Text Mining

2004-2005

Master TKI

Antal van den Bosch en Walter Daelemans

http://ilk.uvt.nl/~antalb/textmining/

Dinsdag, 10.45 - 12.30, SZ33

Timeline (1)

- [1 februari 2005]
  - Introductie (WD)
- [15 februari 2005]
  - Syntactic pipeline 1: Tokenization, POS tagging (AB)
- [22 februari 2005]
  - Concept chunking (Sander Canisius)
- [1 maart 2005]
  - Syntactic pipeline 2: chunking, relation finding (WD)

Timeline (2)

- [8 maart 2005]
  - Named-entity recognition (Toine Borgers)
- [15 maart 2005]
  - Information extraction (WD)
- [5 april 2005]
  - Tools (AB)
- [12 april 2005]
  - Industrial information extraction

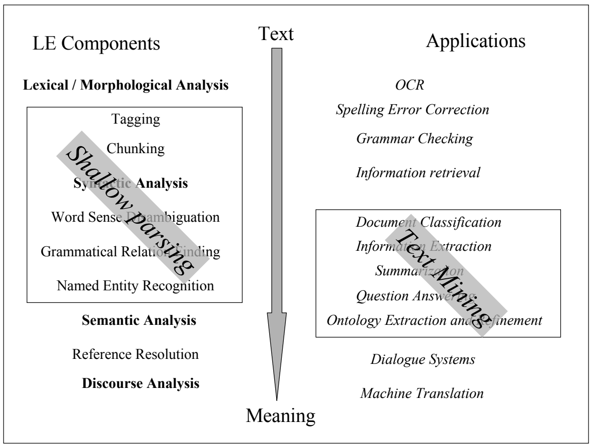
Timeline (3)

- [19 april 2005]
  - Factoids (AB)
- [26 april 2005]
  - Ontology learning (Marie-Laure Reinberger)
- [3 mei 2005]
  - Information extraction from spoken user input (Piroska Lendvai)
- [10 mei 2005]
  - Presentaties

Overview

- What is Text Mining?
- Which Language Technology tools are useful?
- Evaluation:
  - 2 exercises: software assignments (programming / tuning / testing of modules)
  - Final paper and presentation

What is Text Mining?



LE Components

Text

Applications

Lexical / Morphological Analysis

Tagging

Chunking

Syntactic Analysis

Word Sense Disambiguation

Grammatical Relation Finding

Named Entity Recognition

Semantic Analysis

Reference Resolution

Discourse Analysis

Shallow Parsing

OCR

Spelling Error Correction

Grammar Checking

Information retrieval

Document Classification

Information Extraction

Summarization

Question Answering

Ontology Extraction and Refinement

Text Mining

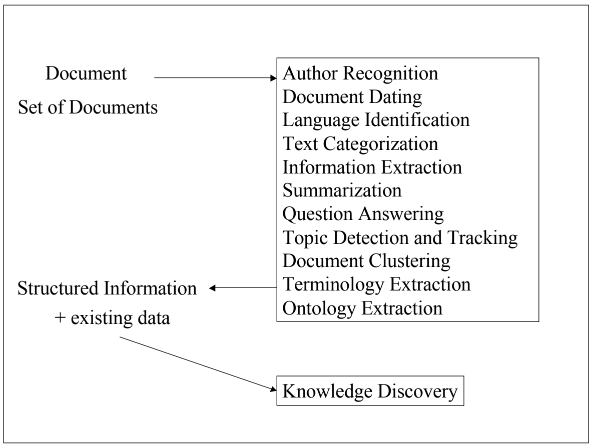
Dialogue Systems

Machine Translation

Meaning

Text Mining

- Automatic extraction of reusable information (knowledge) from text, based on linguistic features of the text
- Goals:
  - Data mining (KDD) from unstructured and semi-structured data
  - (Corporate) Knowledge Management
  - “Intelligence”
- Examples:
  - Email routing and filtering
  - Finding protein interactions in biomedical text
  - Matching resumes and vacancies



Document

Set of Documents

Author Recognition
Document Dating
Language Identification
Text Categorization
Information Extraction
Summarization
Question Answering
Topic Detection and Tracking
Document Clustering
Terminology Extraction
Ontology Extraction

Structured Information

+ existing data

Knowledge Discovery

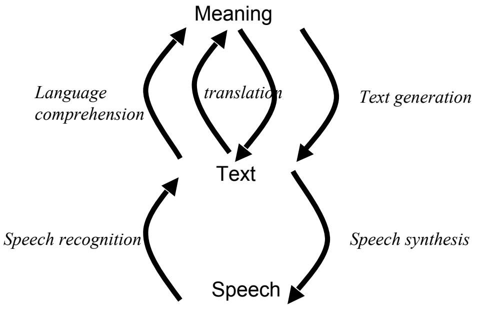
Role of Language Technology: Compute Text Representation Units

- Character n-grams
- Words, phrases, heads of phrases
- POS tags
- Parse tree (fragment)s
- Grammatical Relations
- Frames and scripts
- “meaning” (?)

