NLP Training – Session 1

Dr. Alexandra M. Liguori

Incubio – The Big Data Academy

Barcelona, March 11, 2015

Outline

Introduction

1. Name?
2. Background and current activity?
3. Interest in this NLP training?
4. What do you expect from this training?

Introduction: Intelligent machines?

Video:
https://www.youtube.com/watch?v=dSIKBliboIo

(Stanley Kubrick and Arthur C. Clarke,
screenplay of *2001: A Space Odyssey*)

Introduction: Intelligent machines?

Dave Bowman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'm afraid I can't do that.

(Stanley Kubrick and Arthur C. Clarke,
screenplay of *2001: A Space Odyssey*)

https://www.youtube.com/watch?v=dSIKBliboIo

Introduction: Intelligent machines?

Introduction: Intelligent machines?

1. Phonetics and phonology

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*
3. **Syntax** $\rightarrow$ cfr. *Open the pod bay doors, HAL.*
   vs. *HAL, the pod bay door is open.*
   vs. *HAL, is the pod bay door open?*

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*
3. **Syntax** $\rightarrow$ cfr. *Open the pod bay doors, HAL.*
   vs. *HAL, the pod bay door is open.*
   vs. *HAL, is the pod bay door open?*
4. **Lexical semantics** $\rightarrow$ meaning of component words

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*
3. **Syntax** $\rightarrow$ cfr. *Open the pod bay doors, HAL.*
   vs. *HAL, the pod bay door is open.*
   vs. *HAL, is the pod bay door open?*
4. **Lexical semantics** $\rightarrow$ meaning of component words
5. **Compositional semantics** $\rightarrow$ knowledge of how components combine to form larger meanings

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*
3. **Syntax** $\rightarrow$ cfr. *Open the pod bay doors, HAL.*
   vs. *HAL, the pod bay door is open.*
   vs. *HAL, is the pod bay door open?*
4. **Lexical semantics** $\rightarrow$ meaning of component words
5. **Compositional semantics** $\rightarrow$ knowledge of how components combine to form larger meanings
6. **Pragmatics** $\rightarrow$ cfr. *I'm sorry ... , I'm afraid I can't*
   vs. *No, I won't open the door.*
   vs. *No.*

Introduction: Intelligent machines?

1. Phonetics and phonology
2. **Morphology** $\rightarrow$ produce contractions *I'm* and *can't*
3. **Syntax** $\rightarrow$ cfr. *Open the pod bay doors, HAL.*
   vs. *HAL, the pod bay door is open.*
   vs. *HAL, is the pod bay door open?*
4. **Lexical semantics** $\rightarrow$ meaning of component words
5. **Compositional semantics** $\rightarrow$ knowledge of how components combine to form larger meanings
6. **Pragmatics** $\rightarrow$ cfr. *I'm sorry ... , I'm afraid I can't*
   vs. *No, I won't open the door.*
   vs. *No.*
7. **Discourse conventions** $\rightarrow$ engaging in structured conversation using reference *that* in *I'm sorry Dave, I'm afraid I can't do* ***that***

Natural Language Processing

NLP: techniques that process written human language *as language*.

Natural Language Processing

NLP: techniques that process written human language *as language*.

