Natural Language Processing

Class 01: - Introduction
- Non-randomness of language

VRIJE UNIVERSITEIT BRUSSEL

16 February 2018
Dr. Katrien Beuls

Natural Language Processing

- Lecturer: Katrien Beuls (katrien.beuls@vub.be)
- Assistents: Jens Nevens
  Paul Van Eecke
- Guest lectures by:
  - Luc Steels (Francqui chair)
- Classes:
  - Fridays, **4pm-6pm**
  - Room PL9 3.31
- Slides: At the start of each class on PointCarré
- Evaluation
  - Three assignments during semester (50%)
  - Final written exam (50%)
- Check PointCarré regularly for up-to-date info!

2

Course materials

- Handbook:
  Daniel Jurafsky and James H. Martin (2006, 2nd edition). Speech and language processing. Prentice Hall.
  Third edition (draft):
  https://web.stanford.edu/~jurafsky/slp3/
- Slides: At the start of each class on PointCarré
- Articles: As PDFs on PointCarré

- Recommended for self-study
  - Christopher Manning and Hinrich Schütze (1999). *Foundations of statistical natural language processing*. MIT Press.
  - Steven Bird, Ewan Klein and Edward Loper (2009). *Natural language processing with Python*. O'reilly. +free software: *the NLP toolkit* (www.nltk.org/book/)

3

What is Natural Language Processing?

VRIJE UNIVERSITEIT BRUSSEL

- Can we build useful applications that involve natural language input and output?
- Not necessarily by mimicking the brain, but rather through practical **engineering**.

Natural Language Understanding

vs

Understanding Natural Language

- Can we understand how the human mind as a biological mechanism is able to process language?
- Computational modelling as part of **Cognitive Science**

5

- Can we build useful applications that involve natural language input and output?
- Not necessarily by mimicking the human brain, but rather through practical **engineering**.

Natural Language Understanding

vs

Understanding Natural Language

- Can we understand how the human mind as a biological mechanism is able to process language?
- Computational modelling as part of Cognitive Science

6

- Can we build useful applications that involve natural language input and output?
- Not necessarily by mimicking the human brain, but rather through practical **engineering**.

Natural Language Understanding

Some examples of applications...

7

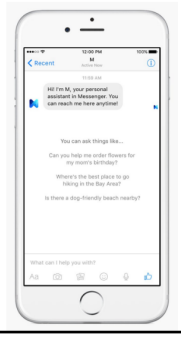
February 2010. Jeopardy!
IBM Watson beats best human player

https://www.youtube.com/watch?v=P18EdAKuC1U

8

Virtual assistants based on NLP

(https://www.wired.com/2015/08/facebook-launches-m-new-kind-virtual-assistant/)



Content analytics on big data

http://www.predictiveanalyticstoday.com/election-tracker-big-data-content-analytics-opentext/

Chatbots relying on NLP

Zuckerberg announces Messenger Platform **chatbots**

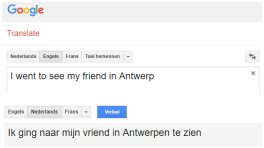
The chief executive of Facebook, Mark Zuckerberg, said outside companies would be able to use **chatbots** on Facebook's Messenger app. (NYTimes April 12, 2016)

Banks Bet on the Next Big Thing: Financial **Chatbots**

really wants to chat. This week Bank of America, MasterCard and several financial start-ups announced new tools — known as **chatbots** — that will allow customers to ask questions about their financial accounts, initiate (NYTimes October 26, 2016)