Text is a special kind of data

- Direct entry, OCR (.99 accuracy), Speech Recognition output (.50-.90 accuracy), …
- What we have:
  - Characters, character n-grams, words, word n-grams, lay-out, counts, lengths, …
- What we want:
  - Meaning (answering questions, relating with previous knowledge)
- Bridging the gap:
  - Tagging, lemmatization, phrase chunking, grammatical relations, … I.e.: *Language Technology*

What is Language Technology?



- Language Technology (Natural Language Processing, Computational Linguistics) is based on the complex transformation of linguistic representations
- Examples
  - from text to speech
  - from words to morphemes
  - from words to syntactic structures
  - from syntactic structures to conceptual dependency networks

- In this transformation, two processes play a role
  - segmentation of representations
  - disambiguation of possible transformations of representation units
- Similar representations at input level correspond to similar representations at the output level
- Complexity because of context-sensitivity (regularities, subregularities, exceptions)

gebruiksvriendelijkheid
ge+bruik+s+vriend+elijk+heid

The old man the boats
det N-plur V-plur det N-plur Punc

The old man the boats
(S (NP (DET the) (N old)) (VP (V man) (NP (DET the) (N boats))))

The old man the boats
De ouden bemannen de schepen

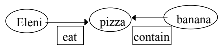
(S (NP (DET the) (N old)) (VP (V man) (NP (DET the) (N boats))))
(man-action (agent (def plur old-person)) (object (def plur boat)))

How to reach Language Understanding ?

- A fundamental solution for the problem of language understanding presupposes
  - Representation and use of knowledge / meaning
  - Acquisition of human-level knowledge

What is meaning ?

Eleni eats a pizza with banana



Semantic networks, Frames

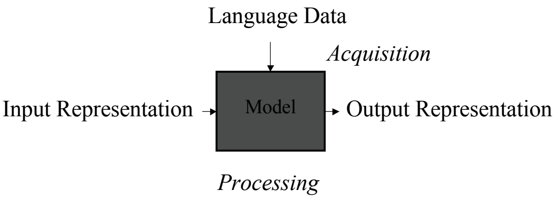
∃(x): pizza(x) ^ eat(Eleni,x) ^ contain(x, banaan) *First-order predicate calculus*

Pizza = {p1, p2, p3, …}
Eat ={<Nicolas,p1>,<Nicolas,p3>,<Eleni,p2>,…}
Contain ={<p1,ansjovis>,<p1,tomaat>,<p2,banaan>,…} *Set theory*
x=p2

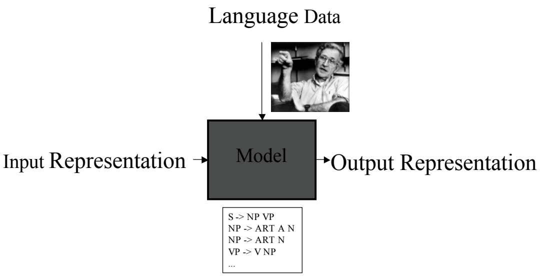
“Symbol grounding” problem
Representation and processing of time, causality, modality, defaults, common sense, …

“Meaning is in the mind of the beholder”

Language Technology



Deductive Route



S -> NP VP
NP -> ART A N
NP -> ART N
VP -> V NP
...

Deductive Route

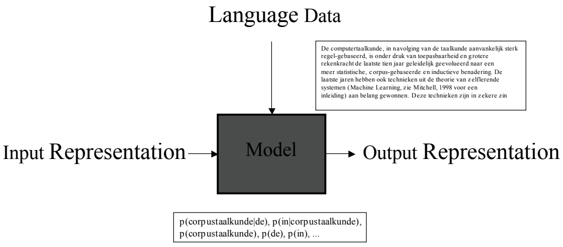
- Acquisition

  Construct a (rule-based) model about the domain of the transformation.

- Processing

  Use rule-based reasoning, deduction, on these models to solve new problems in the domain.

