Natural Language Processing in a nutshell

Veronique Hoste

LT3 Language and Translation Technology Team
Ghent University

Part I

Introduction

Computational linguistics or Natural Language Processing

- science domain that studies the human language system from a computational perspective
- subdiscipline of AI (next to robotics, machine learning, etc.)
- goal: build models of human intelligence??

How long till human-level artificial intelligence?

Central hypothesis in AI

If we succeed in adequately formulating the right knowledge (data structures) and cognitive processes (algorithms), we can make computers intelligent (and let them understand language). This does not have to be a model of how humans do it!

How long till human-level artificial intelligence?

Computers can perform tasks we perceive as intelligent, e.g. medical diagnosis, find oil or gas, play chess, ...

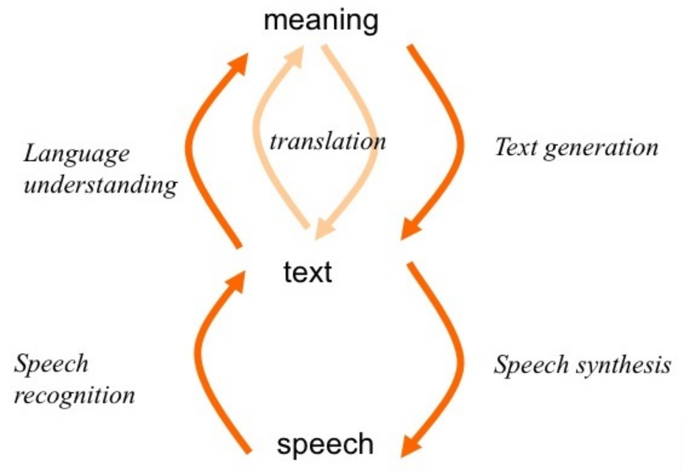
How long till human-level artificial intelligence?

▸ ./jeopardy.jpg

Human Language Technology or Language Engineering

- applied computational linguistics
- develop software that speaks, listens and understands (a bit)
- used in many applications (search engines, mobile telephones, GPS-systems, etc.)

Language and speech technology



Application areas

- Intelligence / military purposes: Intelligent text processing, machine translation, automatic summarization
- Media: e.g. ASR $\Rightarrow$ MT $\Rightarrow$ automatic subtitling
- Medical domain: e.g. automatic information extraction from patient files
- Marketing / CRM: e.g. “sentiment detection” in blogs
- (...)

Sentiment detection

1 in 3 auto buyers say that social media help them make their decision. 51% use it to help narrow their choice, 23% to confirm a choice, 15% to select a top choice. User feedback causes 24% of auto buyers to change their mind about the type of vehicle they select to purchase.
(http://www.attentio.com/industries.htm)

Automatic sentiment detection

- can help companies to gain insight into their customers
- can help companies to gain insight into their products
- can help customers to choose the right products (recommender systems)

The downside of social media: need for automatic filtering

Warning to parents as 'worst ever' internet paedophile jailed

Parents have been warned that their children are not safe from paedophiles even in their own bedrooms as Britain's worst internet child sex abuser was jailed.

The Telegraph

Paedophile who groomed victims on Facebook jailed

Link to this video

By Richard Savill

4:32PM BST 24 Sep 2010

Michael Williams, 28, a pos

Share:

Recommend 206

Tweet 28

Share 0

+1 0

Crime

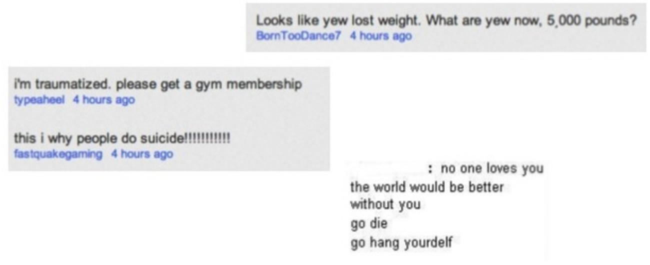
News » UK News »

Technology News »

IN CRIME

Moat 999 call: 'I'm hunting officers'

The downside of social media: need for automatic filtering



The downside of social media: need for automatic filtering

"(...) it is crucial to have an inventory of relevant 'User Generated Content'. 25% of the relevant links on suicide are blogs. (...) In addtion to this, we want to have a clear view on the number of 'pro-suicide' sites, informative sites and prevention sites. Once we know where on the use of suicidal terminology on the Web 2.0, we will be able to put buttons and/or banners on priority blogs, profiles and sites. (...) In 2007, about 87% of the people calling the suicide line were younger than 30 years old." (Grieke Forceville, head of the Belgian Suicide Prevention Centre)

Example

my name is mike iv been depressed for 3 years im 15 nd i fell in love with a beautiful girl named sierra every day i told her how much i loved her nd what id do for her but one day in 7th grade when i told her i love her she said but i dont love u nd to leave her alone ever sense iv been horribly depressd iv tried to kil myself 4 times but never succeeded i hate my life if anyone can help please contact me at (...)

Application areas (ctd.)

Humanities, Social Sciences, etc.

- Political sciences: study political discourse over time
- Economics: event detection + sentiment analysis for market prediction; how do companies present themselves in annual reports or in sustainability reports?
- Communication sciences: e.g. is it possible to build models of emotion for crisis detection / communication?
- Literary studies: who wrote this manuscript?
- (...)

Example

Vodafone gave itself a PR headache after an employee sent out an obscene, homophobic tweet from its official account

Vodafone's followers were shocked to see an obscene tweet in their stream from the brand's official account. The company received hundreds of complaints immediately and the media picked it up. Vodafone quickly had a crisis to manage.

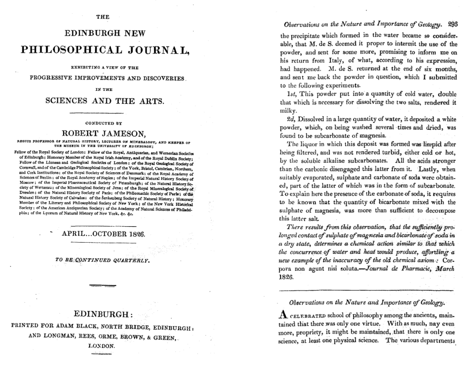
Twitter

The initial assumption was that the brand's account got hacked, but it turns out that the tweet was sent out by one of its own staff. Whatever Vodafone's checks are regarded account access, they weren't enough, but at least it was transparent about what happened.

The employee was later suspended.

Example

- Text : first positive reaction to the controversial evolution theory of Lamarck: Observations on the Nature and Importance of Geology



THE

EDINBURGH NEW

PHILOSOPHICAL JOURNAL,

EXHIBITING A VIEW OF THE

PROGRESSIVE IMPROVEMENTS AND DISCOVERIES

IN THE

SCIENCES AND THE ARTS.

CONDUCTED BY

ROBERT JAMESON,

REGIUS PROFESSOR OF NATURAL HISTORY, LECTURER ON MINERALOGY, AND KEEPER OF THE MUSEUM IN THE UNIVERSITY OF EDINBURGH;

Fellow of the Royal Society of London: Fellow of the Royal, Antiquarian, and Wernerian Societies of Edinburgh; Honorary Member of the Royal Irish Academy, and of the Royal Dublin Society; Fellow of the Linnean and Geological Societies of London; of the Royal Geological Society of Cornwall, and of the Cambridge Philosophical Society; of the York, Bristol, Cambrian, Northern, and Cork Institutions; of the Royal Society of Sciences of Denmark; of the Royal Academy of Sciences of Berlin; of the Royal Academy of Naples; of the Imperial Natural History Society of Moscow; of the Imperial Pharmaceutical Society of Petersburgh; of the Natural History Society of Wetterau; of the Mineralogical Society of Jena; of the Royal Mineralogical Society of Dresden; of the Natural History Society of Paris; of the Philomathic Society of Paris; of the Natural History Society of Calvados; of the Senkenberg Society of Natural History; Honorary Member of the Literary and Philosophical Society of New York; of the New York Historical Society; of the American Antiquarian Society; of the Academy of Natural Sciences of Philadelphia; of the Lyceum of Natural History of New York, &c. &c.