Applications

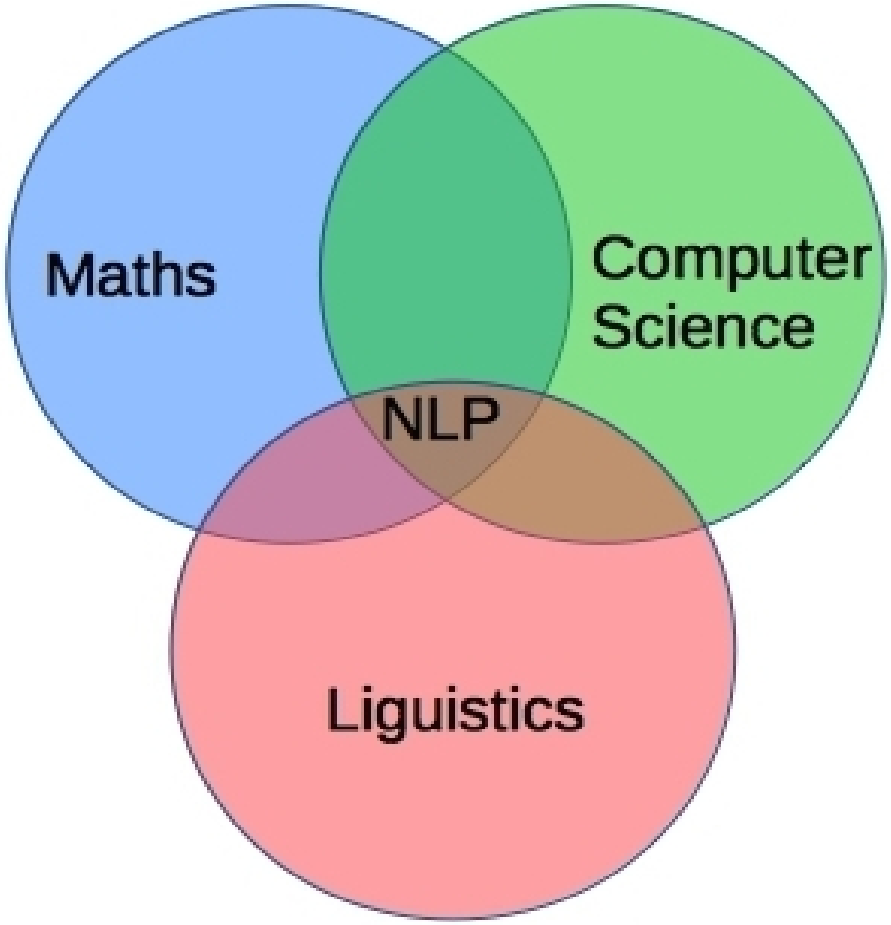
- word counting
- automatic hyphenation
- automated question answering
- named entity extraction (NER)
- information/content extraction
- semantic analysis
- sentiment analysis
- machine translation

Natural Language Processing

NLP: techniques that process written human language *as language*.

Natural Language Processing

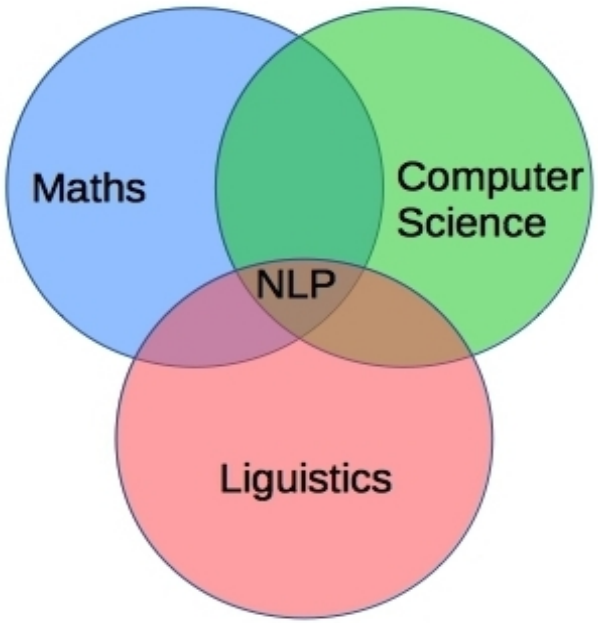
NLP: techniques that process written human language *as language*.



