CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Scaling up without dumbing down

Walter Daelemans
CLiPS Computational Linguistics Group
http://www.clips.ua.ac.be/~walter

CRISSP Seminar May 19, 2014

Structure of the talk

- From NLU to Text Mining
  - Relations between Named Entities
- The problem of scalability
  - Scaling up by dumbing down?
- Becoming less dumb
  - Biograph (biomedical text mining)
    - Deeper text analysis
  - AMiCA (detecting harmful events in social networks)
    - From text categorization to scenario extraction

Background

- General Linguistics (F.G. Droste) & Psycholinguistics (Franz Loosen)
- From mid 1980s
  - Nijmegen Language Technology project (Gerard Kempen)
    - "Scruffy" AI (frames & rules)
    - PhD 1987 (Leuven): Object-Oriented KR for Dutch phonology and morphology
  - AI-LAB Brussels (Luc Steels)
    - Neater scruffy KR (components of expertise, reflection, KRS)
    - Machine Learning (symbolic, genetic algorithms, neural networks)
    - [1988 DATR, Gerald Gazdar, TFS, UBG]

Background

- 1990s
  - Tilburg University (Harry Bunt)
    - "Pioneering" of Machine Learning of NLP in Europe
      - SIGNLL (with David Powers), CoNLL & CoNLL shared task
    - Foundation of ILK research group
      - Memory-Based Language Processing
      - Text analysis components (morphology, syntax, semantics)
- 2000s
  - University of Antwerp
    - "Text Mining"

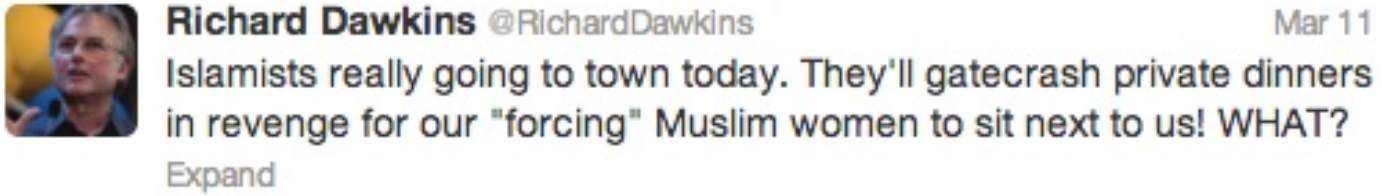
CNTS
CENTRUM VOOR NEDERLANDSE TAAL EN SPRAAK

  - ILK + CNTS / CLiPS
    - Exemplar-Based Models of Language Acquisition and Language Processing
      - With Antal van den Bosch (ILK, now Nijmegen) & Steven Gillis (CNTS & CLiPS)

