

# Introduction to Natural Language Processing

Hongning Wang  
CS@UVa

# 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

A dog is chasing a boy on the playground.

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase

Complex Verb

Noun Phrase

Noun Phrase

Prep Phrase

Verb Phrase

Verb Phrase

Sentence

**Lexical analysis**  
(part-of-speech tagging)

**Syntactic analysis**  
(Parsing)

A person saying this may be reminding another person to get the dog back...

**Pragmatic analysis**  
(speech act)

**Semantic analysis**

Dog(d1).  
Boy(b1).  
Playground(p1).  
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(\_,x,\_).



Scared(b1)

**Inference**





- *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*

If we can do this for all the sentences in all languages, then ...

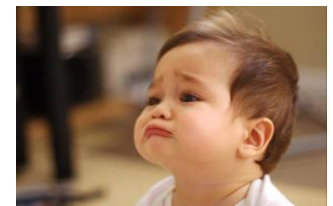
**BAD NEWS**

- Unf
- Gen



**right now.**

”



# 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 with a telescope.” (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.



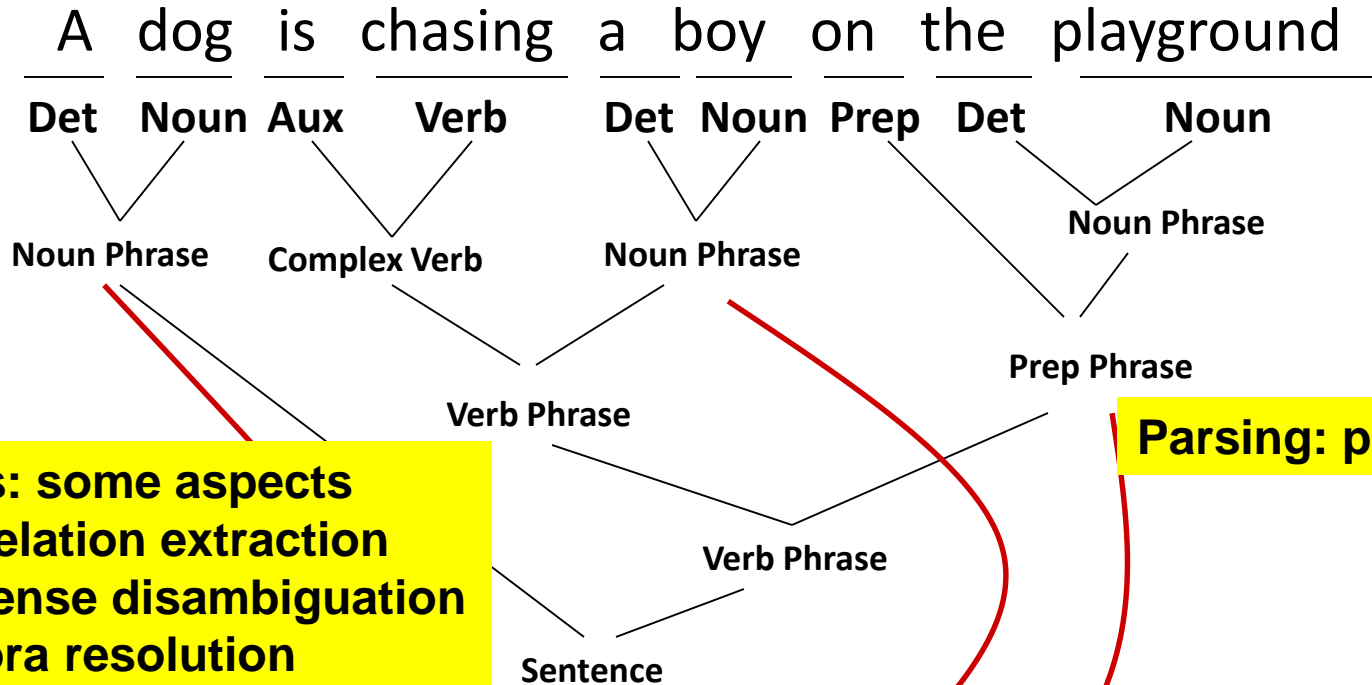
Despite all the challenges, research in NLP has also made a lot of progress...