Inductive Route



De computertaalkunde, in navolging van de taalkunde aanvankelijk sterk regel-gebaseerd, is onder druk van toepasbaarheid en grotere rekenkracht de laatste tien jaar geleidelijk geevolueerd naar een meer statistische, corpus-gebaseerde en inductieve benadering. De laatste jaren hebben ook technieken uit de theorie van zelflerende systemen (Machine Learning, zie Mitchell, 1998 voor een inleiding) aan belang gewonnen. Deze technieken zijn in zekere zin

p(corpustaalkunde|de), p(in|corpustaalkunde), p(corpustaalkunde), p(de), p(in), ...

Inductive Route

- Acquisition

  Induce a stochastic model from a corpus of “examples” of the transformation.

- Processing

  Use statistical inference (generalization) from the stochastic model to solve new problems in the domain.

Advantages

Deductive Route

- Linguistic knowledge and intuition can be used
- Precision

Inductive Route

- Fast development of model
- Good coverage
- Good robustness (preference statistics)
- Knowledge-poor
- Scalable / Applicable

Problems

Deductive Route

- Representation of sub/irregularity
- Cost and time of model development
- (Not scalable / applicable)

Inductive Route

- Sparse data
- Estimation of relevance statistical events

Applications of Text Mining

Question Answering

- Give answer to question
  (document retrieval: find documents relevant to query)
- Who invented the telephone?
  – Alexander Graham Bell
- When was the telephone invented?
  – 1876

QA System: Shapaqa

- Parse question
  When was the telephone invented?
  – Which slots are given?
    - Verb invented
    - Object telephone
  – Which slots are asked?
    - Temporal phrase linked to verb
- Document retrieval on internet with given slot keywords
- Parsing of sentences with all given slots
- Count most frequent entry found in asked slot (temporal phrase)

Shapaqa: example

- *When was the telephone invented?*
- Google: invented AND “the telephone”
  – produces 835 pages
  – 53 parsed sentences with both slots and with a temporal phrase

is through his interest in Deafness and fascination with acoustics that the telephone was invented in 1876 , with the intent of helping Deaf and hard of hearing

The telephone was invented by Alexander Graham Bell in 1876

When Alexander Graham Bell invented the telephone in 1876 , he hoped that these same electrical signals could

Shapaqa: example (2)

- So when was the phone invented?
- Internet answer is noisy, but robust
  – 17: 1876
  – 3: 1874
  – 2:ago
  – 2:later
  – 1:Bell
  – …
- System was developed quickly
- Precision 76% (Google 31%)
- International competition (TREC): MRR 0.45

Who shot Kennedy ?

4 x OSWALD

- * www.anusha.com/jfk.htm
  situation in which Oswald shot Kennedy on November 22 , 1963 .
- * www.mcb.ucdavis.edu/people/hemang/spooky.html
  Lee Harvey Oswald shot Kennedy from a warehouse and ran .
- * www.gallica.co.uk/monarch.htm
  November 1963 U.S. President Kennedy was shot by Lee Harvey Oswald .
- * astrospeak.indiatimes.com/mystic_corner.htm
  Lee Harvey Oswald shot Kennedy from a warehouse and fled .

2 x BISHOP

- * www.powells.com/biblio/0-200/000637901X.html
  The day Kennedy was shot by Jim Bishop .
- * www.powells.com/biblio/49200-49400/0517431009.html
  The day Kennedy was shot by by Jim Bishop .

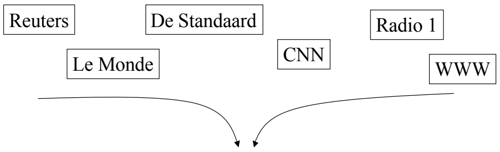
1 x BULLET

- * www.lustlysex.com/index_m.htm
  President John F. Kennedy was shot by a Republican bullet .

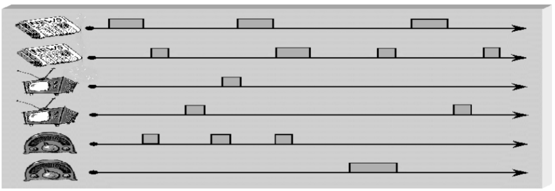
1 x MAN

- * www.ncas.org/condon/text/appndx-p.htm
  KENNEDY ASSASSINATION Kennedy was shot by a man who was not .

Topic Detection and Tracking



First Story Detection
Topic Detection
Topic Tracking
Topic Segmentation
Link Detection



= document classification