APRIL...OCTOBER 1826.

TO BE CONTINUED QUARTERLY.

EDINBURGH:

PRINTED FOR ADAM BLACK, NORTH BRIDGE, EDINBURGH; AND LONGMAN, REES, ORME, BROWN, & GREEN, LONDON.

Observations on the Nature and Importance of Geology. 293

the precipitate which formed in the water became so considerable, that M. de S. deemed it proper to intermit the use of the powder, and sent for some more, promising to inform me on his return from Italy, of what, according to his expression, had happened. M. de S. returned at the end of six months, and sent me back the powder in question, which I submitted to the following experiments.

1*st*, This powder put into a quantity of cold water, double that which is necessary for dissolving the two salts, rendered it milky.

2*d*, Dissolved in a large quantity of water, it deposited a white powder, which, on being washed several times and dried, was found to be subcarbonate of magnesia.

The liquor in which this deposit was formed was limpid after being filtered, and was not rendered turbid, either cold or hot, by the soluble alkaline subcarbonates. All the acids stronger than the carbonic disengaged this latter from it. Lastly, when suitably evaporated, sulphate and carbonate of soda were obtained, part of the latter of which was in the form of subcarbonate. To explain here the presence of the carbonate of soda, it requires to be known that the quantity of bicarbonate mixed with the sulphate of magnesia, was more than sufficient to decompose this latter salt.

There results from this observation, that the sufficiently prolonged contact of sulphate of magnesia and bicarbonate of soda in a dry state, determines a chemical action similar to that which the concurrence of water and heat would produce, affording a new example of the inaccuracy of the old chemical axiom: Corpora non agunt nisi soluta.—*Journal de Pharmacie, March* 1826.

Observations on the Nature and Importance of Geology.

A CELEBRATED school of philosophy among the ancients, maintained that there was only one virtue. With as much, nay even more, propriety, it might be maintained, that there is only one science, at least one physical science. The various departments

- 2 candidates for the authorship: Robert Jameson or Robert Grant (both have written a lot)

Part II

Ambiguity and the waterfall model

Ambiguity

The fundamental problem of language technology

Lexical-morphological

= ambiguity at the word and sub-word level

Morphological

(...) een eengemaakte politiezone (...)
(...) a-made police force area (...)
Fremdzugehen, betrachtet die Familie als eine Schande.
External train marriages, the family considers as a disgrace.
Der Photograf hat das Model abgelichtet.
The photo count photocopied the model.

Morpho-syntactic

= what is the contextually appropriate morphosyntactic category of each word?

Morpho-syntactic

Syntactic

= what is the contextually appropriate syntactic category of each word?

Flying planes can be dangerous.

Ik eat pizza with olives / Ik eat pizza with my friend.

Semantic

= what is the contextually appropriate sense of a given word?

Some of the features of the Yamaha Motor Canada site require that your browser have cookies enabled. For optimum user experience, please configure your browser to accept cookies under your Security Preferences. You may continue without cookies enabled.

Certaines caractéristiques du site de Yamaha Moteur du Canada Ltée exigent que les biscuits magiques de votre navigateur soient activés. Pour assurer une navigation optimale, veuillez configurer votre navigateur de sorte à ce qu'il accepte les biscuits magiques en passant par les préférences de sécurité. Vous pouvez toutefois continuer sans activer les biscuits magiques.

YAMAHA

Semantic

Zalm werd geboren als zoon van een kolenboer.
Salmon was born as son of a coal farmer.

Discourse

= referential ambiguity

The monkey ate the banana because it was hungry.
Der Affe aß die Banane weil er Hunger hatte.

The monkey ate the banana because it was ripe.
Der Affe aß die Banane weil sie reif war.

The monkey ate the banana because it was lunch time.
Der Affe aß die Banane weil es Zeit zum Essen war.

Discourse

The soldiers shot at the women and some of them fell.
Les soldats ont tiré sur les femmes et quelques unes sont tombées.

The soldiers shot at the women and some of them missed.
Les soldats ont tiré sur les femmes et quelques uns ont raté.

Put the paper in the printer. Then switch it on.

World knowledge

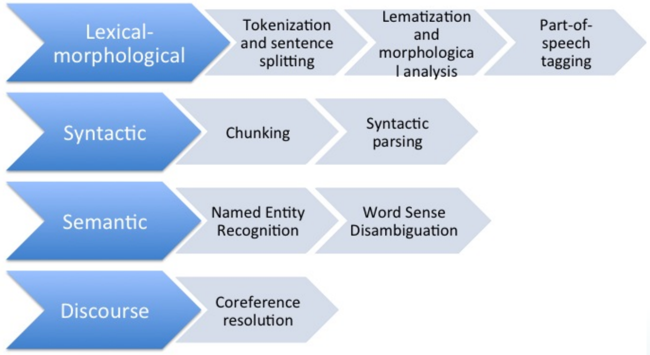
= the real problem!

Tom was unemployed. He took the newspaper.
The fly was getting on Tom's nerves. He took the newspaper.

Recap: Ambiguity in language

- AI complete: learning, common sense, world knowledge
- NLP solution: make models for transformations between different representations

How to solve ambiguity: the waterfall model



How to solve ambiguity: the waterfall model

Every module in the waterfall model transforms a linguistic input representation into a linguistic output representation and does so by using a “model” of this transformation

2 possible approaches

- deductive approach: the scientist builds the information sources and rules which are necessary to implement the desired transformation
- inductive approach: the scientist collects examples of the transformation and uses statistical and learning approaches which enable the computer to build the model by itself

Deductive and inductive computational linguistics

Acquisition

Construct a rule-based model about the domain
vs.
Induce a stochastic model from a corpus of examples

Processing

Use rule-based reasoning, deduction, on these models to solve new problems in the domain
vs.
Use statistical inference (generalization) from the stochastic model to solve new problems in the domain

Deductive and inductive computational linguistics: the clash

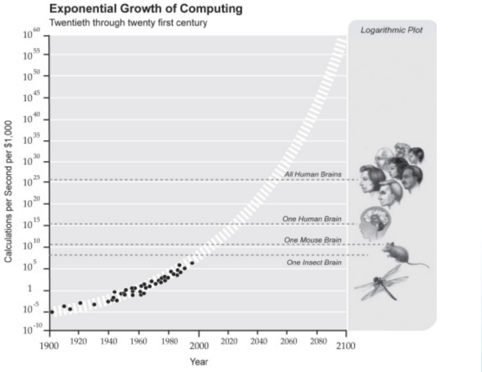
Deductive and inductive computational linguistics: the clash

Every time I fire a linguist, the performance of our speech recognition system goes up. (Fred Jelinek)

The validity of a statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950 () The crude force of computers is not science. (review COLING 1988)

Deductive and inductive computational linguistics: and the winner is ...

- success of statistics in related research domains such as speech technology and IR
- growing processing and storage capacity of computers



Deductive and inductive computational linguistics: and the winner is ... (ctd.)

- success of statistics in related research domains such as speech technology and information retrieval
- growing processing and storage capacity of computers
- more corpora and larger corpora available
- sociological: inductive approaches reach higher accuracies; in a competitive field where research depends on competitive financing, the most powerful methodologies win.

An example: the shift in machine translation

Крабы способны чувствовать боль

28.01.2013

Ученые из Королевского университета в Белфасте (Великобритания) изучили реакцию прибрежных крабов на слабые удары электрическим током и выяснили, что они тоже способны чувствовать боль. Получается, что кидать ракообразных в кипяток, пока они еще живы, не является этичным — такой метод готовки причиняет этим животным непереносимые страдания.

MT Output

Systran

Scientists from royal university to Belfast (Great Britain) studied the reaction of coastal it was crab to the weak impacts by electric current they explained that they are also capable of feeling pain. It it turns out that to throw crustaceans into the boiling water, until they are still living, is not ethical – this method of preparation causes the unbearable sufferings to these animals.