Machine translation

Speech-to-speech machine translation

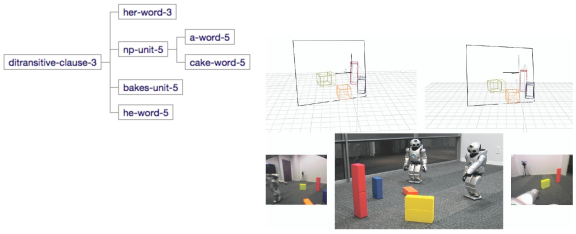


Switch from statistical MT to neural MT

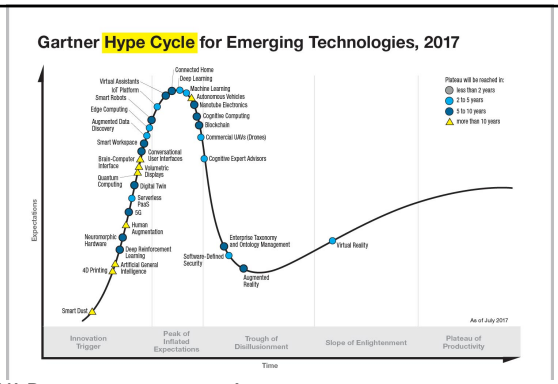
Going **Neural** (NYTimes December 14, 2016)
Apparently **Google Translate**, the company's popular machine-translation service,had suddenly and almost immeasurably improved.

Interactive Language Learning by Robots

Fluid Construction Grammar



https://www.fcg-net.org/



NLP start-ups overviews:

https://www.crunchbase.com/category/natural-language-processing/789bbbefc46e1532a68df17da87090ea
https://angel.co/natural-language-processing

Why is NLP so difficult?

VRIJE UNIVERSITEIT BRUSSEL

Three main problems

Ambiguity
Ambiguity
Ambiguity

Rules, but many exceptions
No clear understanding of how humans process language

Examples of ambiguity

Crash Blossoms
Headlines gone wrong

Word sense ambiguity

I made her duck.

- I cooked for her (duck=food, her=beneficiary)
- I cooked the duck that belonged to her for dinner (her=owner)
- I caused her to move downwards (duck=action)
- I created a duck for her (duck=artefact)
- I magically turned her into a duck (made=spell)

Ambiguity is pervasive

I caused her to quickly lower her head or body.
- Lexical category: "duck" can be a N or V

I cooked waterfowl belonging to her.
- Lexical category: "her" can be a possessive ("of her") or dative ("for her") pronoun

I made the (plaster) duck statue she owns.
- Lexical semantics: "make" can mean "create" or "cook"

Other difficulties

non-standard English
Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either

segmentation issues
the New York-New Haven Railroad
the New York-New Haven Railroad

idioms
dark horse
get cold feet
lose face
throw in the towel

neologisms
unfriend
Retweet
bromance

world knowledge
Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names
Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

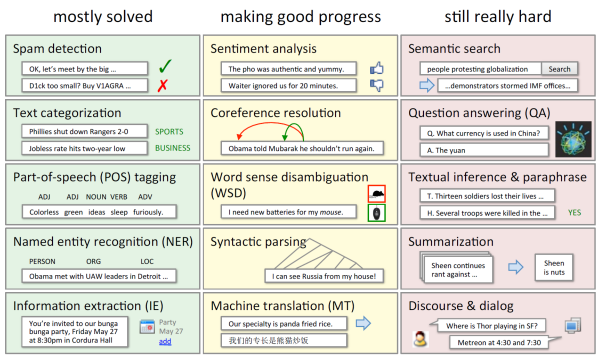
Three laws of computational linguistics

Hugo Brandt Corstius (1978)

1. Whatever you do, semantics will screw things up.
2. Every theory, no matter how explicit it is formulated, will turn out to contain errors when you make a program of it.
3. The first 80% of accuracy takes litte effort, but further diminishing the gap by half takes double the effort of prevous work. (~law of diminished returns)

What is the state-of-the-art?

VRIJE UNIVERSITEIT BRUSSEL



https://www.aclweb.org/aclwiki/index.php?title=State_of_the_art

Many challenges ahead…

- "By ignoring […] finer details, our language-processing systems have been stuck in an "idiot savant" stage where they can find everything but cannot understand anything. The main language processing challenge of the coming decade is to create robust, accurate, efficient methods that learn to understand the main entities and concepts discussed in any text, and the main claims made."
(Fernando Pereira, Google Research Director, at the META-FORUM 2012)

Watson Doesn't Know It Won on 'Jeopardy!'
IBM invented an ingenious program—not a computer that can think.
By **JOHN SEARLE** (Wall Street Journal Feb. 23, 2011)

What will you learn in this course?

