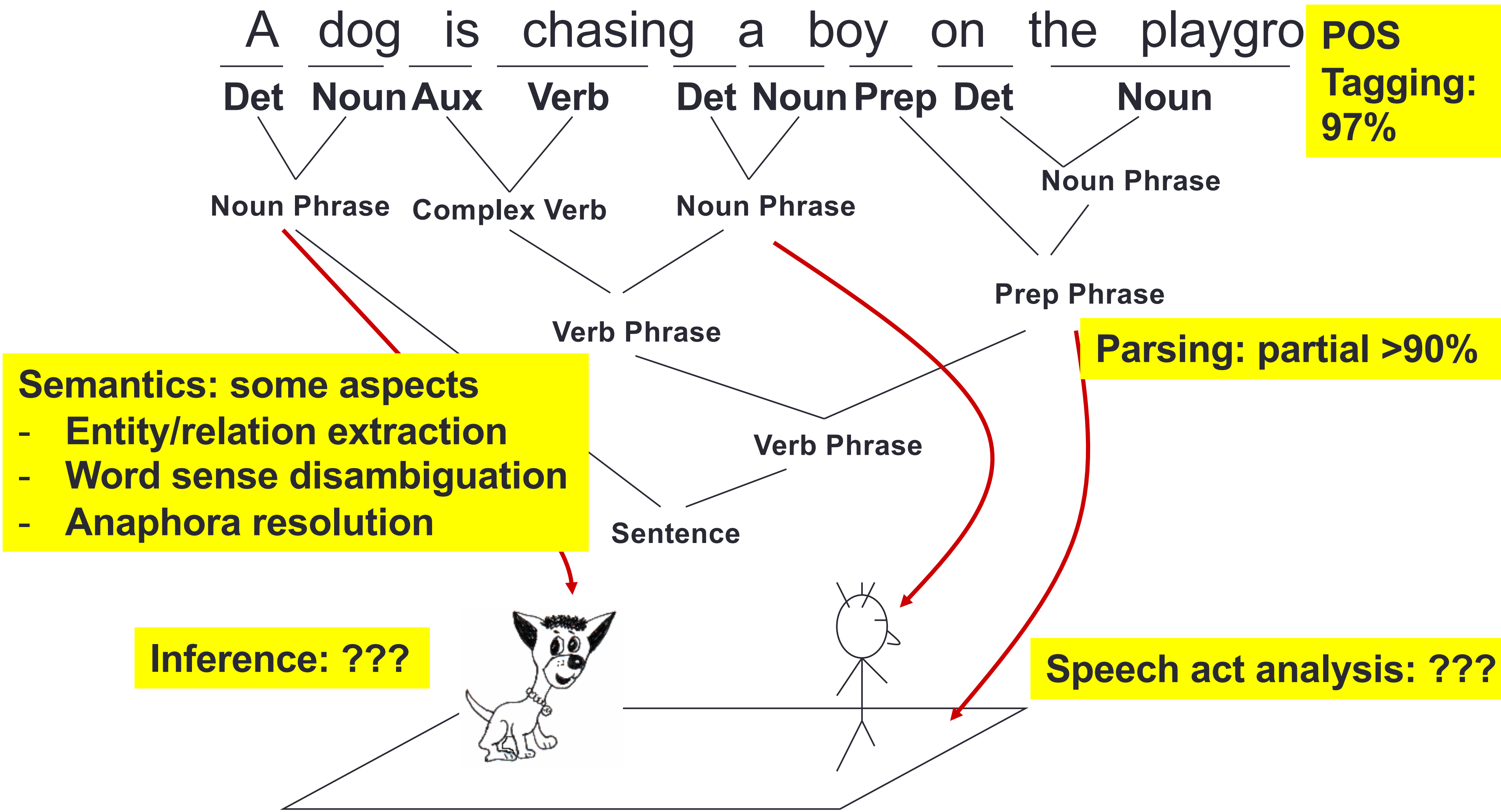
Robust component techniques

Statistical language models

Deep Neural Models

Applications

The state of the art



Machine translation

English | Spanish | French | English - detected

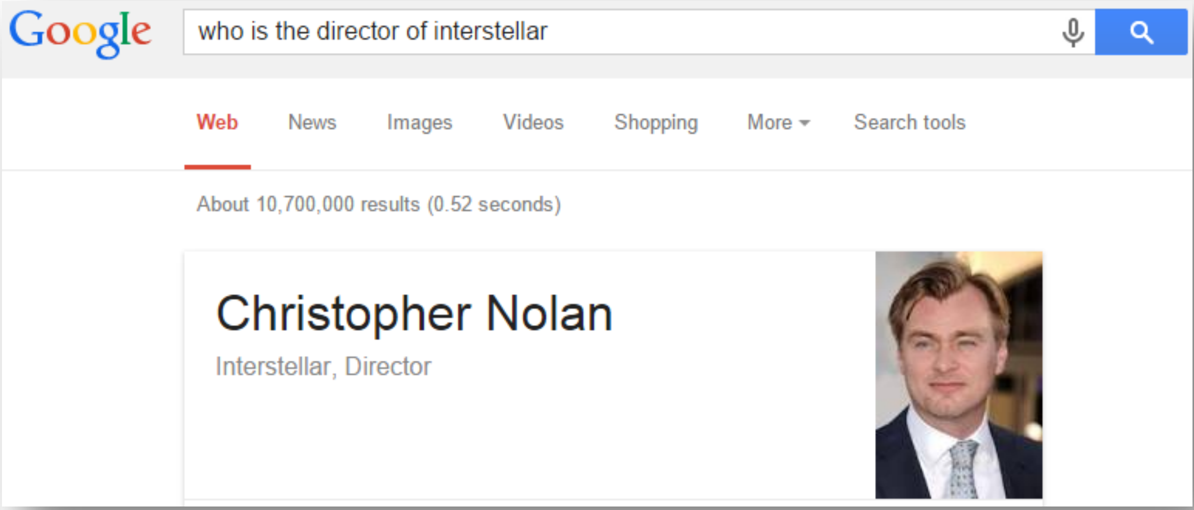
it's a question, but also an expression of disbelief.
Those who get lost driving can use GPS. If you lose your iPhone, there's an app to track it down. Scientists successfully plotted the course for a spacecraft that landed on a speeding asteroid.
How did weather affect AirAsia flight?
But something goes wrong aboard a 123-foot, 67-ton passenger jet and rescuers must resort to scouring the ocean?
"Why is it easier to find an iPhone (than) to find a plane?" one Twitter user, Catalina Buitano, asked.
There are dozens of similar questions on social media. They hint at the same sentiment: in a world where people's locations are tracked for everything from map apps to what ads appear on a web browser, why does Big Brother's gaze avoid the skies?

Google Translate

这是一个问题，但也不敢相信的表情。
这些谁迷路驾驶可以使用GPS。如果你失去了你的iPhone，有一个应用程序来追查。科学家成功绘制过程中的飞船降落在小行星飞驰。
没有天气如何影响亚航的班机吗？
但不顺心的事一艘123英尺，67吨重的喷气式客机和救援人员必须求助于淘海洋？
“为什么更容易找到一个iPhone（比）找到飞机？”1 Twitter的用户，卡塔利娜Buitano，问道。
有几十个在社交媒体上类似的问题。他们暗示相同的感悟：在这个世界上，人的位置进行跟踪，一切从地图应用程序，以广告出现在网页浏览器，为什么大哥的目光避开天空是什么？