Google Translate

Scientists from Queen's University Belfast (UK) studied the response of coastal crab on weak electric shocks and found that they too are capable of feeling pain. It turns out that throwing crustaceans into boiling water, while they are still alive, is not ethical - this method of cooking is causing untold suffering to these animals.

An example: German plural

- input representation: German singular noun
- output representation: plural noun

Example

Frau - Frauen
Nacht - Nächte
Tochter - Töchter
Kind - Kinder
Mann - Männer
etc.

An example: German plural

Deductive approach

- take a German grammar and define rules which describe the problem to be solved
- implement those rules e.g. using regular expressions in Python (input text; apply rules; write to output text)
  e.g. Rule 1: if the noun is feminine, use the suffix . . . , unless in . . . , . . . , . . .
- Problem: many subregularities and exceptions in language (e.g. loan words, unproductive regularities, neologisms, etc.) leading to complex rule sets
- “Bounded rationality” of the linguist

An example: German plural

Deductive approach (ctd.)

It is rather easy to build a working system which describes regularities in language. It is however very hard to improve the accuracy of the system to an acceptable level due to the complexity of sub regularities and exceptions and rules contradicting each other.

An example: German plural

Inductive approach

- the computational linguist collects examples of singular nouns and their plural
- these examples are given to the learning system in a consistent way using a feature vector
  - e.g. the singular noun is represented as a set of features describing the phonological structure of the last syllabe of the word (onset, nucleus en coda) and the gender of the noun
  - e.g. the class: some kind of formula to build the plural: ”Umlaut + er”

An example: German plural

Inductive approach (ctd.)

M,a,nn,masculine,Umlaut+er
t,e,r,feminine,Umlaut
N,a,cht,feminine,Umlaut+e
F,r,au,feminine,+en

To convert your data (nouns in singular and plural) into a feature vector, you can use Python.

It is crucial to define relevant features for solving your problem: garbage in, garbage out

An example: German plural

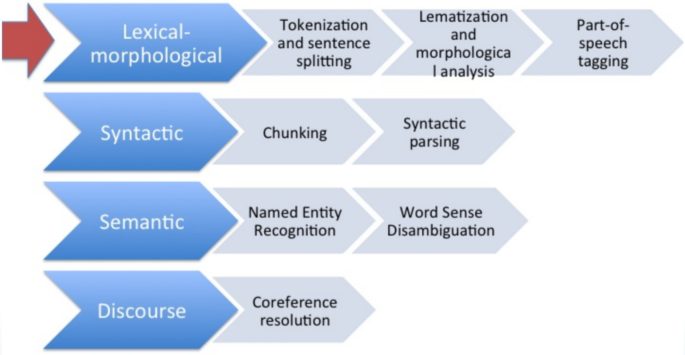
Inductive approach (ctd.)

- once the data set is ready, it is presented to the learning system which tries to induce a model from the data
- Key to the success of the learner: it should also generalise to unseen words (current acc. $>$ 95%)

Learned rule for the suffix -er

IF gender = neuter, AND
coda=lt or nt AND nucleus=i or e OR
coda=cht, onset=l OR
coda=t, nucleus=i, onset=l OR
nucleus=au, onset=kl OR
onset=br, nucleus=e OR
onset=gl, nucleus=i OR
THEN +er

How to solve ambiguity: the waterfall model



Lexical-morphological level

Tokenization and sentence splitting

- = automatic word recognition and sentence splitting
- “token”: Series of letters and numbers separated by interpunction, white space or mark-up.
- sentence: Series of words starting with a capitalized word and ending with a question mark, full stop, or exclamation mark
- All approaches are rule-based/deductive
- Try it out:
  To use in Python: NLTK package, pattern
  Or: Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml), TreeTagger

Lexical-morphological level

Example

- http://text-processing.com/demo/tokenize/
- http://www.ncbi.nlm.nih.gov/books/NBK195574/

Lexical-morphological level

Tokenization and sentence splitting

Challenges:

1. Abbreviations with punctuation
   Ave., Corp., Gov., Dept., Vol., 29 Oct., ...
2. Punctuation as part of a word
   C++
   C#
   B-52
   M*A*S*H

Lexical-morphological level

Stemming/lemmatization

- = removing and replacing word suffixes to arrive at a common root form of the word
- Lemmas differ from stems in that a lemma is a canonical form of the word, while a stem may not be a real word.
- Both deductive and inductive approaches.
- Try it out:

  http://text-processing.com/demo/stem/

  To use in Python: NLTK package, pattern
  Or: Stanford CoreNLP
  (http://nlp.stanford.edu/software/corenlp.shtml)

Lexical-morphological level

Part of speech tagging

- **Goal:** Assign the contextually appropriate morpho-syntactic category to a given word
- How?
  - **deductive**: define a set of rules to determine the PoS-categorie for each word in a given sentence
  - **inductive**: “train” a statistical system on the basis of a corpus labeled with PoS information
- Try it out:

  http://text-processing.com/demo/tag/

  To use in Python: NLTK package, pattern
  Or: Stanford CoreNLP
  (http://nlp.stanford.edu/software/corenlp.shtml)

Stochastic POS tagging: a simple bigram tagger

- Starting point: the tagging problem can be solved by looking at the words in the local context.

Example

He is expected to race tomorrow .
"race" as noun or as verb??

- How? Tag sequence probability * word probability

Example

He is expected to race tomorrow .
$P(VB|TO)P(race|VB)$
$P(NN|TO)P(race|NN)$

Jurafsky and Martin, "SPEECH and LANGUAGE PROCESSING", 2009.

Example training corpus

NNP/ Houston ,/ , NNP/ Monday ,/ , NNP/ July CD/ 21 :/ -- NN/ Men VBP/ have VBD/ landed CC/ and VBD/ walked IN/ on DT/ the NN/ moon ./ . CD/ Two NNPS/ Americans ,/ , NNS/ astronauts IN/ of NNP/ Apollo CD/ 11 ,/ , VBD/ steered PRP$/ their JJ/ fragile JJ/ four-legged NN/ lunar VB/ module RB/ safely CC/ and RB/ smoothly TO/ to DT/ the JJ/ historic NN/ landing NN/ yesterday IN/ at NN/ 4:17:40 NNP/ P.M. ,/ , NNP/ Eastern NN/ daylight NN/ time ./ . NNP/ Neil NNP/ A. NNP/ Armstrong ,/ , DT/ the JJ/ 38-year-old JJ/ civilian NN/ commander ,/ , VBD/ radioed TO/ to NN/ earth CC/ and DT/ the NN/ mission NN/ control NN/ room RB/ here :/ : ``/ " NNP/ Houston ,/ , NNP/ Tranquility NNP/ Base RB/ here :/ ; DT/ the NNP/ Eagle VBZ/ has VBN/ landed ./ . "/ "

Stochastic POS tagging

- 1. Tag sequence probability $P(t_i|t_{i-1})$
  - How probable is it that a POS is a noun or verb given the previous POS tag?
  - Starting point: a verb is more probable eg. "to walk", "to eat", "to have" versus "go to school"
  - Calculations on the basis of the Brown corpus:
    - P(NN|TO) = 0.021
    - P(VB|TO) = .34

  ⇒"to" is more often followed by a verb than a noun

Jurafsky and Martin, "SPEECH and LANGUAGE PROCESSING", 2009.