# 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 **Deep understanding in limited domain**
  - Dialogue (Eliza) **Shallow understanding**
  - Inference and domain knowledge (SHRDLU="block world")
- Real world evaluation (late 1970's – now)
  - Story understanding (late 1970's & early 1980's) **Knowledge representation**
  - Large scale evaluation of speech recognition, text retrieval, information extraction (1980 – now) **Robust component techniques**
  - Statistical approaches enjoy more success (first in speech recognition & retrieval, later others) **Statistical language models**
- Current trend:
  - Boundary between statistical and symbolic approaches is disappearing.
  - We need to use all the available knowledge **Applications**
  - Application-driven NLP research (bioinformatics, Web, Question answering...) **9**

# The state of the art



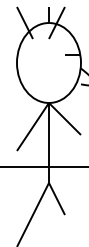
**POS  
Tagging:  
97%**

**Parsing: partial >90%**

**Semantics: some aspects**

- Entity/relation extraction
- Word sense disambiguation
- Anaphora resolution

**Inference: ???**



**Speech act analysis: ???**

# 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?



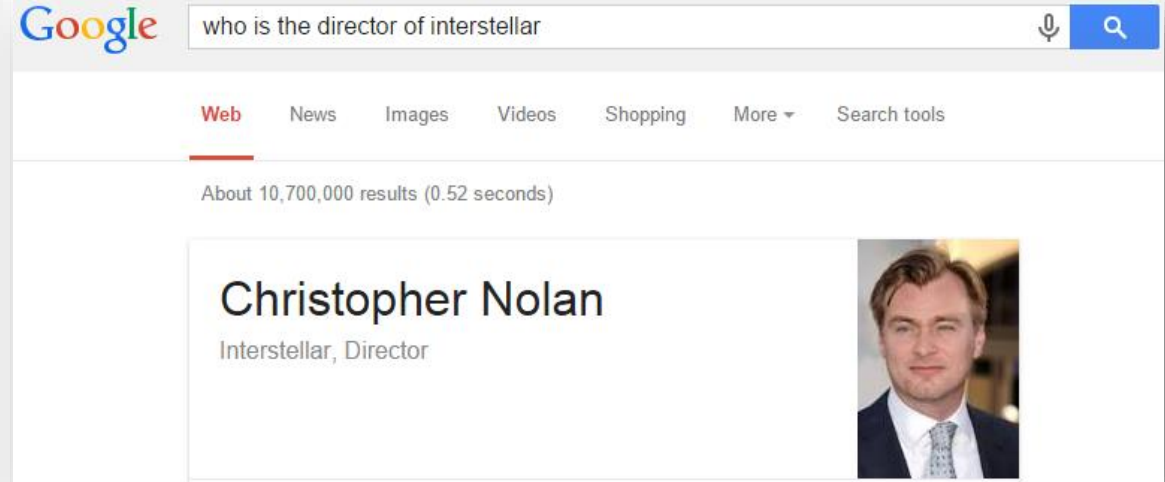
这是一个问题，但也不敢相信的表情。  
这些谁迷路驾驶可以使用GPS。如果你失去了你的iPhone，有一个应用程序来追查。科学家成功绘制过程中的飞船降落在小行星飞驰。  
没有天气如何影响亚航的班机吗？  
但不顺心的事一艘123英尺，67吨重的喷气式客机和救援人员必须求助于淘海洋？  
“为什么更容易找到一个iPhone（比）找到飞机？”1 Twitter的用户，卡塔利娜Buitano，问道。  
有几十个在社交媒体上类似的问题。他们暗示相同的感悟：在这个世界上，人的位置进行跟踪，一切从地图应用程序，以广告出现在网页浏览器，为什么大哥的目光避开天空是什么？



# Dialog systems



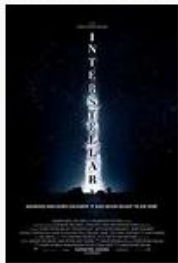
Apple's siri system



Google search

# Information extraction

## Interstellar (2014)



**PG-13** · 2hr 49min · Science Fiction

IMDb

8.9/10 ★★★★★

Rotten Tomatoes

73% ★★★★★

In the near future around the American Midwest, Cooper an ex-science engineer and pilot, is tied to his farming land with his daughter Murph and son Tom. As devastating sandstorms ravage earths crops, the people of Earth realize their life here ... +

[en.wikipedia.org](#)

Boxoffice gross: \$779 million USD

Estimated budget: \$165 million USD

Release date: Nov 05, 2014

Director: [Christopher Nolan](#)

Screenwriters: [Christopher Nolan](#) · [Jonathan Nolan](#)

Music by: [Hans Zimmer](#)

Watch movie

[Watch trailer on YouTube](#)

Cast

[See all \(20+\)](#)