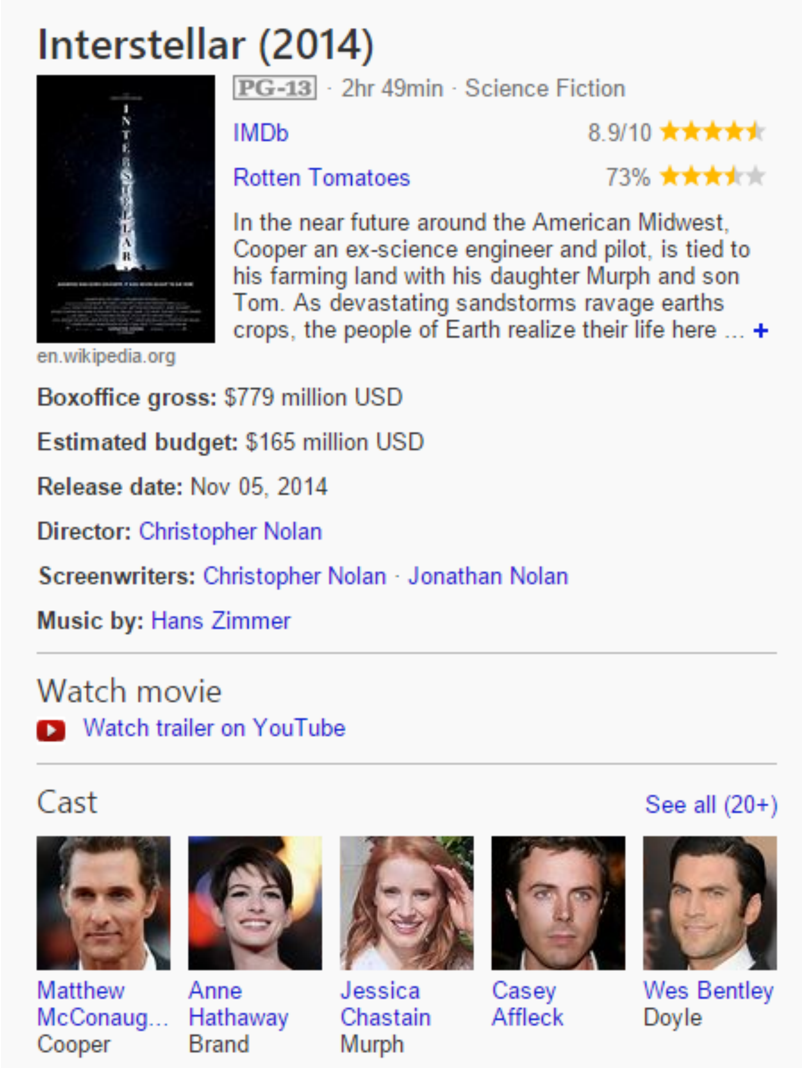
Dialog systems

Apple's siri system

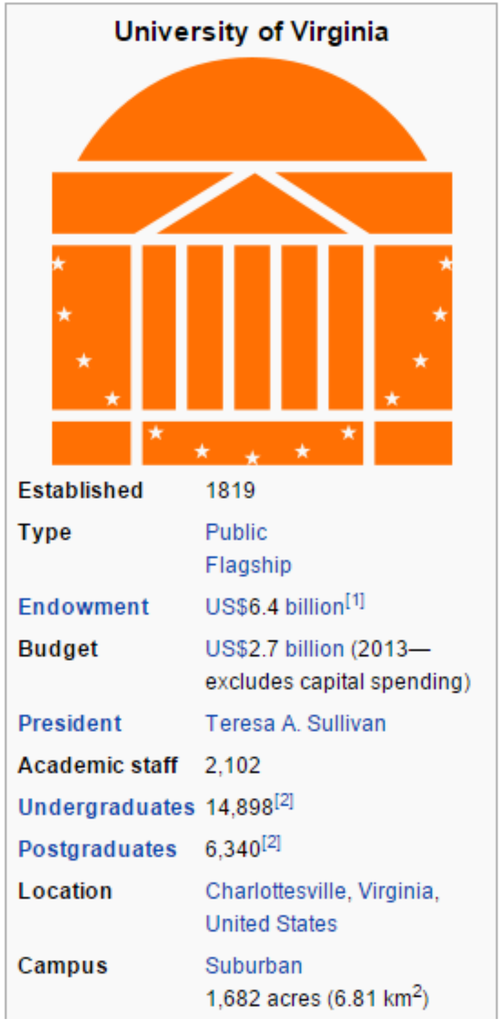


Google search

Information extraction



Google Knowledge Graph



Wiki Info Box

Information extraction

Search: eng

<Albert_Einstein>

← <Elsa_Einstein>
← <Mileva_Marić>
<isMarriedTo>

"albert. ainctain"@jbo
"Albert Einstein"@afr

Recently-Learned Facts

twitter

Refresh

instance	iteration	date learned	confidence
tear_drop_tomatoes is an agricultural product	887	27-nov-2014	93.1
ryan_mckenzie is a professor	886	21-nov-2014	90.2
fiorina_161 is a planet	889	07-dec-2014	92.8
critical_thinking_in_health_science is a cognitive action	886	21-nov-2014	99.0
fateful_new_year is a monarch	886	21-nov-2014	99.0
tony_martin has been charged with murder	890	11-dec-2014	100.0
sen__joe_biden is a U.S. politician who holds the office of vice_president	887	27-nov-2014	93.8
hat is a clothing item to go with blue_jeans	889	07-dec-2014	93.8
statistics is headquartered in the country the_usa	891	18-dec-2014	98.4
eoin_colfer wrote the book artemis_fowl	886	21-nov-2014	100.0