Stochastic POS tagging

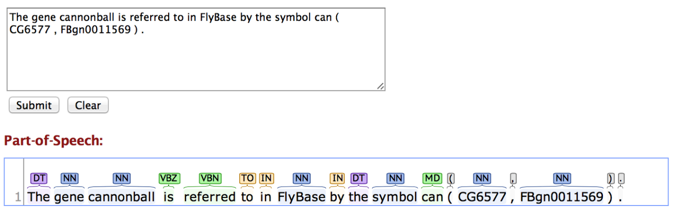
- 2. Word probability
  - If we expect a noun/verb: how probable is it that the verb/noun will be "race"?
  - Calculations on the basis of the Brown corpus:
    - $P(race|NN) = .00041$
    - $P(race|VB) = .00003$

  $\Rightarrow$ "race" occurs more often as noun than as verb
- 3. Combination
  $P(VB|TO)P(race|VB) = .0000102$
  $P(NN|TO)P(race|NN) = .00000861$

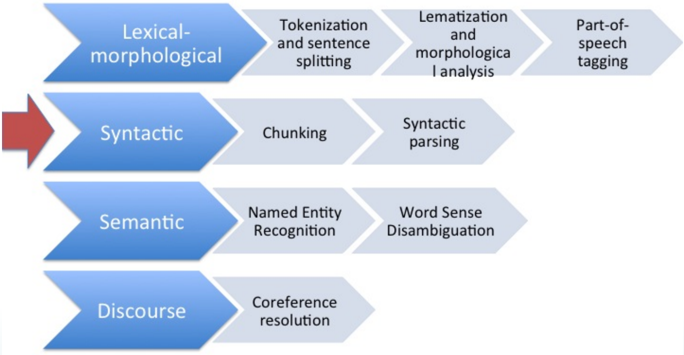
Jurafsky and Martin, "SPEECH and LANGUAGE PROCESSING", 2009.

Lexical-morphological level

Part of speech tagging



How to solve ambiguity: the waterfall model



How to solve syntactic ambiguity?

Parsing

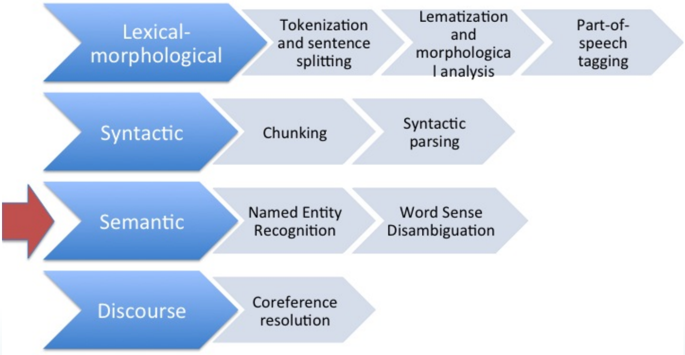
- **Goal**: build a syntactic structure/tree of a sentence (divide in constituents)
- 2 possible search strategies:
  - Top down: parser builds a tree starting from the sentence node and searches for possible edges
  - Bottom up: parser starts from the words in the sentence

  Mostly shallow syntax through chunking and dependency parsing .
- Demo:

  ▸ http://nlp.stanford.edu/software/corenlp.shtml

  ▸ http://cogcomp.cs.illinois.edu/demo/shallowparse/?id=7

How to solve ambiguity: the waterfall model



How to solve semantic ambiguity?

Automatic named entity recognition (NER)

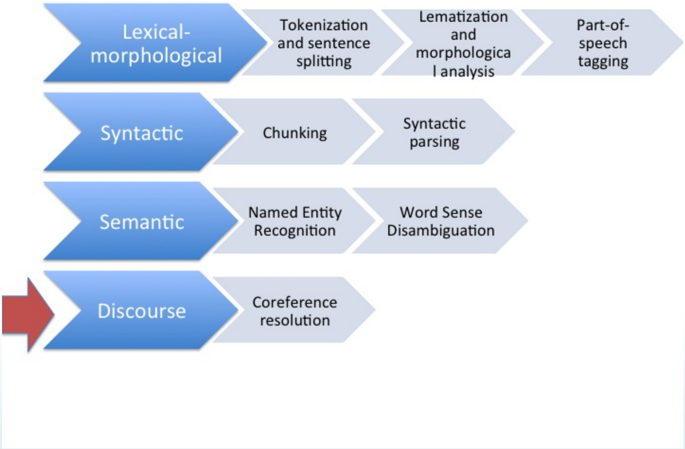
- **Goal:** Determine whether a name refers to a person, organisation, location, etc.

Example

An armed gang has stolen **[CARDINAL four]** paintings worth some $160m by the great painters **[PERSON Cezanne]** , **[PERSON Degas]** , **[PERSON Van Gogh]** and **[PERSON Monet]** from a museum in **[GPE Zurich]** .

- Demo:
  - http://nlp.stanford.edu/software/corenlp.shtml
  - http://cogcomp.cs.illinois.edu/demo/ner/?id=8

How to solve ambiguity: the waterfall model



How to solve semantic ambiguity?

Word Sense Disambiguation

Polysemy: most words have many possible meanings. A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human.

Example

(a) I bought myself a new bass guitar.
(b) They like grilled bass.

Word sense disambiguation

WSD => select the correct sense of an ambiguous word in a given context

Word sense representations

With respect to a dictionary

chair = a seat for one person, with a support for the back; ”he put his coat over the back of the chair and sat down”

chair = the position of professor; ”he was awarded an endowed chair in economics”

With respect to its translation

chair = chaise
chair = directeur

With respect to the context where it occurs (discrimination)

Sit on a chair Take a seat on this chair
The chair of the Math Department The chair of the meeting

Granularity of sense distinctions

- John is rich.
- This is my house.
- Where is my umbrella?
- Is there a God?
- There were two hundred people at his funeral?
- This money is my only income.
- She is our resident philosopher.

Granularity of sense distinctions

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: be [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) beryllium, **Be**, glucinium, atomic number 4 (a light strong brittle grey toxic bivalent metallic element)

Verb

- S: (v) **be** (have the quality of being; (copula, used with an adjective or a predicate noun)) *"John is rich"; "This is not a good answer"*
- S: (v) **be** (be identical to; be someone or something) *"The president of the company is John Smith"; "This is my house"*
- S: (v) **be** (occupy a certain position or area; be somewhere) *"Where is my umbrella?" "The toolshed is in the back"; "What is behind this behavior?"*
- S: (v) exist, **be** (have an existence, be extant) *"Is there a God?"*
- S: (v) **be** (happen, occur, take place) *"I lost my wallet; this was during the visit to my parents' house"; "There were two hundred people at his funeral"; "There was a lot of noise in the kitchen"*
- S: (v) equal, **be** (be identical or equivalent to) *"One dollar equals 1,000 rubles these days!"*
- S: (v) constitute, represent, make up, comprise, **be** (form or compose) *"This money is my only income"; "The stone wall was the backdrop for the performance"; "These constitute my entire belonging"; "The children made up the chorus"; "This sum represents my entire income for a year"; "These few men comprise his entire army"*
- S: (v) **be**, follow (work in a specific place, with a specific subject, or in a specific function) *"He is a herpetologist"; "She is our resident philosopher"*
- S: (v) embody, **be**, personify (represent, as of a character on stage) *"Derek Jacobi was Hamlet"*
- S: (v) **be** (spend or use time) *"I may be an hour"*
- S: (v) **be**, live (have life, be alive) *"Our great leader is no more"; "My grandfather lived until the end of war"*
- S: (v) **be** (to remain unmolested, undisturbed, or uninterrupted -- used only in infinitive form) *"let her be"*
- S: (v) cost, **be** (be priced at) *"These shoes cost $100"*

WordNet home page

Granularity of sense distinctions (ctd.)

The required granularity of sense distinctions might depend on the application.

Example

head (En.) $\Rightarrow$ hoofd (Nl.)

How to solve pragmatic ambiguity?

Coreference resolution

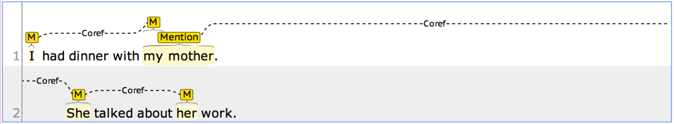
- **Goal:** The meaning, referent of some words, such as pronouns (e.g. “he”, “his”, “her”) depends on the meaning of their antecedent. Automatic anaphora resolution is the task of automatically determining the antecedent of a given anaphor.
- How?: very complex problem. Can only be solved through a combination of lexical and morphological knowledge, syntactic knowledge, semantic and world knowledge.
- Demo:

  http://nlp.stanford.edu/software/corenlp.shtml

Automatic anaphora resolution

Example

Coreference:



Part III

NLP applications

NLP applications

text

To train modules

Lexical/morphological analysis

Tagging
Chunking
Parsing
Word sense disambiguation
Named entity recognition

Semantic role labeling

Coreference resolution

Discourse analysis

To construct LT applications

Spelling correction
Grammar checking

Information retrieval
Document classification
Information extraction
Summarization
Question answering

Sentiment detection
Authorship recognition

Machine translation

etc.

meaning

Part IV

Sentiment analysis

Traditional versus social media

https://www.youtube.com/watch?v=vDGrfhJH1P4

Social media?