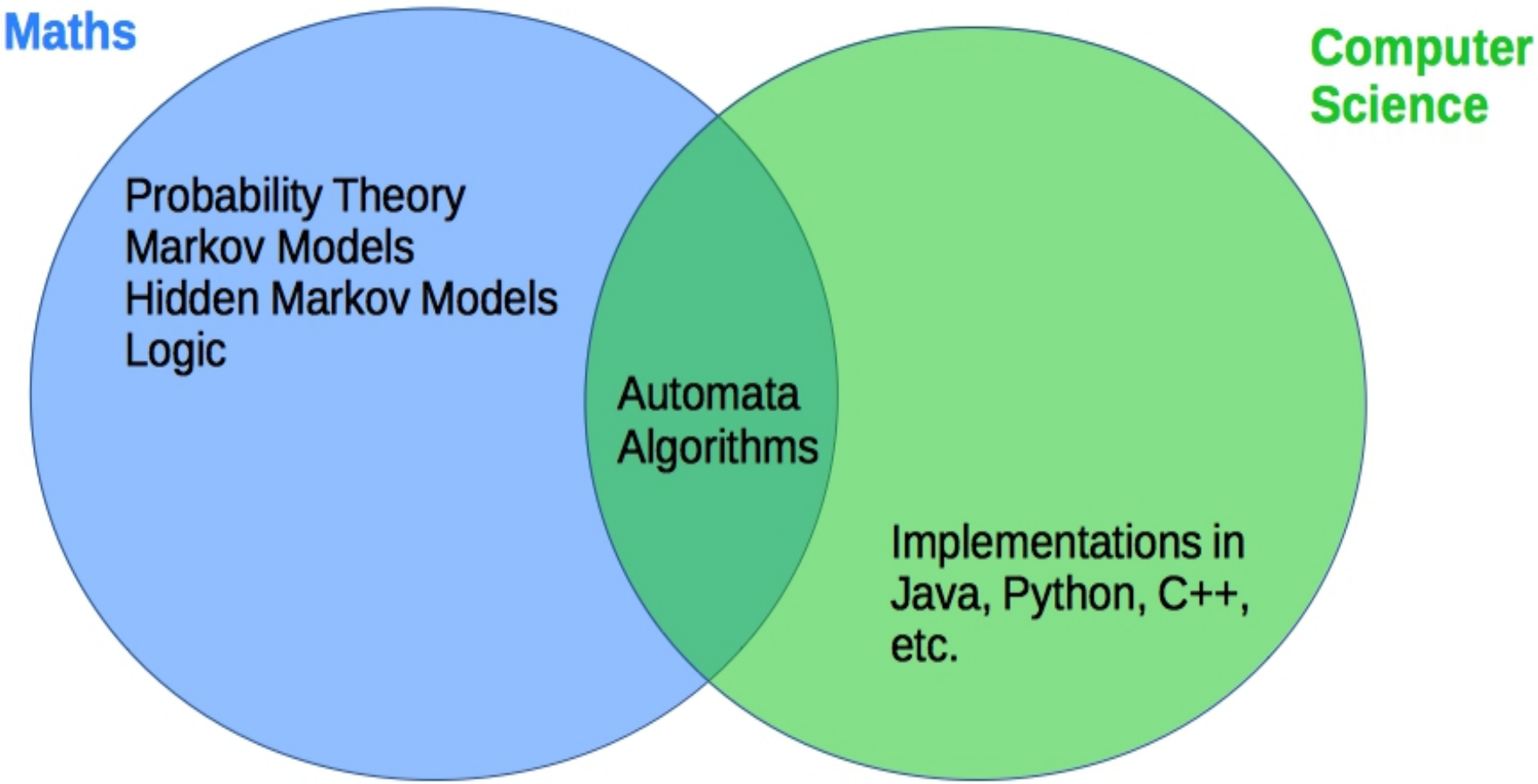
Natural Language Processing

An ideal NLP team is very interdisciplinary, including:

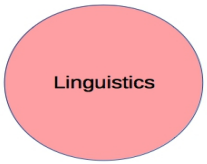
- Language experts (linguists)
- Maths experts (mathematicians, physicists, statisticians)
- Programmers (computer scientists)



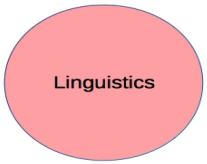
NLP: Maths & Computer Science



NLP: Six categories of linguistic knowledge

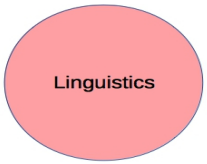


NLP: Six categories of linguistic knowledge



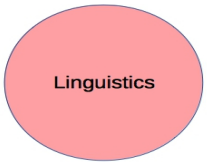
1. **Phonetics and phonology** $\leftrightarrow$ *red* - *read* - *read*; *sleigh* - *slay*

NLP: Six categories of linguistic knowledge



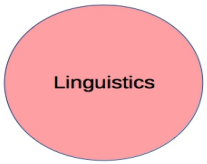
1. **Phonetics and phonology** $\leftrightarrow$ *red* - *read* - *read*; *sleigh* - *slay*
2. **Morphology** $\leftrightarrow$ *I/you/we/you/they walk_* - *he/she/it walks*; *walked*; *walking*

NLP: Six categories of linguistic knowledge



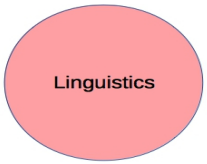
1. **Phonetics and phonology** $\leftrightarrow$ *red - read - read*; *sleigh - slay*
2. **Morphology** $\leftrightarrow$ *I/you/we/you/they walk_ - he/she/it walks*; *walked*; *walking*
3. **Syntax** $\leftrightarrow$ *She ate a mammoth breakfast* - *She eating a mammoth breakfast*

NLP: Six categories of linguistic knowledge



1. **Phonetics and phonology** $\leftrightarrow$ *red - read - read*; *sleigh - slay*
2. **Morphology** $\leftrightarrow$ *I/you/we/you/they walk_ - he/she/it walks*; *walked*; *walking*
3. **Syntax** $\leftrightarrow$ *She ate a mammoth breakfast* - *She eating a mammoth breakfast*
4. **Semantics** $\leftrightarrow$ *book (verb) - book (noun)*; *duck (verb) - duck (noun)*