[Matthew McConaughey](#)  
Cooper



[Anne Hathaway](#)  
Brand



[Jessica Chastain](#)  
Murph



[Casey Affleck](#)



[Wes Bentley](#)  
Doyle

## University of Virginia



Established	1819
Type	Public Flagship
Endowment	US\$6.4 billion <sup>[1]</sup>
Budget	US\$2.7 billion (2013— excludes capital spending)
President	<a href="#">Teresa A. Sullivan</a>
Academic staff	2,102
Undergraduates	14,898 <sup>[2]</sup>
Postgraduates	6,340 <sup>[2]</sup>
Location	<a href="#">Charlottesville, Virginia,</a> <a href="#">United States</a>
Campus	Suburban 1,682 acres (6.81 km <sup>2</sup> )

# Information extraction

Search:  eng <Albert\_Einstein>

<Elsa\_Einstein>
<isMarriedTo>
"albert. ainctain"@jbo
"Albert Einstein"@afr

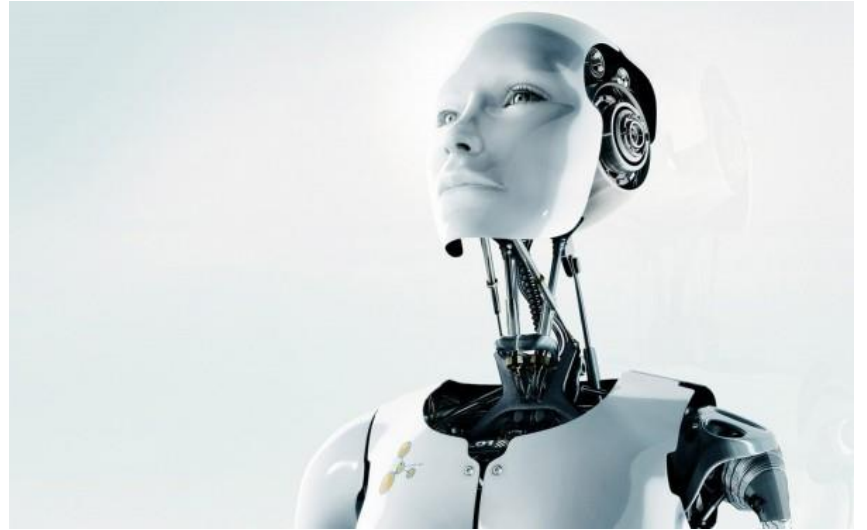
Recently-Learned Facts [twitter](#) [Refresh](#)

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

<Abelardo\_I
<Abraham\_Pais>
<Albert\_Einstein's\_brain>
<Abram\_L.\_Sachar>
<Alfred\_Kleiner>
<Absent-minded\_professor>
<Amnen\_der\_Physik>
<Absorption\_refrigerator>
<Amnis\_Mirabilis\_papers>

CMU Never-Ending Language Learning

YAGO Knowledge Base



# Building a computer that 'understands' text: The NLP pipeline

# Tokenization/Segmentation

- Split text into words and sentences
  - Task: what is the most **likely** segmentation/tokenization?

There was an earthquake near  
D.C. I've even felt it in  
Philadelphia, New York, etc.

There + was + an + earthquake  
+ near + D.C.

I + ve + even + felt + it + in +  
Philadelphia, + New + York, + etc.



# Part-of-Speech tagging

- Marking up a word in a text (corpus) as corresponding to a particular part of speech
  - Task: what is the most **likely** tag sequence

A + dog + is + chasing + a + boy + on + the + playground

A	+	dog	+	is	+	chasing	+	a	+	boy	+	on	+	the	+	playground
Det		Noun		Aux		Verb		Det		Noun		Prep		Det		Noun

# Named entity recognition

- Determine text mapping to proper names
  - Task: what is the most **likely** mapping

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**

# Recap: perplexity

- The inverse of the likelihood of the test set as assigned by the language model, normalized by the number of words

$$PP(w_1, \dots, w_N) = \sqrt[N]{\frac{1}{\prod_{i=1}^N p(w_i | w_{i-1}, \dots, w_{i-n+1})}}$$

↖ N-gram language model

# Recap: an example of NLP

A dog is chasing a boy on the playground.

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase

Complex Verb

Noun Phrase

Noun Phrase

Prep Phrase

Verb Phrase

Verb Phrase

Sentence

**Lexical analysis**  
(part-of-speech tagging)

**Syntactic analysis**  
(Parsing)

A person saying this may be reminding another person to get the dog back...