Twitter

Facebook

Pinterest

Social media and politics

beheer

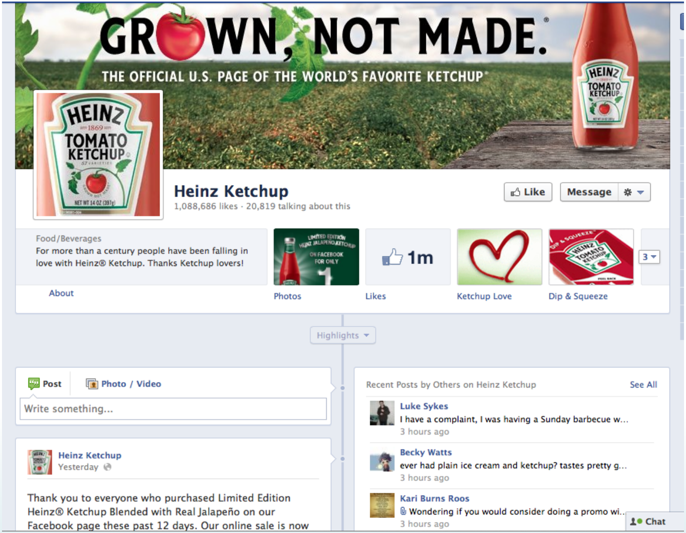
Companies on social media

3. Starbucks Asks for Your Advice

Social media isn't only about using existing websites, but sometimes creating your own. To get a better handle on consumer feedback, Starbucks did just that with "My Starbucks Idea."

The site allows users to submit suggestions to be voted on by Starbucks consumers, and the most popular suggestions are highlighted and reviewed. Starbucks then took it a step further and added an "Ideas in Action" blog that gives updates to users on the status of changes suggested.

Companies on social media



SNS and personalised advertisements

SNS and personalised advertisements

Another application domain

Financial markets

News has a strong impact on the volatility on the financial markets

- Evans en Lyons (2003): news can lead to ca. 30% price variation
- Prast en de Vor (2001): different reaction to “good” and “bad” news
- Bollen et al. (2011): Can Twitter predict the financial markets?

What information do we find on social media?

- mouse clicks
- purchase behaviour
- what people think about products, hotels, etc.
- age
- gender
- region
- photos, blogs, tweets contain a wealth of information!

How can we determine gender?

1. corpus of texts written by male and female authors (e.g. 1000 docs male + 1000 docs female)
2. determine which words are more used by male than by female authors e.g. through calculating **Mutual Information (MI)**
3. the MI values determine for a new text whether it is written by a male or female author

How can we determine gender?

MI

$$MI(t,c) = p(t|c)p(c)log\frac{p(t|c)}{p(t)} \tag{1}$$

$$p(t|c) = \frac{count(c,t)}{count(c)} \tag{2}$$

$$p(c) = \frac{count(c)}{n} \tag{3}$$

$$p(t) = \frac{count(t)}{n} \tag{4}$$

How can we determine gender?

Words with the highest MI values for male and female authors in a MySpace dataset (Caverlee and Webb, 2008):

Male		Female	
dating	sport	love	people
networking	metal	dancing	life
serious	football	shopping	can
relationships	s***	girl	family
single	wars	hearts	being
straight	band	have	notebook
video	f***	are	dance
guitar	gay	favorite	things

How can we determine age?

high	high	college	graduate	networking	parent	parent
school	school	someday	college	graduate	proud	proud
hearts	someday	student	networking	parent	married	president
junior	love	love	grad	proud	networking	swinger
single	best	straight	professional	married	kids	his
best	boy	caucasian	relationship	grad	great	married
hair	ever	white	traveling	professional	our	kids
friend	hair	like	some	art	divorced	united
lol	lol	girl	reading	cure	daughter	began
play	single	know	working	travel	years	retired

Do the test!

http://www.bookblog.net/gender/genie.html

So far ...

text = bag of words

"I poop what I eat"
"I eat what I poop"
(Jurafsky and Martin, 2009)

https://engagor.com

Do we need semantic and syntactic knowledge for sentiment analysis?

Sentiment detection:

Wiebe, J., Wilson, T. & Cardie, C.: 2005, Annotating Expressions of Opinions and Emotions in Language, Language Resources and Evaluation, 164-210)

Tasks:

- Subjective / objective
- Sentiment classification: positive, negative en neutral; even implicit sentiment (Van de Kauter, M., Desmet, B. & Hoste, V.: to appear, The Good, the Bad and the Implicit: A Comprehensive Approach to Annotating Explicit and Implicit Sentiment)
- Detection of opinion strength

Do we need semantic and syntactic knowledge for sentiment analysis?

Most current SA approaches

NO need for semantic and syntactic knowledge.
But:

- fail to model who is positive/negative about what (aspect-based sentiment mining)
- fail to model why someone is positive/negative (argumentation mining)

Automatic sentiment detection

2 approaches:

- Lexicon based
- Corpus based
  3 base components [Liu, 2012]:
  - Opinion holder
  - Object
  - Opinion

  vb. I love Wall-E. I adore the character and the film that tells his story.
  vb. “ik hate my life too”

Automatic sentiment detection

Lexicon based approach

The semantic characteristics of individual words are good predictors of the semantic characteristics of a phrase or text Problem: Need for sentiment lexicons: e.g.
http://www.cs.pitt.edu/mpqa/
http://sentiwordnet.isti.cnr.it/ (SentiWordNet)

Automatic sentiment detection

Learning approach

Cynthia Van Hee, Marjan Van de Kauter, Orphe De Clercq, Els Lefever and Vronique Hoste, "LT3: Sentiment Classification in User-Generated Content Using a Rich Feature Set", CoLING, 2014.

Preprocessing

- Manual replacement of non-UTF-8 characters
- Tokenization & PoS-tagging[2]
- Dependency parsing[3]
- Named Entity Recognition[4]
- *neutral*
  objective
  neutral-OR-objective } *neutral*

Feature Extraction/Groups

- **N-gram** features
  - Word token n-grams (1g, 2g, 3g)
  - Character n-grams (3g, 4g)
  - Normalized n-grams
- **Word shape** features
  - Character flooding (numeric)
  - Punctuation flooding (numeric)
  - Punctuation of the last token (binary)
  - Token capitalization (numeric)
  - Hashtags (numeric)
- Sentiment **lexicons** features (numeric)
  - Three general (AFINN, GenInq, MPQA)
  - Three Twitter-specific (NRC, Bing Liu, Bounce)
  - One emoticon list (based on training)
- Syntactic features
  - **PoS**-tags (binary, ternary, absolute, frequency)
  - **Dependency** relations (binary)
- **Named entity** features (binary, absolute, absolute & frequency tokens)
- **PMI** features based on NRC lexicon & training data (numeric)

Automatic sentiment detection

Try it out yourself!

- http://text-processing.com/demo/sentiment/

But what about Mike?

Example

my name is mike iv been depressed for 3 years im 15 nd i fell in love with a beautiful girl named sierra every day i told her how much i loved her nd what id do for her but one day in 7th grade when i told her i love her she said but i dont love u nd to leave her alone ever sense iv been horribly depressd iv tried to kil myself 4 times but never succeeded i hate my life if anyone can help please contact me at (...)

Problematic for both the lexicon-based and learning-based approach.
The same problem, although less clear, exists in case of genre shifts.

Part V

Machine translation from a different angle . . .

But what about Mike?