CLiPS
COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER

Knowledge from Text

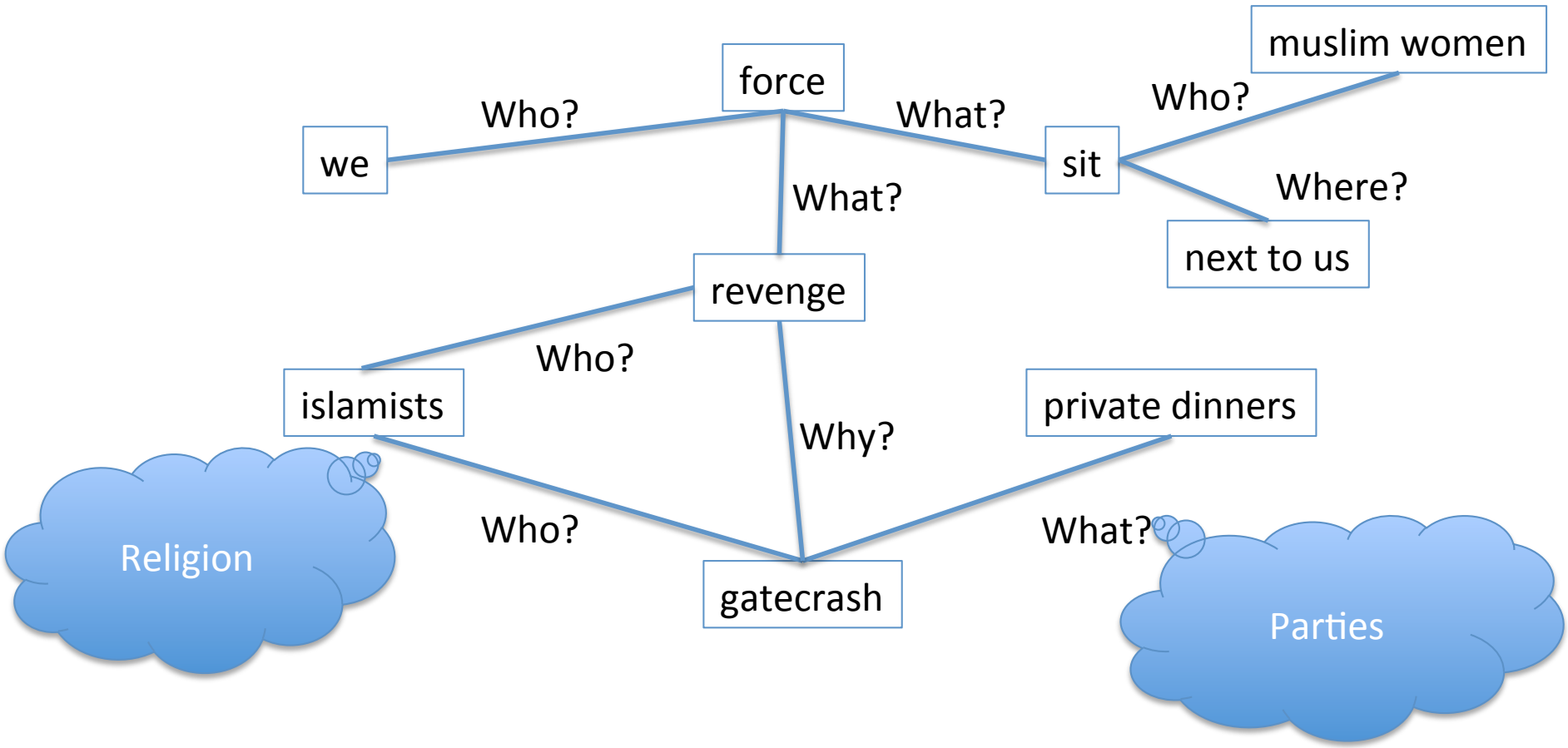
- Contents (Text Mining, Text Analytics)
  - Objective
    - Facts, concepts, properties of concepts, relations between concepts, events, …
      - Who does what, when, where, how and why?
  - Subjective
    - Opinion, sentiment
      - Who thinks what about what?
- Profiles (“metadata”)
  - Authorship, age, gender, personality, …
    - What do we know about the author of a text?



Richard Dawkins @RichardDawkins Mar 11

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Expand



Objective

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us!

WHAT?

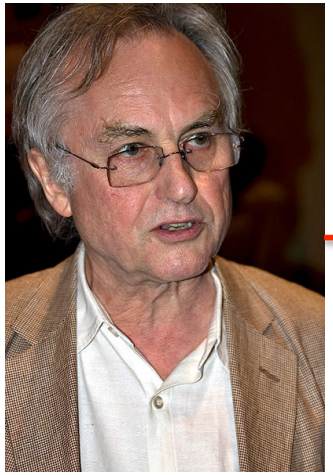
4 events, 5 concepts, 9 relations (one causal)