**Pragmatic analysis**  
(speech act)

**Semantic analysis**

Dog(d1).  
Boy(b1).  
Playground(p1).  
Chasing(d1,b1,p1).

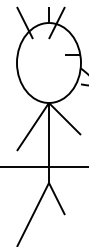
+

Scared(x) if Chasing(\_,x,\_).



Scared(b1)

**Inference**

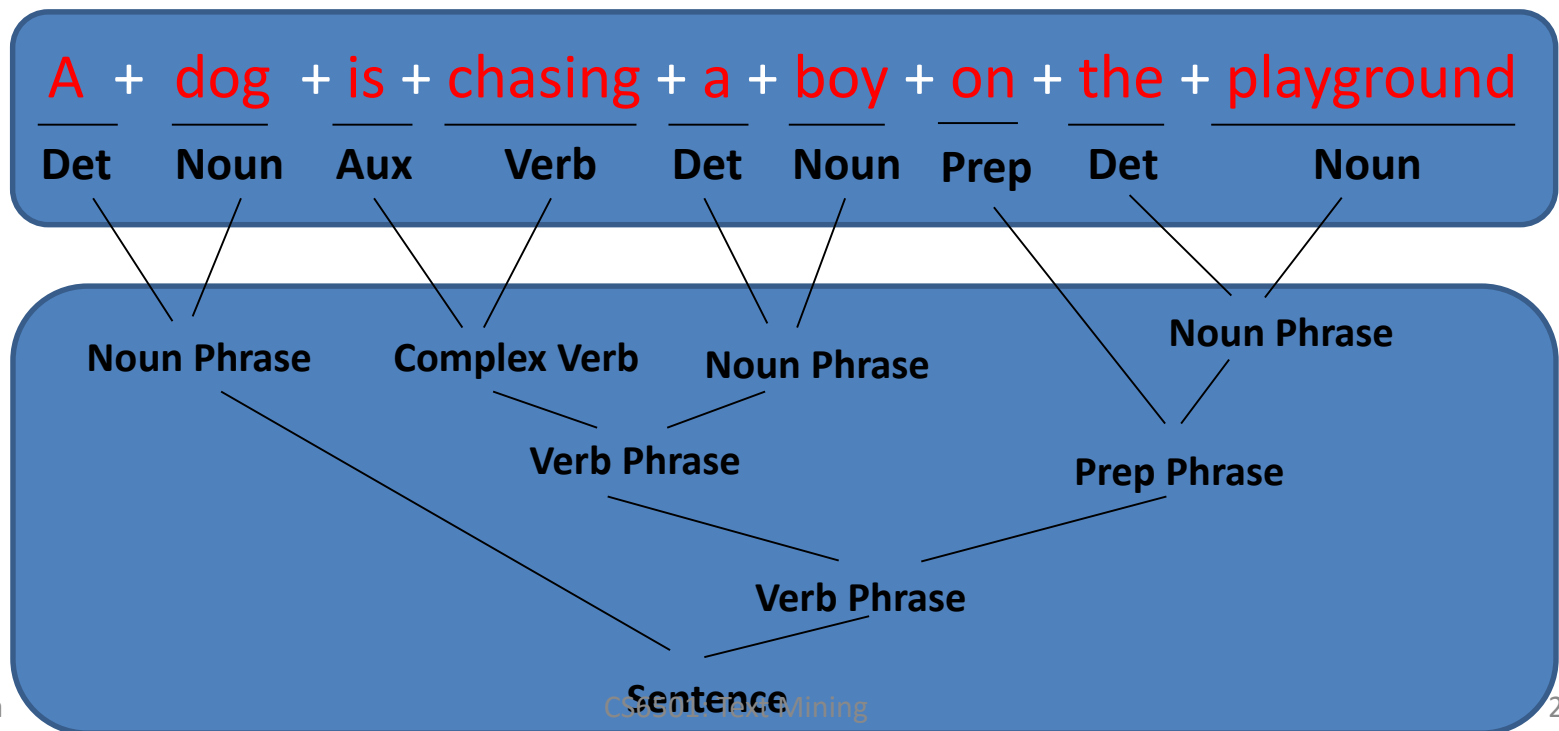


# Recap: NLP is difficult!!!!!!

- 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

# 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.

$\exists x (\text{Is\_Person}(x) \ \& \ \text{Is\_President\_Of}(x, \text{'U.S.'}) \ \& \ \text{Is\_Member\_Of}(x, \text{'Board of Visitors'}))$

