Introduction

Razvan C. Bunescu
Department of Computer Science @ CCI

razvan.bunescu@uncc.edu
What is Natural Language Processing?

- **Natural Language Processing (NLP)** = developing computer systems that can *process, understand, or communicate* in natural language (*text* or *speech*):
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - **Natural Languages** are significantly more difficult to process than Artificial Languages!

- **What about Computational Linguistics (CL)?**
  - Computational Linguistics is focused on the *study of language*, using computational tools.
  - NLP is focused on solving language tasks such as *machine translation*, *information extraction*, *question answering*, *taking instructions*, *holding conversations*, …
NLP Application: Question Answering

• Input:
  – A question:
    • What is the meaning of life?
  – A large collection of text documents:
    • all books from UNCC Library.

• Output:
  – An answer, or list of answers.
    • Found by ‘mining’ the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)
NLP Application: Question Answering

• Input:
  – A question:
    • What is the meaning of life?
  – A large collection of text documents:
    • all books from UNCC Library.

• Output:
  – An answer, or list of answers.
    • Found by ‘mining’ the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)
NLP Application: Question Answering

• Input:
  – A question:
    
    Try simple pattern matching: “the meaning of life is <?>”

• Output:
  – An answer, or list of answers.
    • Found by ‘mining’ the documents in the collection.

What is the meaning of life?

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

Tomorrow, and tomorrow, and tomorrow, Creeps in this petty pace from day to day, To the last syllable of recorded time; And all our yesterdays have lighted fools The way to dusty death. Out, out, brief candle! Life's but a walking shadow, a poor player That struts and frets his hour upon the stage And then is heard no more. It is a tale Told by an idiot, full of sound and fury Signifying nothing.
NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*
  - A large collection of text documents:
    - *all books from OU Library.*

- Output:
  - An answer, or list of answers.
    - Found by ‘mining’ the documents in the collection.

**Word Sense Disambiguation:**

*meaning ≃ signifying*

**What is the meaning of life?**

_Tomorrow, and tomorrow, and tomorrow,_  
_Creeps in this petty pace from day to day,_  
_To the last syllable of recorded time;_  
_And all our yesterdays have lighted fools_  
_The way to dusty death. Out, out, brief candle!_  
_Life's but a walking shadow, a poor player_  
_That struts and frets his hour upon the stage_  
_And then is heard no more. It is a tale_  
_Told by an idiot, full of sound and fury_  
_Signifying nothing._

_Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)
NLP Application: Question Answering

• Input:
  – A question:
    • What is the meaning of life?
  – A large collection of text documents:
    • all books from UNCC Library.

• Output:
  – An answer, or list of answers.
    • Found by ‘mining’ the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)
What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

Coreference Resolution:
{Life, it, tale} are coreferent.
What is the **meaning** of **life**?

*Tomorrow, and tomorrow, and tomorrow,*  
*Creeps in this petty pace from day to day,*  
*To the last syllable of recorded time;*  
*And all our yesterdays have lighted fools*  
The way to dusty death. Out, out, brief candle!  
*Life's but a walking shadow, a poor player*  
*That struts and frets his hour upon the stage*  
*And then is heard no more. It is a tale*  
*Told by an idiot, full of sound and fury*  
**Signifying nothing.**

Shakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)
NLP Application: Question Answering

What is the **meaning** of **life**?

*Tomorrow, and tomorrow, and tomorrow,* 
*Creeps in this petty pace from day to day,* 
*To the last syllable of recorded time;* 
*And all our yesterdays have lighted fools* 
The way to dusty death. Out, out, brief candle!  
**Life**'s but a walking shadow, a poor player 
That struts and frets his hour upon the stage 
And then is heard no more. **It is a tale**  
Told by an idiot, full of sound and fury  
**Signifying nothing.**

Skakespeare’s Macbeth (Act 5, Scene 5, lines 17-28)

Fundamental NLP tasks:

- **Tokenization.**
- **Syntactic Analysis:**
  - Part of Speech Tagging.
  - Dependency Parsing.
- **Word Sense Disambiguation.**
- **Coreference Resolution.**
- Semantic Role Labeling.
- Semantic Parsing.
Fundamental NLP Tasks in Text Analysis

- Tokenization
- Morphological Analysis
- Part of Speech Tagging
- Syntactic Parsing
- Word Sense Disambiguation
- Semantic Role Labeling
- Semantic Parsing
- Anaphora/Coreference Resolution
Tokenization

- **Tokenization** = segmenting text into words and sentences.
  - A crucial first step in most text processing applications.
  - Recent SoA models use *subword* tokenization.