Coreference resolution

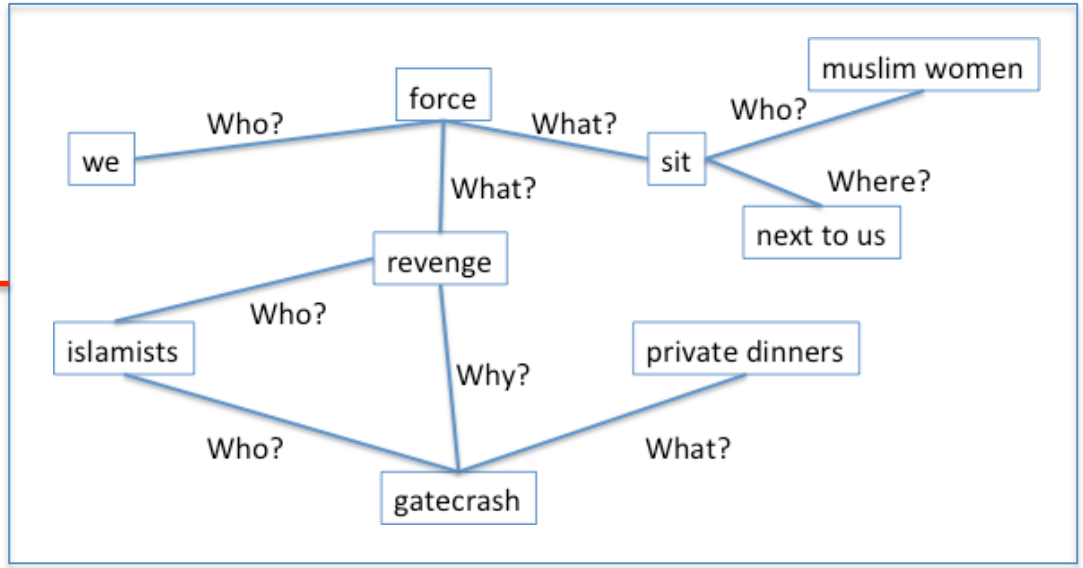
Richard Dawkins @RichardDawkins Mar 11

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Expand



Indignation
Derision
Anger



Subjective

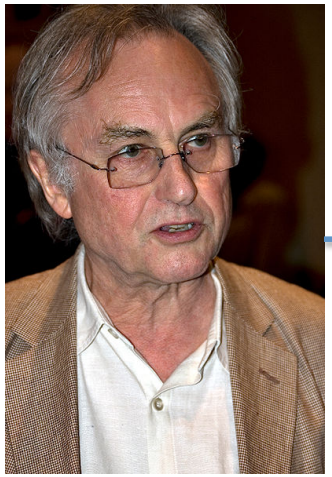
Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Slang, derisory quotes, uppercase letters, exclamation mark, …

Richard Dawkins @RichardDawkins Mar 11

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Expand



Man
Highly educated
60+
BrE native speaker
Introverted

On the basis of a set of tweets

Profiling

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Punctuation, uppercase, function words (especially pronouns) …

Extraction of deeper knowledge

www.biograph.be

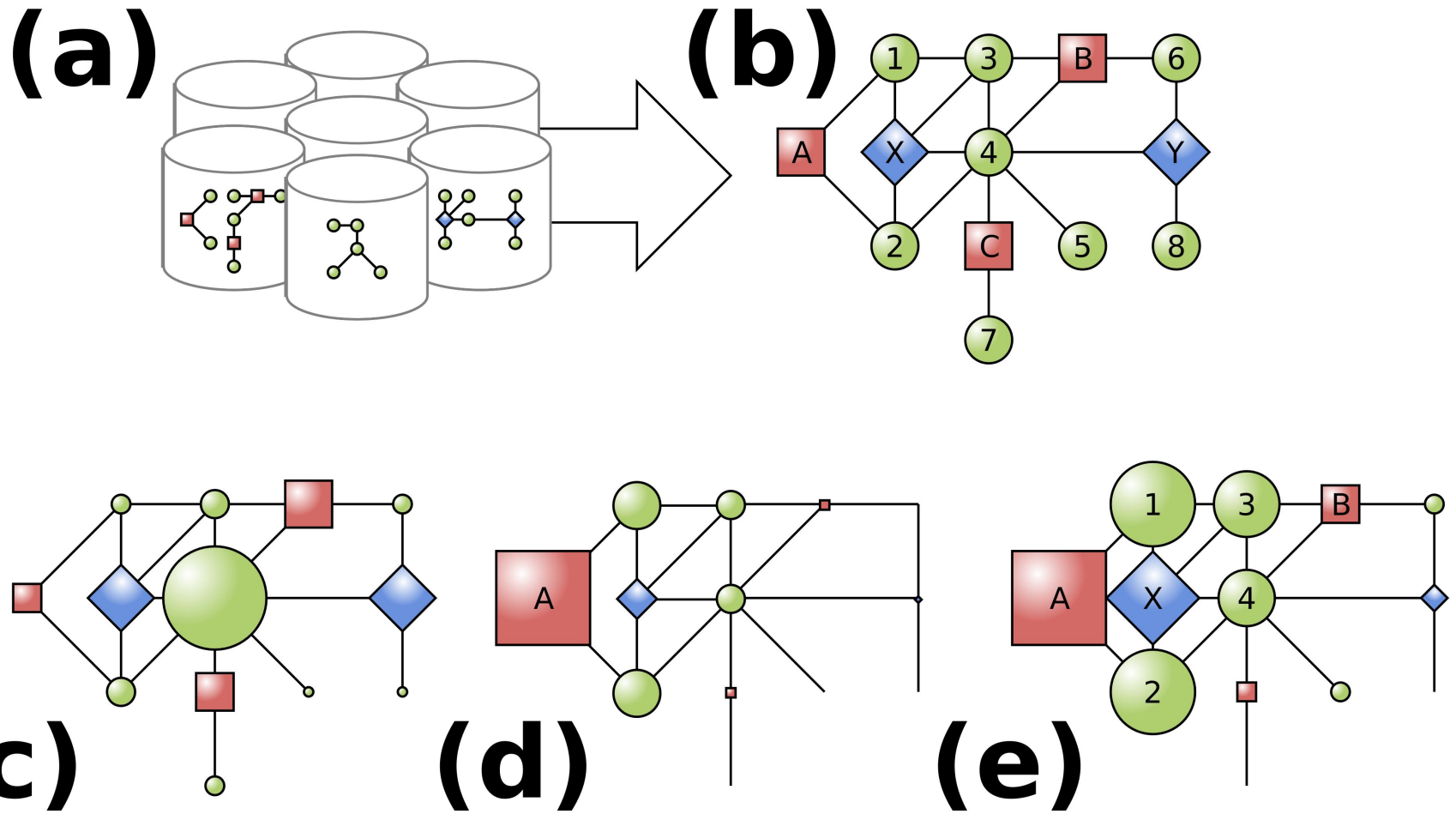
- Funded by University of Antwerp (BOF-GOA)
- CLiPS (text mining); ADReM (graph data mining); AMG (molecular genetics)
- Goals
  - Assisting researchers in ranking candidate genes that cause disease
  - Providing accurate relations automatically extracted from databases and text and weighted according to their reliability
    - Negation and modality