NLP: Six categories of linguistic knowledge



1. **Phonetics and phonology** $\leftrightarrow$ *red - read - read*; *sleigh - slay*
2. **Morphology** $\leftrightarrow$ *I/you/we/you/they walk_ - he/she/it walks*; *walked*; *walking*
3. **Syntax** $\leftrightarrow$ *She ate a mammoth breakfast* - *She eating a mammoth breakfast*
4. **Semantics** $\leftrightarrow$ *book (verb) - book (noun)*; *duck (verb) - duck (noun)*
5. **Pragmatics** $\leftrightarrow$ *open the door - can you open the door? - could you open the door, please?*

NLP: Six categories of linguistic knowledge

- **Discourse**

NLP: Six categories of linguistic knowledge

- **Discourse** 

  Gracie: Oh yeah... And then Mr. and Mrs. Jones were having matrimonial trouble, and my brother was hired to watch Mrs. Jones.
  George: Well, I imagine she was a very attractive woman.
  Gracie: She was, and my brother watched her day and night for six months.
  George: Well, what happened?
  Gracie: She finally got a divorce.
  George: Mrs. Jones?
  Gracie: No, my brother's wife.

NLP: Six categories of linguistic knowledge

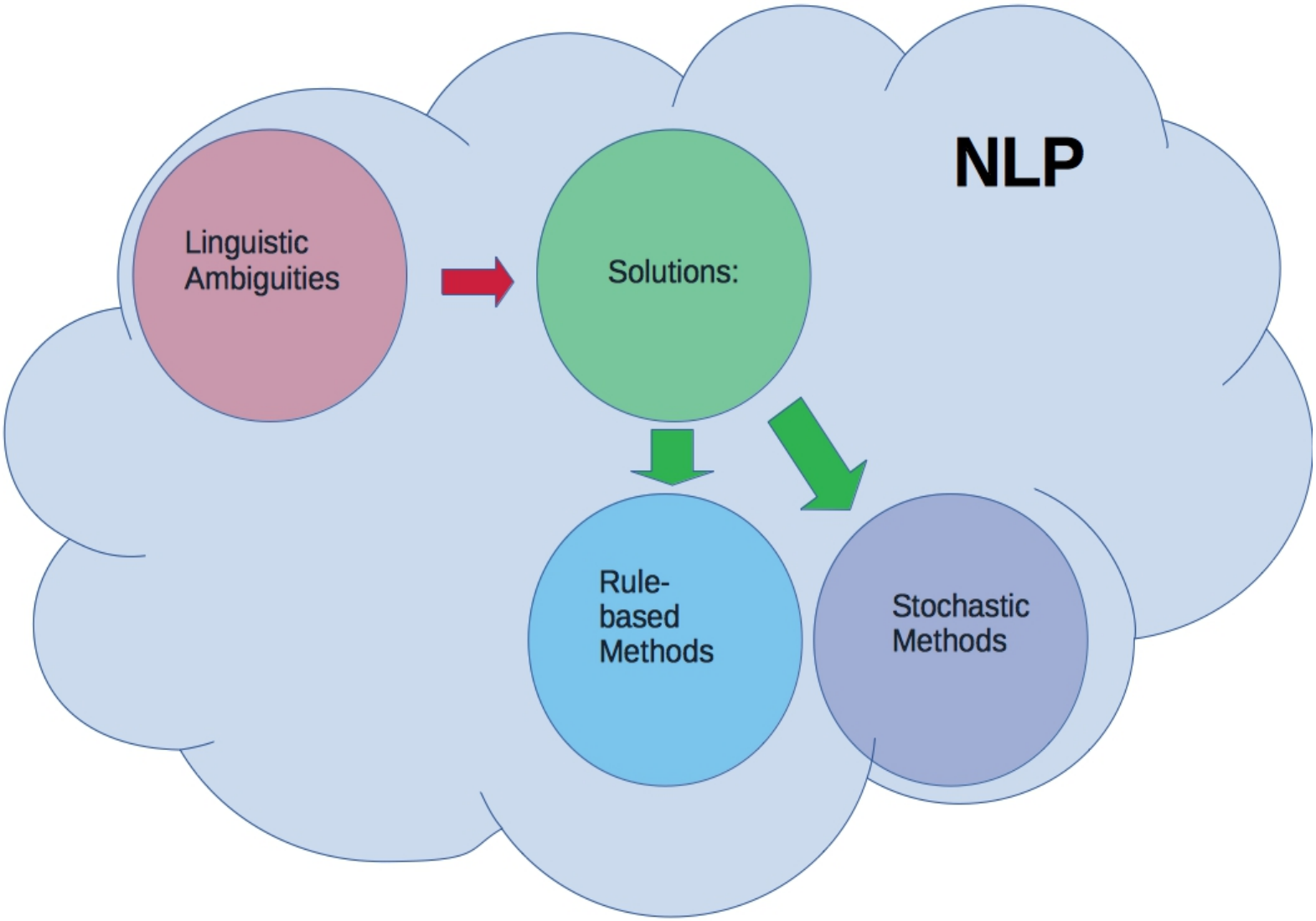
- **Discourse** 

  Gracie: Oh yeah... And then Mr. and Mrs. Jones were having matrimonial trouble, and my brother was hired to watch Mrs. Jones.
  George: Well, I imagine she was a very attractive woman.
  Gracie: She was, and my brother watched her day and night for six months.
  George: Well, what happened?
  Gracie: She finally got a divorce.
  George: Mrs. Jones?
  Gracie: No, my brother's wife.

  John went to Bill's car dealership to check out an Acura Integra. He looked at it for about an hour.

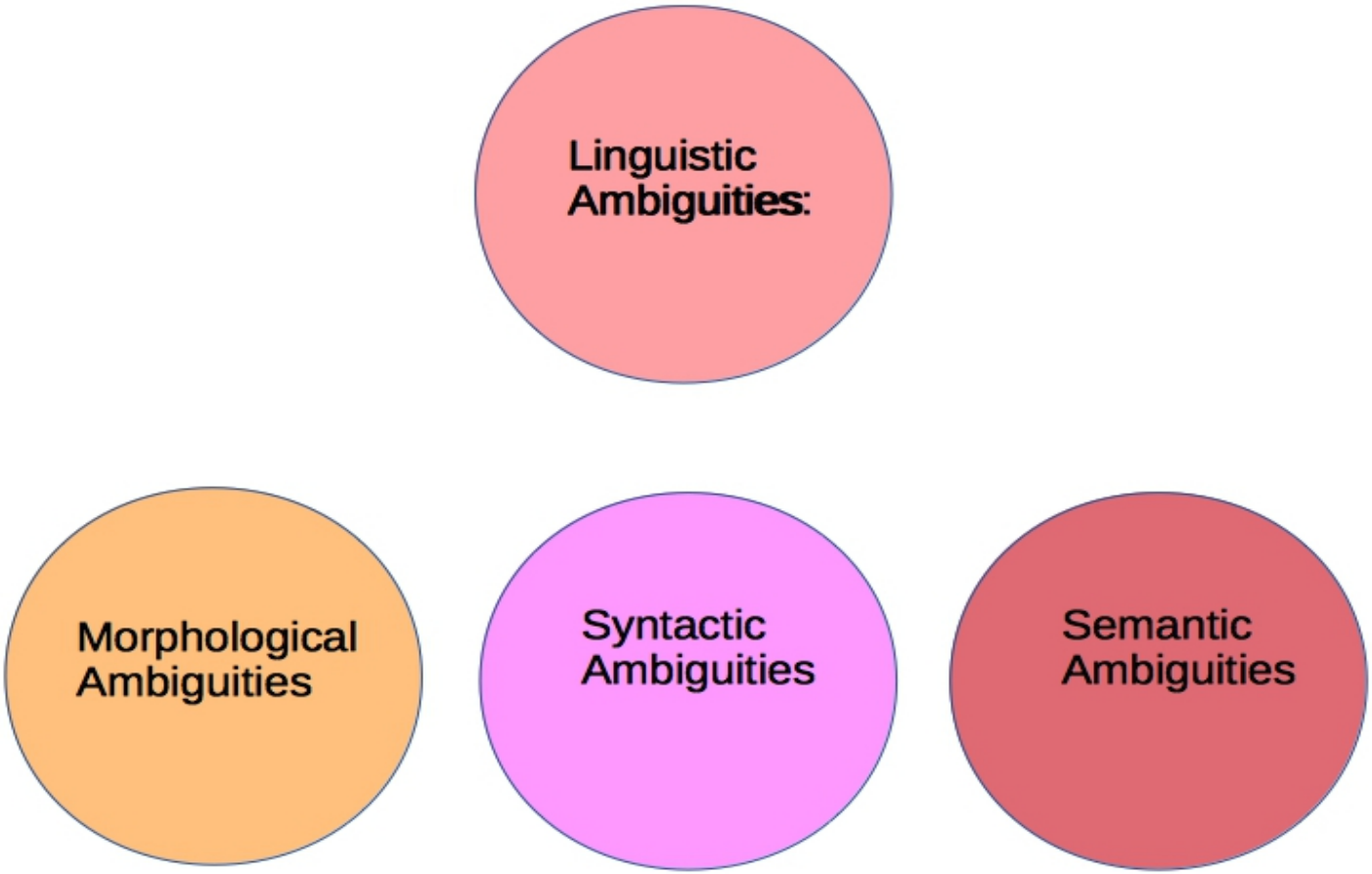
NLP: Ambiguities and Solutions

NLP: Ambiguities and Solutions



Linguistic Ambiguities

Linguistic Ambiguities



Linguistic Ambiguities

Linguistic Ambiguities

Example

I made her duck.