- Whitespace indicative of word boundaries?
  - Yes: English, French, Spanish, …
  - No: Chinese, Japanese, Thai, …

- Whitespace is not enough:
  - ‘What’re you? Crazy?’ said Sadowsky. ‘I can’t afford to do that.’

  ⇒ ‘what’re you? crazy? Sadowsky. ‘I can’t that.’
Word Segmentation

• In English, characters other than whitespace can be used to separate words:
  – , ; . - : ()”

• But punctuation often occurs inside words:
  – m.p.h., Ph.D., AT&T, 01/02/06, google.com, 62.5
  – Homework: design regular expressions to match constructions where punctuation does not split:
    – acronyms, dates, web addresses, numbers, etc.
    – https://docs.python.org/3/howto/regex.html

• Expansion of clitic constructions:
  – he’s happy ⇒ he is happy
  – Need ambiguity resolution between clitic construction, possessive markers, quotative markers:
    • he’s happy vs. the book’s cover vs. ‘what are you? crazy?’
Sentence Segmentation

• Generally based on punctuation marks: ? ! .
  – Periods are ambiguous, as sentence boundary markers and abbreviation/acronym markers:
    • Mr., Inc., m.p.h.
  – Sometimes they mark both:
    • SAN FRANCISCO (MarketWatch) – Technology stocks were mostly in positive territory on Monday, powered by gains in shares of Microsoft Corp. and IBM Corp.

• Tokenization approaches:
  – Regular Expressions.
  – Machine Learning (state of the art).
**Morphology** = the field of linguistics that studies the internal structure of words.

- **Morpheme** is the smallest linguistic unit that has semantic meaning:
  - **stems**: “carry”, “depend”, “Google”, “lock”
  - **affixes**: “pre”, “ed”, “ly”, “s”

**Morphological analysis** = segmenting words into morphemes:

- carried ⇒ carry + ed (past tense)
- independently ⇒ in + (depend + ent) + ly
- Googlers ⇒ (Google + er) + s (plural)
- unlockable ⇒ un + (lock + able) ? (un + lock) + able ?
Morphological Analysis: Stemming

• In IR applications such as Web search, useful to know if two words have the same stem:
  – Boolean Query: “marsupial OR kangaroo OR koala”.
  ⇒ stemming, i.e. given a word, extract the stem:
    • marsupials => marsupial
    • played, playing, player, plays => play

• Porter stemmer – a series of simple cascaded rewrite rules:
  – ATIONAL => ATE (e.g. relational => relate)
  – ING => e (e.g. motoring => motor)
  – SSES => SS (e.g. grasses => grass)
Part of Speech (POS) Tagging

• Annotate each word in a sentence with its POS:
  – nouns, verbs, adjectives, adverbs, pronouns, prepositions, …

  PRP  VBD  TO  VB  TO  DT  NN  IN  NN  VBD  VBG

  They used to object to the use of object oriented programming

  obJECT  OBject

• Useful for many NLP tasks downstream:
  – speech recognition and synthesis, syntactic parsing, word sense disambiguation, information retrieval, …

• Nowadays superseded in many tasks by (contextualized) word embeddings.
Syntactic Analysis

- Compute the *phrase structure* of a sentence:

- Corresponding *dependency structure*:
Word Sense Disambiguation

- Words in natural language may have multiple meanings:
  - he cashed a check at the bank
  - he sat on the bank of the river and watched the currents
  - they built a large plant to manufacture automobiles
  - chlorophyll is generally present in plant leaves

- Use lexical resources such as WordNet that map words to their meanings.

- Identifying the meaning of a word is useful for:
  - machine translation, information retrieval, question answering, text classification, …

- Nowadays superseded in many tasks by (contextualized) word embeddings.
Semantic Role Labeling

• For each clause, determine the semantic role played by each noun phrase that is an argument to the verb:
  - agent  patient  source  destination  instrument
  - John drove Mary from Charlotte to Asheville in his Honda Accord.
  - The hammer broke the window.

• Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”.

Semantic Role Labeling