You will learn how to…

- Evaluate which problems need to be solved for a given NLP task;
- Choose an appropriate representation language for solving the task;
- Develop a model in that language;
- Perform inferences with the model;
- Evaluate the model's performance

Foundations

- Formal grammars ("Chomsky hierarchy")
  - Type 0 unrestricted grammars
  - Type 1 context-sensitive grammars
  - Type 2 context-free grammars
  - Type 3 regular grammars
- Automata
  - Turing machines
  - Linearly-bounded automata
  - Pushdown automata
  - Finite-state automata

  (and probabilistic counterparts)
- Shannon Information Theory
  - Perplexity
  - Entropy
  - Noisy channel
- Linear Algebra
  - Vector representation
  - Matrix calculus
  - Singular value decomposition
- Bayesian statistics
  - Inference
  - Prior distributions
  - EM-algorithm
  - Gibbs sampling
- Neural networks
  - Feed-Forward networks
  - Recurrent Recursive networks
  - Long Short-Term Memory

Preliminary Schedule

Date	Class	Lecturer	Linguistic phenomenon	techniques	Applications	Assignments
16/02/2018	1	Beuls	introduction / non-randomness	Frequency Distributions: words & collocations	Overview of applications	
23/02/2018	2	Beuls	morphology	regular expressions, finate state automata, string edits	tokenization, lemmatization, spelling correction	
02/03/2018	3	Beuls	n-grams	language models, entropy, perplexity	language guessing,	Assignment 1
09/03/2018	4	Van Eecke	Parts of speech	Hidden Markov Models	Labelling tasks	
16/03/2018	5	Beuls	syntax	Dynamic programming	Dependency parsing	
23/03/2018	6	Beuls	Fluid Construction Grammar	Feature structures, unification	Semantic parsing	
30/03/2018	7	Beuls	Pragmatics	inferencing	Language grounding	Assignment 2
06/04/2018	Easter	break				
13/04/2018	Easter	break				
20/04/2018	8	Beuls	Document semantics	Vector space models, topic models	Information retrieval	
27/04/2018	9	Beuls	Lexical semantics	word spaces, word embeddings	thesaurus extraction, lexical substitution	
04/05/2018	10	Beuls	compositional semantics	formal semantics, distrubutional semantics	summarization, logical inference	
11/05/2018	11	Beuls	opinion mining	sentiment analysis, paraphrase detection	content analytics	Assignment3
18/05/2018	12	Beuls	Machine translation	rule-based, statistical, neural MT	general vs domain-specific MT	
25/05/2018	13	Beuls	reserve class/ question time			

Levels of linguistic description and corresponding methods

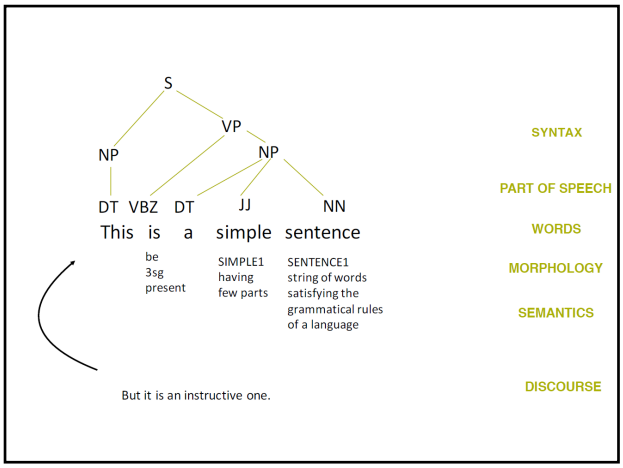
Language Processing

Linguistics

words
morphology
parts of speech
syntax
senses
semantics
discourse

Methods

finite state models
hidden Markov models
grammars and parsing
statistical parsing
sense disambiguation



Non-randomness of Natural Language

VRIJE UNIVERSITEIT BRUSSEL

Non-randomness 1

What are the relevant units in Natural Language?