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

1. I cooked waterfowl for her.

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

1. I cooked waterfowl for her.
2. I cooked waterfowl belonging to her.

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

1. I cooked waterfowl for her.
2. I cooked waterfowl belonging to her.
3. I created the (plaster?) duck she owns.

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

1. I cooked waterfowl for her.
2. I cooked waterfowl belonging to her.
3. I created the (plaster?) duck she owns.
4. I caused her to quickly lower her head or body.

Linguistic Ambiguities

Example

I made her duck.

Five possible interpretations:

1. I cooked waterfowl for her.
2. I cooked waterfowl belonging to her.
3. I created the (plaster?) duck she owns.
4. I caused her to quickly lower her head or body.
5. I waved my magic wand and turned her into undifferentiated waterfowl.

Linguistic Ambiguities

Linguistic Ambiguities

Morphological ambiguity

- *duck*: verb or noun
- *her*: dative pronoun or possessive pronoun

Linguistic Ambiguities

Morphological ambiguity

- *duck*: verb or noun
- *her*: dative pronoun or possessive pronoun

Syntactic ambiguity: make

- transitive: taking a single direct object (case 2)
- ditransitive: taking two objects, meaning that the first object (*her*) got made into the second object (*duck*)
- taking a direct object and a verb, meaning that the object (*her*) got caused to perform the verbal action (*duck*)

Linguistic Ambiguities

Morphological ambiguity

- *duck*: verb or noun
- *her*: dative pronoun or possessive pronoun

Syntactic ambiguity: make

- transitive: taking a single direct object (case 2)
- ditransitive: taking two objects, meaning that the first object (*her*) got made into the second object (*duck*)
- taking a direct object and a verb, meaning that the object (*her*) got caused to perform the verbal action (*duck*)

Semantic ambiguity: make

- *cook*
- *create*

Typical NLP tasks: Basic and simpler tasks

Typical NLP tasks: Basic and simpler tasks

Tokenization

Typical NLP tasks: Basic and simpler tasks

Tokenization

RegEx

Typical NLP tasks: Basic and simpler tasks

Tokenization

Sentence splitting

RegEx

Typical NLP tasks: Basic and simpler tasks

Tokenization	RegEx
Sentence splitting	RegEx

Typical NLP tasks: Basic and simpler tasks

Tokenization	RegEx
Sentence splitting	RegEx
POS-tagging	

Typical NLP tasks: Basic and simpler tasks

Tokenization	RegEx
Sentence splitting	RegEx
POS-tagging	POS-tagging algorithms and tag sets

Typical NLP tasks: Complex tasks

Typical NLP tasks: Complex tasks

Lemmatization or Stemming

Typical NLP tasks: Complex tasks

Lemmatization or Stemming

Implementations of Porter Stemmer (e.g. in Java), Stanford NLP tool, GATE, ...

Typical NLP tasks: Complex tasks

Lemmatization or Stemming

Implementations of Porter Stemmer (e.g. in Java), Stanford NLP tool, GATE, ...

Syntactic parsing

Typical NLP tasks: Complex tasks

Lemmatization or Stemming	Implementations of Porter Stemmer (e.g. in Java), Stanford NLP tool, GATE, ...
Syntactic parsing	Early algorithm, CYK algorithm, GHR algorithm, Stanford Parser (Java implementation of probabilistic algorithm)

Typical NLP tasks: Complex tasks

Lemmatization or Stemming

Implementations of Porter Stemmer (e.g. in Java), Stanford NLP tool, GATE, ...