<Abelardo_l

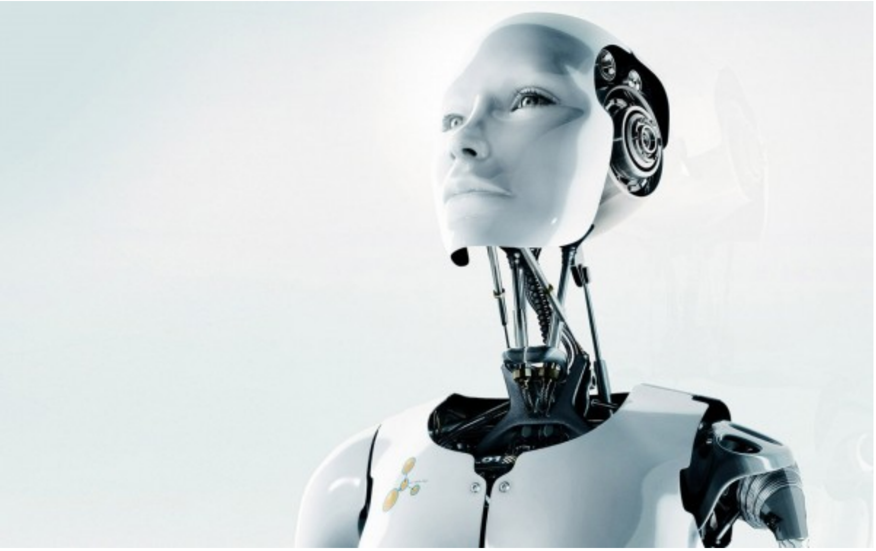
896–1954)>

<Abraham_Pais>
<Abram_L._Sachar>
<Absent-minded_professor>
<Absorption_refrigerator>

<Albert_Einstein's_brain>
<Alfred_Kleiner>
<Annalen_der_Physik>
<Annus_Mirabilis_papers>