VRIJE UNIVERSITEIT BRUSSEL

Primary data in Natural Language:

How can we find recurrent patterns

- **Spoken**: Sound waves, amplitudes varying across time
  - Acoustic signal processing
  - Transcription of indiviual phonemes

- **Written**: Strings of characters
  - Words: strings separated by spaces, but…
  - Latin: HORVMOMNIVMFORTISSIMISVNTBELGAEPROPTEREAQVODACVLTV ATQUEHUMANITATEPROVINCIAELONGISSIMEABSVNT
  - Chinese: 伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎
  - Punctuation: >,?!;:
  - Hyphens: sugar-free / sugarfree / sugar free
  - Clitics: Tom's / Kris' / Kris's
  - Compounds: esp. in languages like Dutch & German
    Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

Non-randomness 2

How repetitive is natural language?

VRIJE UNIVERSITEIT BRUSSEL

Word counts:

Not every word is equally frequent!

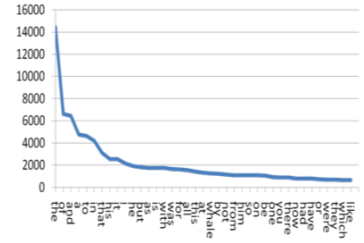
EUROPARL CORPUS
Transcriptions of debates in the European parliament

Most frequent words are ***function words***
They do not convey meaning but functioning as the 'glue' in a language

any word		nouns	
Frequency in text	Token	Frequency in text	Token
1,929,379	the	129,851	European
1,297,736	,	110,072	Mr
956,902	.	98,073	commission
901,174	of	71,111	president
841,661	to	67,518	parliament
684,869	and	64,620	union
582,592	in	58,506	report
452,491	that	57,490	council
424,895	is	54,079	states
424,552	a	49,965	member

Contrast with ***lexically full words***, like nouns

Only a few words occur very frequently.. …many words occur only once.



- Long right tail of distribution
- 33,447 words occur once, for instance
  - cornflakes
  - mathematicians

Zipf's law

$$f \times r = k$$

- f = frequency of a word
- r = rank of a word (if sorted by frequency)
- k = a constant

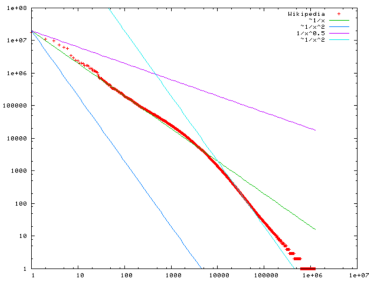
Frequency * frequency-rank: Linear relation on log-log scale

Zipf's law as a graph

Linear relationship

$f \times r = k$
$f = k / r$
$\log f = \log k - \log r$

=> "power law"



Non-randomness 3

How do words co-occur with each other?

VRIJE UNIVERSITEIT BRUSSEL

Bigrams: pairs of words
N-grams: sequence of n-words

- Most frequent bigrams are not very interesting: simply combinations of high frequent words (function words)

ALTERNATIVE

- Look at bigrams that are combinations of lexically full words (nouns, adjectives, verbs) => multi-word expressions
- Look at bigrams that are combinations of words that co-occur more often than expected by chance => collocations

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Combinations of lexically full words

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Multi-word-expressions:
Combination of words that refer to

- a single entity in reality (Named Entities)
- a single concept that is thought of as a whole, "unit status" ≈ compounds

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Collocation: Test if 2 words co-occur more often than expected by chance

Null-hypothesis:
- P(w1w2) = P(w1)P(w2)

Test/quantify deviance
- T-test, Chi-square, log LLH
- Pointwise mutual information

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Words can co-occur above chance even if they are not right next to each other
- He knocked yesterday afternoon at my door
- Window-based co-occurrence frequencies

Window 10 left, 10 right

	Cooccurrence count	Candidate count	T-score
door	803	22,997	28.305
knock	81	1,685	8.993
doors	98	4,407	9.882
unconscious	47	1,336	6.848
off	545	67,192	23.232
knocked	47	2,194	6.843
down	664	90,098	25.631
sideways	30	924	5.471
knocking	27	917	5.189
opened	75	10,825	8.611
stuffing	23	297	4.793
wall	68	9,971	8.199