Syntactic parsing

Early algorithm, CYK algorithm, GHR algorithm, Stanford Parser (Java implementation of probabilistic algorithm)

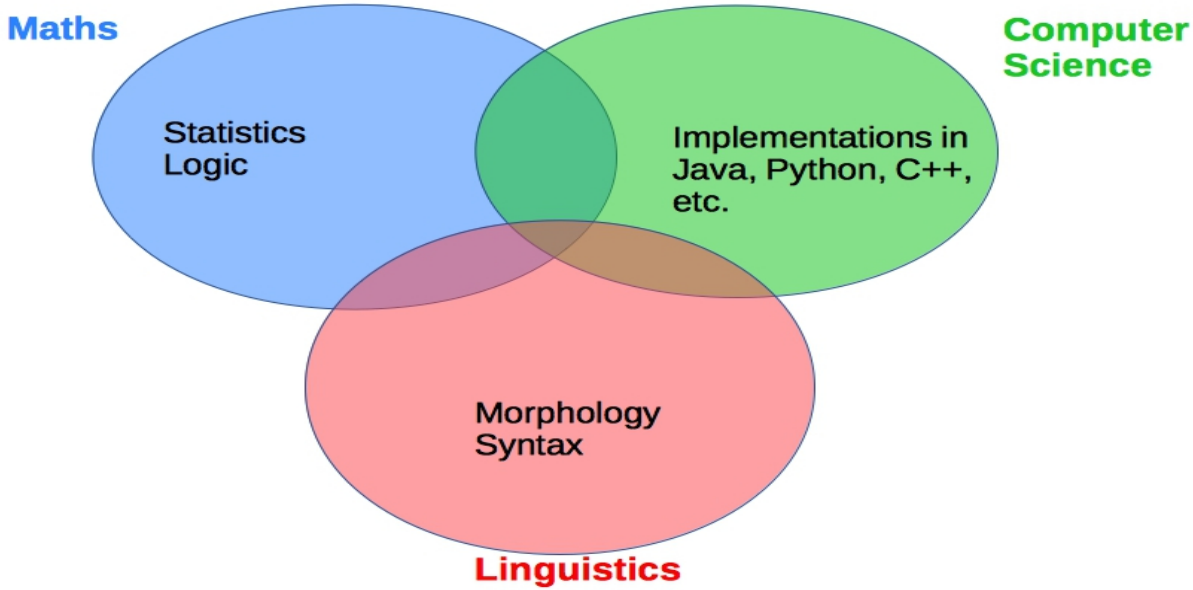
- Topic extraction
- NER
- Semantic analysis
- ...

Typical NLP tasks: Complex tasks

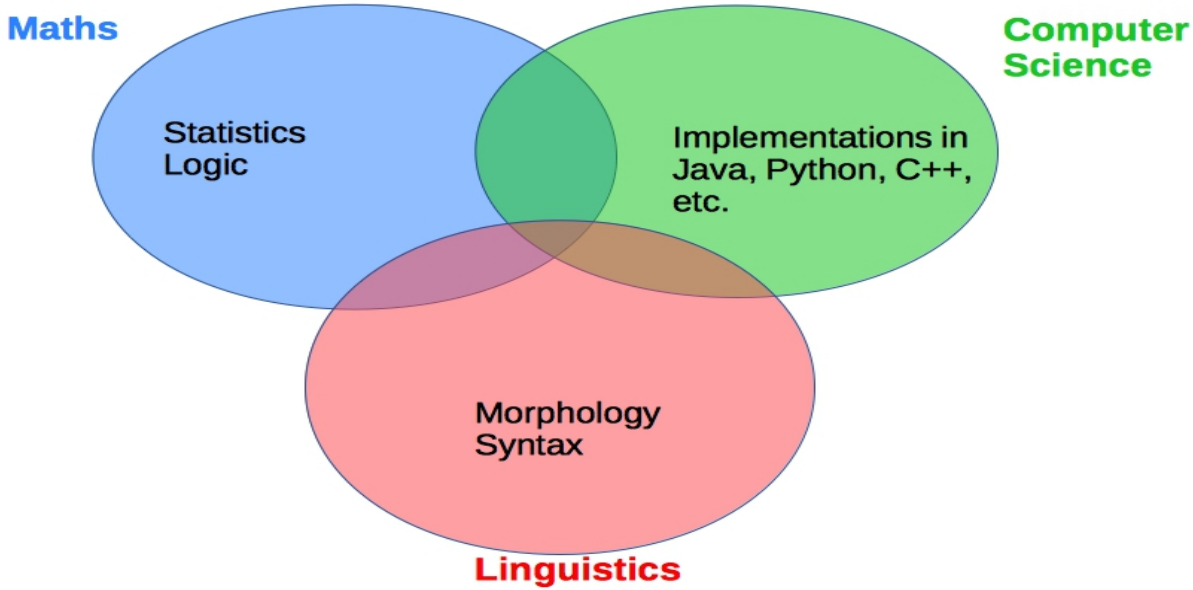
Lemmatization or Stemming	Implementations of Porter Stemmer (e.g. in Java), Stanford NLP tool, GATE, ...
Syntactic parsing	Early algorithm, CYK algorithm, GHR algorithm, Stanford Parser (Java implementation of probabilistic algorithm)
• Topic extraction • NER • Semantic analysis • ...	Ad hoc tools, e.g. dictionaries, ontologies, Frames, GATE, NLTK, ...