CMU Never-Ending Language Learning

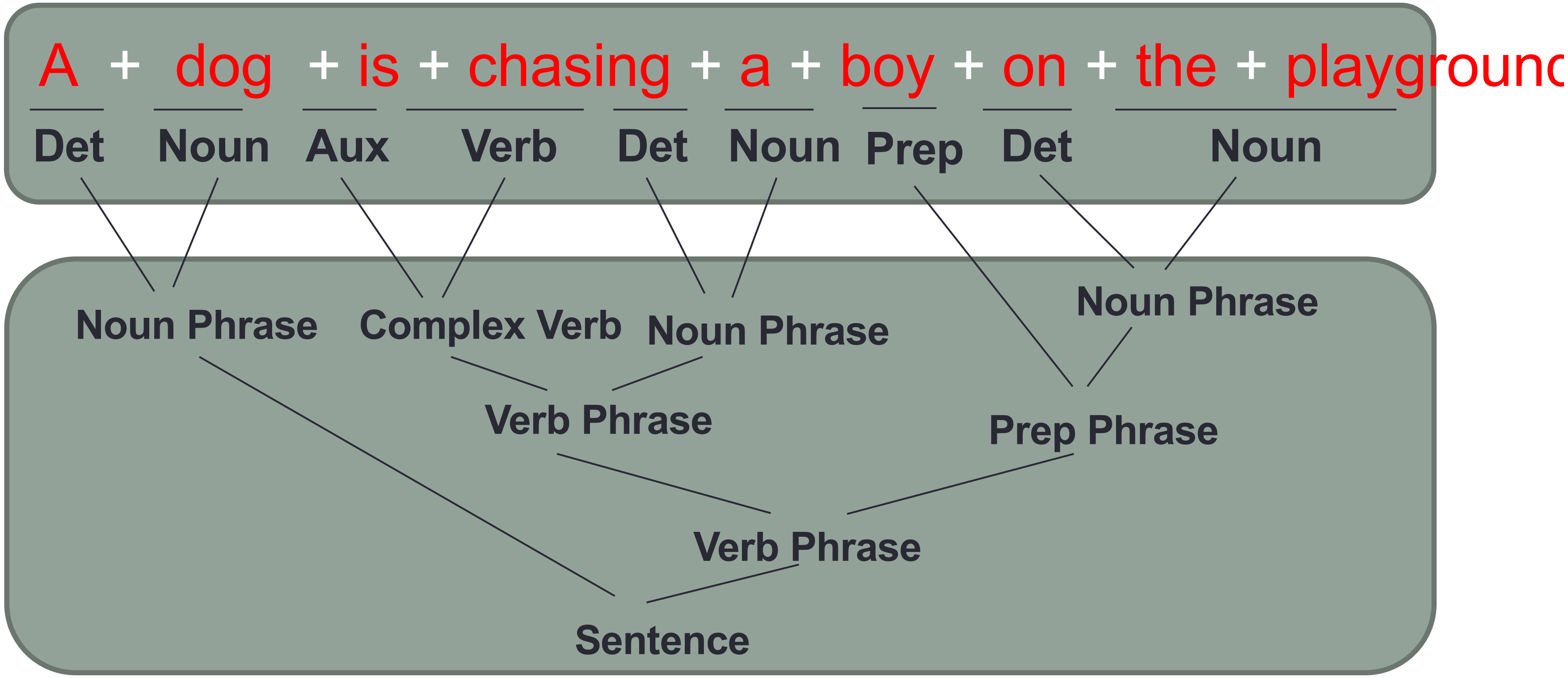
YAGO Knowledge Base



Building a computer that ‘understands’ text: The NLP pipeline

Syntactic parsing

- Grammatical analysis of a given sentence, conforming to the rules of a formal grammar
  - Task: what is the most likely grammatical structure



Relation extraction

- Identify the relationships among named entities
  - Shallow semantic analysis

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

1. Thomas Jefferson Is_Member_Of Board of Visitors
2. Thomas Jefferson Is_President_Of U.S.

Logic inference

- Convert chunks of text into more formal representations
  - Deep semantic analysis: e.g., first-order logic structures

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

$\exists x$ (Is_Person(x) & Is_President_Of(x,'U.S.') & Is_Member_Of(x,'Board of Visitors'))

Towards understanding of text

More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field -- a multi-millionaire.

- Who is Carl Lewis?
- Did Carl Lewis break any records?