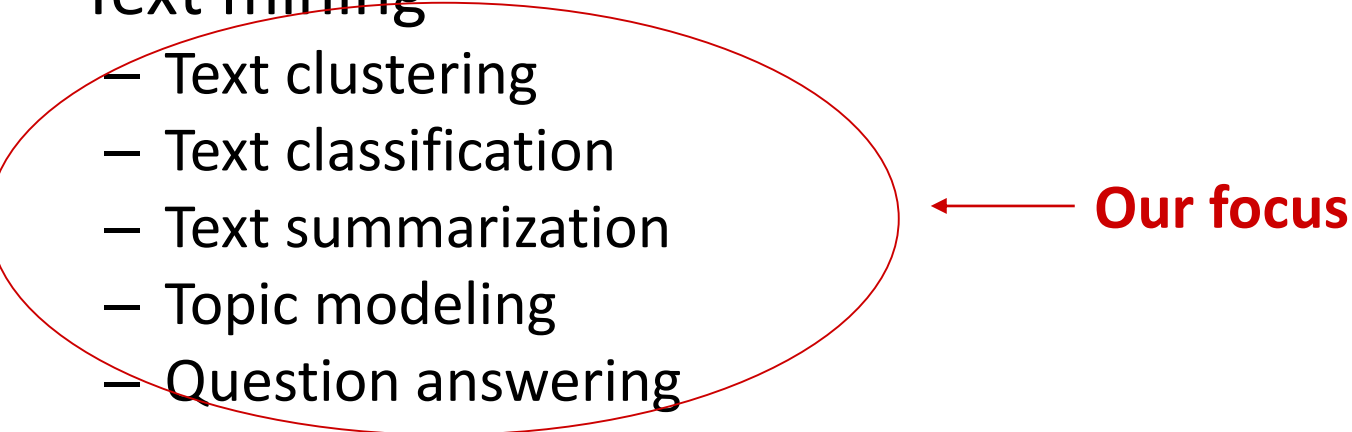


# Towards understanding of text

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.**

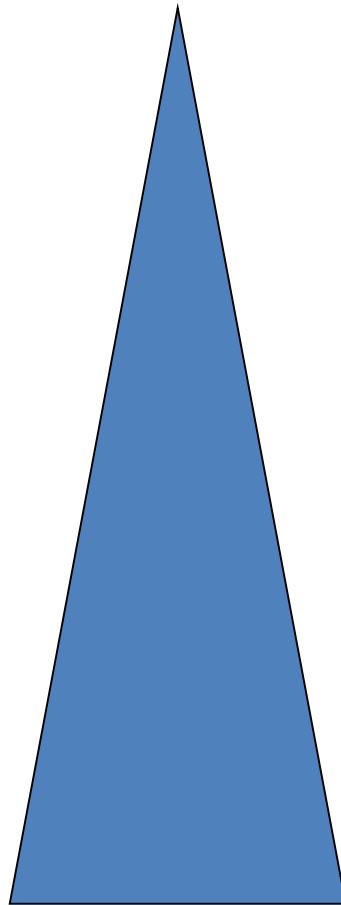
**Errors in NLP can hurt text mining performance...**

# How much NLP is really needed?

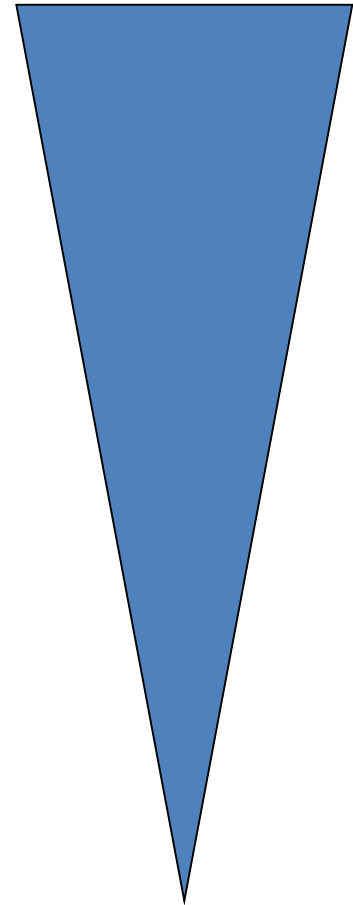
## Tasks

Classification  
Clustering  
Summarization  
Extraction  
Topic modeling  
Translation  
Dialogue  
Question  
Answering  
Inference  
Speech Act

## Dependency on NLP



## Scalability



## 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

# What you should know

- Different levels of NLP
- Challenges in NLP
- NLP pipeline