Example

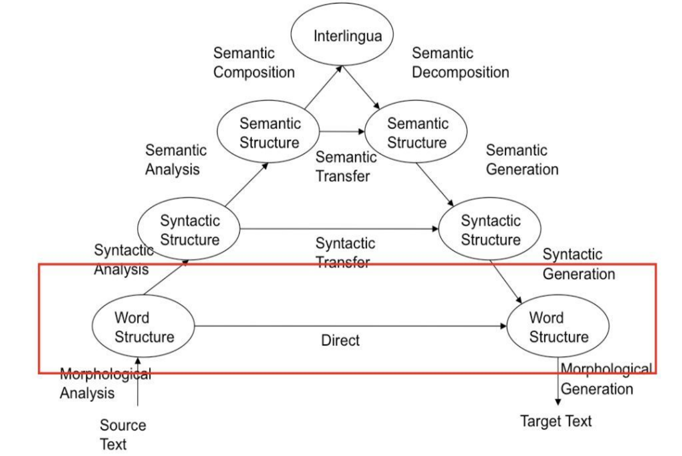
my name is mike iv been depressed for 3 years im 15 nd i fell in love with a beautiful girl named sierra every day i told her how much i loved her nd what id do for her but one day in 7th grade when i told her i love her she said but i dont love u nd to leave her alone ever sense iv been horribly depressd iv tried to kil myself 4 times but never succeeded i hate my life if anyone can help please contact me at (...)

Example

My name is Mike. I've been depressed for 3 years. I'm 15 and I fell in love with a beautiful girl named Sierra. Every day I told her how much I loved her and what I would do for her, but one day in 7th grade, when I told her I love her, she said: "But I don't love you" and to leave her alone. Ever since, I have been horribly depressed. I have tried to kill myself 4 times but never succeeded. I hate my life. If anyone can help, please contact me at (...)

Two possible methodologies: rules

[Dorr, Bonnie 2007]



Two possible methodologies: statistics

- Corpus based
- Language independent
  - Fast prototype
- Fully automatic alignment of words and phrases
- Probabilities are automatically determined on the basis of the training data
- Phrase-based SMT = state-of-the-art

SMT

Base principles

- SMT wants to maximise two factors:
  - **Faithfulness** = how close is the meaning of the translation to the original?
  - **Fluency** = how fluent is the translation?

SMT

Three components in an SMT system

- **Translation model**
  - Higher probability for sentences with the same meaning
  - Use of **bilingual** corpora to estimate probabilities
- **Language model**
  - Higher probability for grammatically correct sentences
  - Use of **monolingual corpora** to estimate probabilities
- **Decoder**
  - Combines translation and language model
  - Searches the sentence with the highest probability
  - Probability of target language sentence T, given source language sentence S
    $P(T|S) = \text{Faithfulness}(S,T) * \text{Fluency}(T)$

SMT

Taalmodel

- Component which takes care of word order
- Probabilities are estimated on the basis of large monolingual corpora in the target language
- **N-gram**models: standard is a **trigram** language model
  - sometimes bigger n-grams
  - sparde
- Trigram probabilities
  - $p(w_3 \mid w_1 w_2) = \frac{Count(w_1 w_2 w_3)}{Count(w_1 w_2)}$

SMT

Example

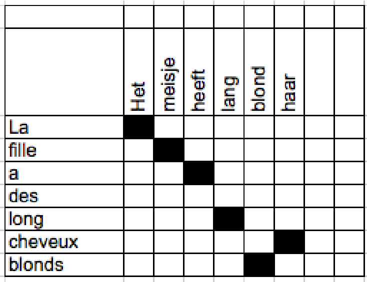
- Given 2 sentences
  1. That car was almost crash onto me
  2. That car almost hit me
- How can you quantify that 1 sentence is worse than sentence 2?
- Trigram model
  - p(That car almost hit me) $>$ p(That car was almost crash onto me)
  - p(That car was almost crash onto me) = p(That$|$ $\emptyset$ $\emptyset$) x p(car$|$ $\emptyset$ That) x p(was$|$That car) x p(almost$|$car was) x p(crash$|$was almost) x p(onto$|$almost crash) x p(me$|$crash onto)
  - p(That car almost hit me) = p(That$|$ $\emptyset$ $\emptyset$) x p(car$|$ $\emptyset$ That) x p(almost$|$That car) x p(hit$|$car almost) x p(me$|$almost hit)

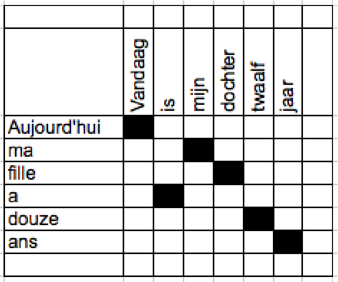
SMT

Quantify faithfulness

- How close is the meaning of the translation to the meaning of the original?
- Example
  - “Dat bevalt me”
  - (1) “That pleases me”
    (2) “I like it”
    (3) “I’ll take that one”
- How to quantify automatically?
  - Intuition: degree in which words in the source and target sentence are translations
  - Formal: product of the probabilities
  - Based on word alignment on parallel corpora

SMT

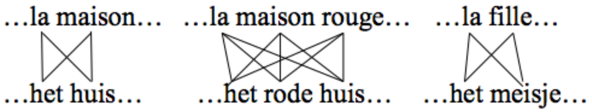

Het
meisje
heeft
lang
blond
haar
La
fille
a
des
long
cheveux
blonds


Vandaag
is
mijn
dochter
twaalf
jaar
Aujourd'hui
ma
fille
a
douze
ans

SMT

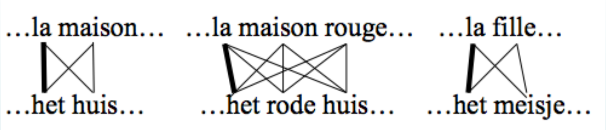
Initialisation: uniform distribution

- Link every source language word to every target language word
- Every word alignment in equally plausible



After 1 iteration

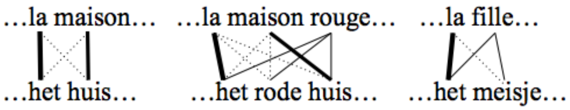
- Model learns that word pair “la” – “the” occurs often



SMT

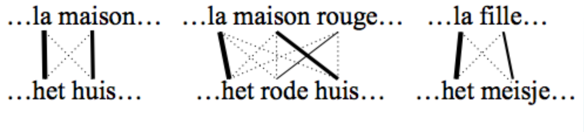
Next iteration

- The alignment "maison" – "house" becomes more plausible



End of process

- Until model converges (mostly after 4 to 5 steps)



SMT

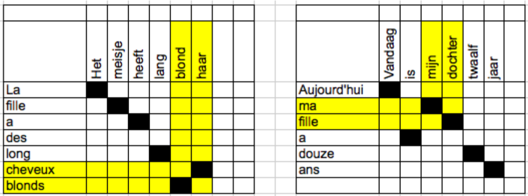
IBM models

- Model 1: lexical model
- Model 2: can take into account absolute order
  - E.g. probability that a word on position three is translated by a word at position 6
- Model 3: adds a fertility model
  - E.g. probability that word x is translated by 3 words
- Model 4: can take into account relative order

SMT

Phrase based translation models

- Phrases induced from word alignments



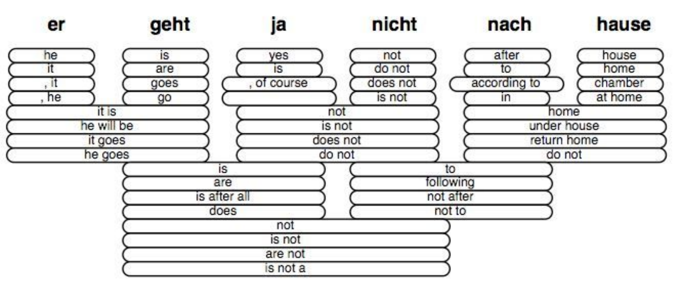
SMT

Decoding = search process

- Search all translations of all “phrases” in the phrase table
- Find the optimal combination: maximise translation prob * language model prob.