MWE and collocations: More than sum of parts

- **Idiomatic:** convential way of saying things
  - Let's have a cup of strong tea
- **Non compositionality**: meaning of the expression cannot be fully predicted from its parts
  - Collocations put constraints on meaning: strong tea taste
  - Difficult to subsitute with near-synonyms powerful tea
  - MWEs refer to concepts, units: e.g. real estate
  - Idioms are extreme examples of non-compositionality
    - He kicked the bucket
    - She was caught between two stools
    - It cost me an arm and a leg
  - Not freely modifiable He kicked the water bucket??
  - Not literally translatable Il a lancé le seau??

NLP with Python Natural Language ToolKit (NLTK)

VRIJE UNIVERSITEIT BRUSSEL

What is NLTK?

Language processing task	NLTK modules	Functionality
Accessing corpora	corpus	standardized interfaces to corpora and lexicons
String processing	tokenize, stem	tokenizers, sentence tokenizers, stemmers
Collocation discovery	collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	tag	n-gram, backoff, Brill, HMM, TnT
Machine learning	classify, cluster, tbl	decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	chunk	regular expression, n-gram, named-entity
Parsing	parse, ccg	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	sem, inference	lambda calculus, first-order logic, model checking
Evaluation metrics	metrics	precision, recall, agreement coefficients
Probability and estimation	probability	frequency distributions, smoothed probability distributions
Applications	app, chat	graphical concordancer, parsers, WordNet browser, chatbots
Linguistic fieldwork	toolbox	manipulate data in SIL Toolbox format

Checklist

- Texts are represented in Python using lists: ['Monty', 'Python']. We can use indexing, slicing, and the len() function on lists.
- A word "token" is a particular appearance of a given word in a text; a word "type" is the unique form of the word as a particular sequence of letters. We count word tokens using len(text) and word types using len(set(text)).
- We obtain the vocabulary of a text t using sorted(set(t)).
- To derive the vocabulary, collapsing case distinctions and ignoring punctuation, we can write set(w.lower() for w in text ifw.isalpha()).
- A frequency distribution is a collection of items along with their frequency counts (e.g., the words of a text and their frequency of appearance).

Homework

- Install NLTK
- Go through commands and exercises in Chapter 1
- http://www.nltk.org/book/ch01.html

Next week

Date	Class	Lecturer	Linguistic phenomenon	techniques	Applications	Assignments
16/02/2018	1	Beuls	introduction / non-randomness	Frequency Distributions: words & collocations	Overview of applications	
23/02/2018	2	Beuls	morphology	regular expressions, finate state automata, string edits	tokenization, lemmatization, spelling correction	
02/03/2018	3	Beuls	n-grams	language models, entropy, perplexity	language guessing,	Assignment 1
09/03/2018	4	Van Eecke	Parts of speech	Hidden Markov Models	Labelling tasks	
16/03/2018	5	Beuls	syntax	Dynamic programming	Dependency parsing	
23/03/2018	6	Beuls	Fluid Construction Grammar	Feature structures, unification	Semantic parsing	
30/03/2018	7	Beuls	Pragmatics	Inferencing	Language grounding	Assignment 2
06/04/2018	Easter	break				
13/04/2018	Easter	break				
20/04/2018	8	Beuls	Document semantics	Vector space models, topic models	Information retrieval	
27/04/2018	9	Beuls	Lexical semantics	word spaces, word embeddings	thesaurus extraction, lexical substitution	
04/05/2018	10	Beuls	compositional semantics	formal semantics, distrubutional semantics	summarization, logical inference	
11/05/2018	11	Beuls	opinion mining	sentiment analysis, paraphrase detection	content analytics	Assignment3
18/05/2018	12	Beuls	Machine translation	rule-based, statistical, neural MT	general vs domain-specific MT	
25/05/2018	13	Beuls	reserve class/ question time			