Major NLP applications

- Speech recognition: e.g., auto telephone call routing
- Text mining
  - Text clustering
  - Text classification
  - Text summarization
  - Topic modeling
  - Question answering
- Language tutoring
  - Spelling/grammar correction
- Machine translation
  - Cross-language retrieval
  - Restricted natural language
- Natural language user interface

Our focus

NLP & text mining

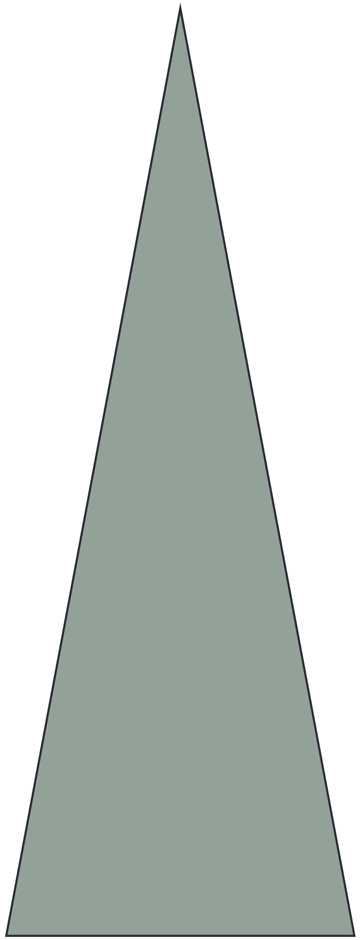
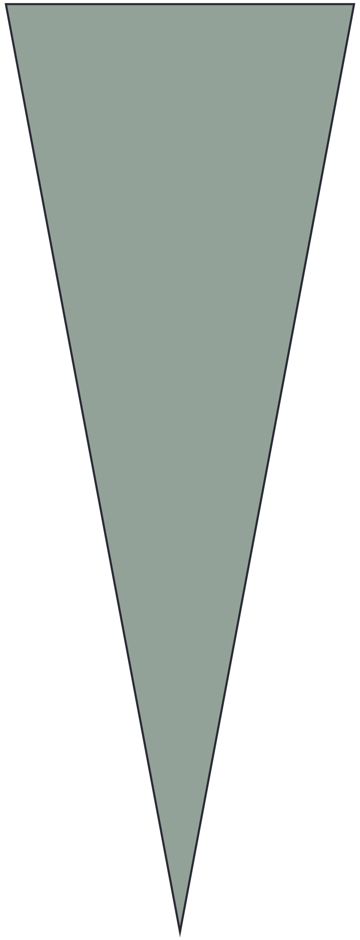
- Better NLP => Better text mining

- Bad NLP => ~~Bad text mining~~?

Robust, shallow NLP tends to be more useful than deep, but fragile NLP.

Errors in NLP can hurt text mining performance…

How much NLP is really needed?

Tasks	**Dependency on NLP**	**Scalability**
Classification Clustering Summarization Extraction Topic modeling Translation Dialogue Question Answering Inference Speech Act		

So, what NLP techniques are the most useful for text mining?

- Statistical NLP in general.
- The need for high robustness and efficiency implies the dominant use of simple models