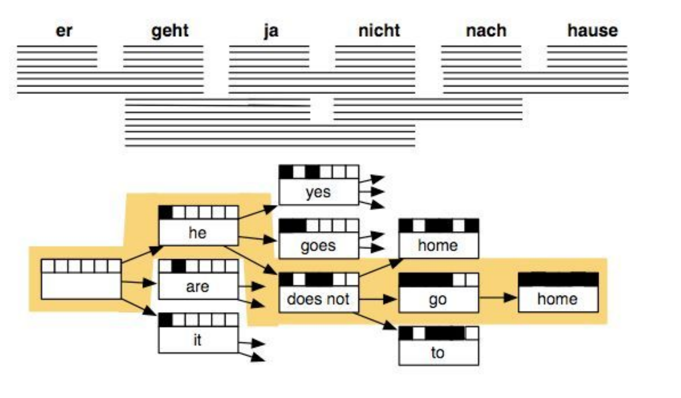
SMT

Decoding



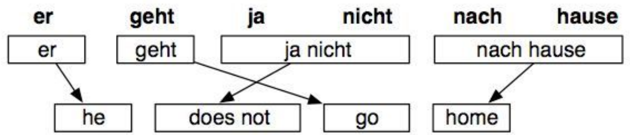
SMT

Decoding



SMT

Decoding – best combination



SMT

Try it out yourself!

- Google Translate has an API
- Moses open source SMT toolkit

Part VI

Computational Stylometry

Wat is jouw schrijfstijl?
http://www.clips.ua.ac.be/cgi-bin/stylene.html

Computational stylometry

- What can we learn from text, not about its content, but about its author and the context of writing (meta)?
- Is there a "human stylome", a linguistic fingerprint, that can be measured, is largely unconscious, and is constant?
- Is text categorization the right hammer for this nail?

Definition of Style

- There is linguistic variability in text (different ways of expressing the same)
- Variation due to “style”: A combination of specific, invariant decisions in language generation at all linguistic levels (discourse, syntactic structures, lexical choice, ...)
- Computational stylometry is the attribution of texts to individual authors using computational models that recognize writing style

The human stylome

- If style is unique (like fingerprint or genome):(...) authors can be distinguished by measuring specific properties of their writings, their stylome as it were (Van Halteren et al., JQL, 2005)
- Application: Authorship Attribution

Other sources of variation

- Language variability is also a property of groups of individuals:
  - People with the same personality, Alzheimer disease patients, Schizophrenia patients, depressed people, ...
  - People of the same gender, age, education level, region of language acquisition, non-native speakers, ...
- Studied in sociolinguistics: youth language, gender studies ...
- Studied in language psychology: effect of Alzheimer on language use, personality and language use, ...

More sources of variation

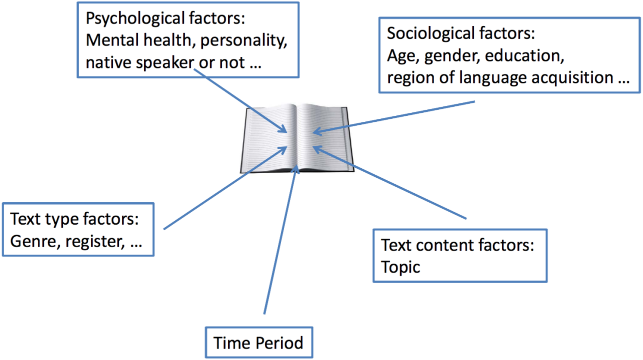
- Register and genre: level of formality (letter, academic text, essay, blog, manual, description, narrative, persuasive text, poetry, ...) Application: genre categorization
- Domain: what the text is about (Sports, economics, Belgian politics, ...) Studied in information retrieval. Application: text categorization (classical)

More sources of variation

- Period in which a text was written: Language (vocabulary) change, spelling change,
  Studied in literary science, philology
  Application: text datingDetection of anachronism

Does a human stylome exist?

Author stylome is composition of several sociological and psychological variables and interacts with variation due to topic, register, genre, etc. How can they be disentangled?



How do we measure the stylome?

- Some authorship attribution experiments are very successful. But what does discriminating between two or a few authors tell us about their stylome?
- Doesnt provide an explanation, its a trick!
- In which linguistic information do we find the stylome? Character n-grams, lexical properties, syntactic properties, metaphorical language use, discourse properties, combination of all ...
- Is the stylome beyond conscious control or can it be faked?
- Is authorial style constant over time?

Faking style

- Style = beyond conscious control?
- Can you make your style unrecognizable?
  - Use Machine Translation (works, but bad idea anyway)
  - Change your text in the features that are mostly used in stylometry (function words)
  - e.g. Gilbert Adairs pastiche of Alice in Wonderland. Some techniques unable to differentiate between both
  - e.g. Brennan and Greenstadt (2009): work on deception

weka!

Table V: The table shows performance of different feature sets in detecting regular and adversarial writing samples. The Writeprints feature set with SVM classifier provides the best performance in detecting deception.

Dataset	Feature set, Classifier	Type	Precision	Recall	F-measure	Overall F-measure
Extended-Brennan-Greenstadt	**Writeprints, SVM**	**Regular**	**97.5%**	**98.5%**	**98%**	**96.6%**
		Imitation	**87.2%**	**82.9%**	**85%**	
		Obfuscation	**93.2%**	**86.1%**	**89.5%**	
	Lying-detection, J48	Regular	95.2%	96.2%	95.7%	92%
		Imitation	80.6%	70.7%	75.3%	
		Obfuscation	60.3%	59.5%	59.9%	
	9-feature set, J48	Regular	92.3%	96.8%	94.5%	89%
		Imitation	52.9%	43.9%	48%	
		Obfuscation	61.9%	32.9%	43%	
Amazon Mechanical Turk	Writeprints, SVM	Regular	96.5%	98.6%	97.5%	95.6%
		Imitation	82.3%	72.9%	77.3%	
		Obfuscation	96.4%	79.1%	86.9%	
	Lying-detection, J48	Regular	94.2%	96.2%	95.2%	90.9%
		Imitation	71.7%	54.3%	61.8%	
		Obfuscation	58.5%	56.7%	57.6%	
	9-feature set, J48	Regular	92.5%	96.3%	94.3%	88%
		Imitation	45.5%	35.7%	40%	
		Obfuscation	45.5%	29.9%	36%	
Brennan-Greenstadt	Writeprints, SVM	Regular	94%	100%	96.9%	94.7%
		Imitation	100%	83.3%	90.9%	
		Obfuscation	100%	50%	66.7%	
	Lying-detection, J48	Regular	90%	92.9%	91.4%	85.3%
		Imitation	90.9%	83.3%	87%	
		Obfuscation	11.1%	8.3%	9.5%	
	9-feature set, J48	Regular	89.4%	93.7%	91.5%	84%
		Imitation	25%	25%	25%	
		Obfuscation	83.3%	41.7%	55.6%	

Constancy of style

Does style remain constant over time in an individual?

Unlikely

- People change
- Language changes (vocabulary, spelling, ...)
- Vocabulary growth and decline
- Grammatical decline (See nuns study)
- Age groups have specific characteristics (see yesterday). Books written by same author in different periods can be discriminated (Can Patton, 2004)

Practical problems

- Simple hypotheses dont work: e.g. style = function words, topic = content words

Very Brief History: Phase 1 (from 19th century)

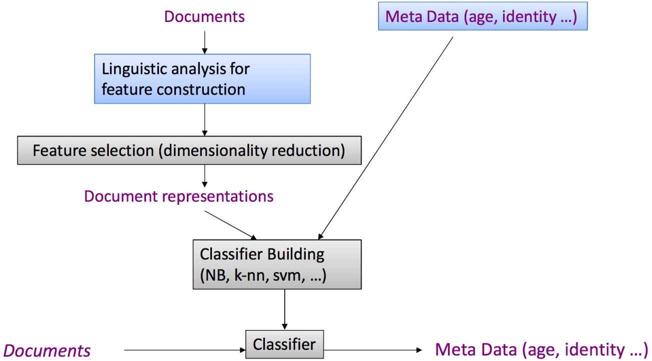
- Expert-based, manual approach to authorship attribution
- E.g. Shakespeare studies
- Unmasking the Unabomber: Theodore Kaczynski (Professor Mathematics Berkeley). Bomb letters against universities and airlines. Unabomber manifesto (35,000 words). Anti-technology. Word use recognized by family member