POS-tagging

POS-tagging


Maths
Statistics
Logic
Computer
Science
Implementations in
Java, Python, C++,
etc.
Morphology
Syntax
Linguistics

POS-tagging



Example with Penn Treebank POS-tags:

A/**DT** woman/**NN** came/**VBD** from/**IN** the/**DT** back/**NN** of/**IN** the/**DT** store/**NN** ./**.** She/**PP** appeared/**VBD** to/**TO** be/**VB** sleepy/**JJ** and/**CC** quite/**RB** a/**DT** bit/**NN** younger/**JJR** than/**IN** Mr./**NNP** Dobbs/**NNP** and/**CC** to/**TO** be/**VB** wearing/**VBG** too/**RB** much/**RB** makeup/**NN** ./**.**

POS-tagging

Example of ambiguity:

POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**

POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**

2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

POS-tagging

Three main tagging algorithms or methods:

1. rule-based tagging, e.g. ENGTWOL
2. stochastic tagging, e.g. HMM tagger
3. transformation-based tagging, e.g. Brill tagger

Rule-based POS-tagging

Rule-based POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

Rule-based POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

Large database of hand-written disambiguation rules, e.g.:

Rule-based POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

Large database of hand-written disambiguation rules, e.g.:

- TO + VB $\rightarrow$ YES
- TO + NN $\rightarrow$ NO
- DT + NN $\rightarrow$ YES
- DT + VB $\rightarrow$ NO

Stochastic POS-tagging

Stochastic POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

Stochastic POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./**.**

Training corpus to compute probability of given word having given tag in given context, e.g.:

Stochastic POS-tagging

Example of ambiguity:

1. Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./.
2. People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN** ./.

Training corpus to compute probability of given word having given tag in given context, e.g.:

- is/**VBZ** expected/**VBN** to/**TO** race/**VB** $\rightarrow 98\%$
- is/**VBZ** expected/**VBN** to/**TO** race/**NN** $\rightarrow 2\%$
- reason/**NN** for/**IN** the/**DT** race/**NN** $\rightarrow 97\%$
- reason/**NN** for/**IN** the/**DT** race/**VB** $\rightarrow 3\%$

Transformation-based tagging POS-tagging

Transformation-based tagging POS-tagging

Example of ambiguity:

- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
- People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN**

Transformation-based tagging POS-tagging

Example of ambiguity:

- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
- People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN**

Rules automatically induced from data using Machine Learning techniques, e.g.:

Transformation-based tagging POS-tagging

Example of ambiguity:

- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./**.**
- People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN**

Rules automatically induced from data using Machine Learning techniques, e.g.:

1. a priori, prob(race $=$ **NN**)$=65\%$, prob(race $=$ **VB**)$=35\%$
   $\rightarrow$ system would always take race $=$ **NN**

Transformation-based tagging POS-tagging

Example of ambiguity:

- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./.
- People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN**

Rules automatically induced from data using Machine Learning techniques, e.g.:

1. a priori, prob(race $=$ **NN**)$= 65\%$, prob(race $=$ **VB**)$= 35\%$ $\rightarrow$ system would always take race $=$ **NN**
2. apply Machine Learning techniques and learn the conditional probabilities:

Transformation-based tagging POS-tagging

Example of ambiguity:

- Secretariat/**NNP** is/**VBZ** expected/**VBN** to/**TO** race/**VB** tomorrow/**NN** ./.
- People/**NNS** continue/**VBP** to/**TO** inquire/**VB** the/**DT** reason/**NN** for/**IN** the/**DT** race/**NN** for/**IN** outer/**JJ** space/**NN**

Rules automatically induced from data using Machine Learning techniques, e.g.:

1. a priori, prob(race = **NN**)= 65%, prob(race = **VB**)= 35%
   $\rightarrow$ system would always take race = **NN**
2. apply Machine Learning techniques and learn the conditional probabilities:
3. is/**VBZ** expected/**VBN** to/**TO** race/**VB** $\rightarrow$ 98%
   reason/**NN** for/**IN** the/**DT** race/**NN** $\rightarrow$ 97%

POS-tagging

POS-tagging tools for English:

- Brill tagger
- Stanford Log-linear POS-tagger (Java)
- POS-tagger integrated in GATE (Java)
- POS-tagger with NLTK (Python)

What next?

Topics for the next sessions:

1. Semantic analysis
2. Question answering
3. Reference resolution
4. Named Entity Recognition (NER)
5. Keyword / topic / information extraction