Very Brief History: Phase 2

- Step to quantitative data (computed by hand or computer-aided)
- Seminal study: The federalist papers (Mosteller Wallace 1964)
    - 64 essays, 3 authors (Jay, Hamilton, Madison) (1787-1788)
    - Bayesian analysis using highly frequent words
    - From single features (word length, vocabulary richness, etc.) to feature combinations
- “Silver bullet” feature does not exist
- Increasing attention for evaluation and benchmark construction, E.g. Frank Juola benchmarks (2004)

Very Brief History: Phase 3

- Text categorization model
- e.g. Argamon et al., Stamatatos,
- Machine Learning
- Extension of the model from authorship attribution to profiling (gender, age, etc.)
- More realistic set-ups: many authors, short texts


Documents
Meta Data (age, identity …)
Linguistic analysis for feature construction
Feature selection (dimensionality reduction)
Document representations
Classifier Building (NB, k-nn, svm, …)
Documents
Classifier
Meta Data (age, identity …)

Applications

- Literary Research: Disputed authorship
- Forensics
- Profiling (gender, age, ...) + link opinion to profiles
- Plagiarism
- Suicide / blackmail letters
- Link ideology and cultural behaviour to profiles
- Monitoring, e.g. pedophiles

Forensics

Solving a murder with linguistics?

- 1992: student Michael Hunter dies after an injection with lidocaine, Benadryl and Vistaril.
- Suicide letters are found, both printed and on his computer
- Analysis of syntactic patterns in the letters and in other texts of the victim and of his roommates is performed by Carole Chaski (http://www.linguisticevidence.org). She concludes that the letters could not be written by Hunter (99.9% certainty) and were probably written by one of the roommates.
- Roommate gets 7 years for manslaughter

Authorship Attribution versus Verification

- Attribution is easy: Given texts of author A and B and an unknown text X, decide whether it was written by A or B
- Verification is difficult: Given a text X and a candidate author A, decide whether A has written X. No negative information: Problematic in learning

Authorship Verification

- e.g. Koppel et al. 2007
- Did A (e.g. suspect) write X? A = Text of our author of interest; X = our mystery text
- Train classifier to discriminate between A and X
- Train classifiers to discriminate between A and impostors (similar authors)
- Iteratively remove most informative features and retrain and retest
- Compare the curves
- Slow and gradual deterioration: then impostors; Sudden and dramatic deterioration: then same author
- If written by same author, then number of differences will be relatively low

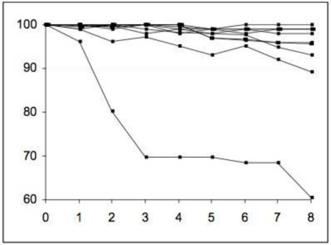
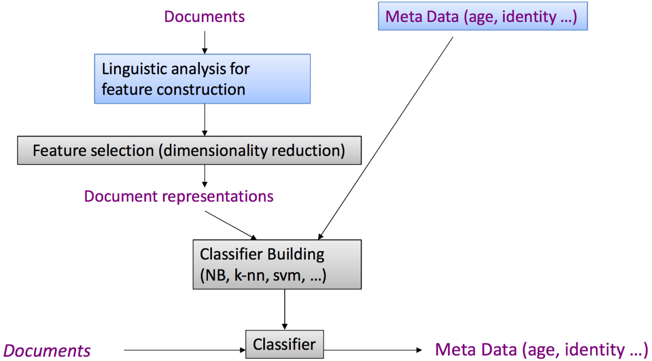


Figure 2. Unmasking *An Ideal Husband* against each of the ten authors (n=250, k=3). The curve below all the authors is that of Oscar Wilde, the actual author. (Several curves are indistinguishable.)


Documents
Meta Data (age, identity …)
Linguistic analysis for feature construction
Feature selection (dimensionality reduction)
Document representations
Classifier Building (NB, k-nn, svm, …)
Documents
Classifier
Meta Data (age, identity …)

Stylometry document representations

- Select information sources and construct features (automatic text analysis)
- Features: Character level, word level, syntactic structure, semantics, document level (readability, text structure). N-grams, frequency (distributions). Anything you can think of!
- Feature selection: infogain / chi-square / frequency (bands) / tf.idf

A brief catalogue of stylometry features

- Letter frequency
- Punctuation
- Character n-grams
- Complexity (readability, word length, sentence length, ...)
- Syllable length, word length, sentence length (averages or distributions)
- Vocabulary richness: type token ratio
- Word frequency distributions: function words, content words, frequent words, pronouns, ...
- etc.

A brief catalogue of stylometry features

- Morphology: prefixes and suffixes
- Syntax: POS tag (distributions), chunks
- Semantics: semantic subclasses (wordnet), case frame distributions
- Stable words (stay the same if translated and translated back, will probably survive editing)
- etc.

Why do char n-grams often work so well?

- Good trade-off between sparseness and information
- Implicit punctuation, morphology, semantics, lexical items (function words are often short words)
- Tolerant to errors (two spelling variants still share many character n-grams)

Training data representation

- Profile-based (all texts are combined into one training text from which the features are extracted)
- Text-based (different texts receive a different representation / feature vector)
- Can be artificially created: e.g. segments of 1000 words, paragraphs, even sentences

One more example to finish

Personality from Text

- Personae corpus: Collected November 2006; 200,000 words, Dutch
- 145 BA students (from a population of 200) in a course on interdisciplinary linguistics
- Voluntarily watched the same documentary on Artificial Life (but received 2 cinema tickets as incentive). Topic, genre, register, age held constant.
- Wrote a text of   1200 words: Factual description + Opinion
- Did an on-line personality test
- Submitted their profile, the text and some user information via a web-site
- All text processed with Tokenizer / Tagger / Chunker / Relation Finder

Are personality traits such as extraversion reflected in writing style?

- Seminal work by Gill and Oberlander on extraversion and neuroticism
- Parallel work on prediction: Argamon et al., 2005; Nowson Oberlander, 2007; Mairesse et al., 2007, ...
- Previous hypotheses and observations:
    - Extraverts use fewer hedges (confidence)
    - More verbs, adverbs and pronouns (vs. nouns, adjectives, prepositions)
    - Less formal
    - Fewer negative emotion words; more positive emotion words
    - More present tense verbs
    - etc.

Meyers-Briggs Forced-choice test

- Carl Jungs personality typology
- Categorization according to 4 preferences: Introversion Extraversion (attitudes) ; iNtuition Sensing (information-gathering) ; Feeling Thinking (decision-making) ;Judging Perceiving (lifestyle)
- Leads to 16 types: ENTJ (1.8%) ... ESFJ (12.3%). Validity and reliability have been questioned
- Typical student: Flemish girl from around Antwerp who likes people and is warm, sympathetic, helpful, cooperative, tactful, down-to-earth, practical, thorough, consistent, organized, enthusiastic, and energetic. She enjoys tradition and security, and will seek a stable life that is rich in contact with friends and family.

Task	Feature set	Precision	Recall	F-score
Introverted	word 3-grams	87.69%	58.16%	69.94%
	random	*44.1%*	*46.2%*	
Extraverted	POS 3-grams	100.00%	56.74%	72.40%
	random	*54.6%*	*52.5%*	
iNtuitive	POS 3-grams	84.62%	64.71%	73.33%
	random	*48.7%*	*48.7%*	
Sensing	POS 3-grams	85.07%	56.44%	67.86%
	random	*40.3%*	*40.3%*	
Feeling	readability	100.00%	73.43%	84.68%
	random	*72.6%*	*73.3%*	
Thinking	word 2-grams	72.50%	39.19%	50.88%
	random	*28.2%*	*27.5%*	
Judging	word 3-grams	81.82%	100.00%	90.00%
	random	*77.6%*	*76.9%*	
Perceiving	word 2-grams	60.71%	36.96%	45.95%
	random	*6.9%*	*7.1%*	

Getting Started yourself

- Linguistic analysis using NLP tools (tokenization, PoS tagging, lemmatisation, parsing, etc.)
- Make feature vectors
- Use weka