Gene Prioritization

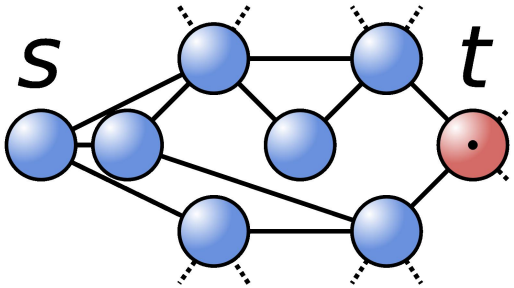
- Candidate region
  - Genes responsible for a disease (e.g. schizophrenia or Alzheimer) are in known areas of the genome
  - Many genes (> 200) are typically in such a candidate region
    - Very expensive to validate experimentally
- Combine information in literature and in databases
  - Which genes in the candidate region could be most relevant for the disease and why?
    - ranking problem
  - Find *indirect* functional relations between the disease and its putative disease genes (explanation)

Graph Datamining

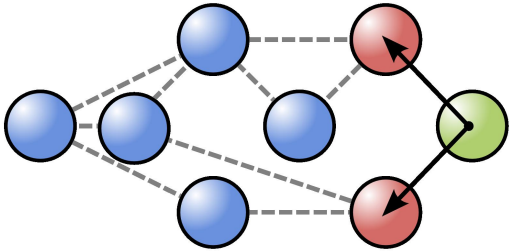
A.M.L. Liekens, J. De Knijf J, W. Daelemans, B. Goethals, P. De Rijk, J. Del-Favero J, BioGraph: Unsupervised Biomedical Knowledge Discovery via Automated Hypothesis Generation, *Genome Biology* 12, 2011



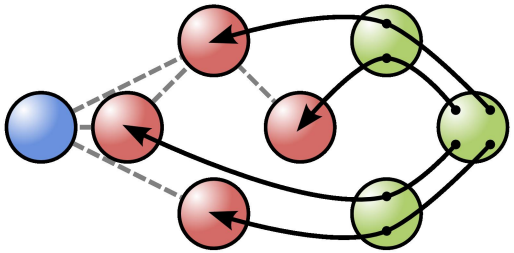
Graph Datamining



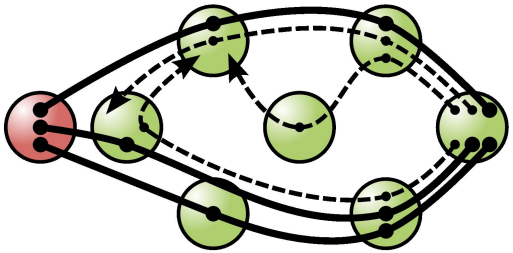
(a)



(b)



(c)



(d)

Text Relations in the Biograph

- Extract (positive) relations of any kind between biomedical concepts found in biomedical abstracts and full papers
- Add to the Biograph
- Evaluation
  - Can text-mined knowledge add to the curated database information?
  - Can text-mined knowledge replace the need for curated database information?

BiographTA Modules

http://www.clips.uantwerpen.be/BiographTA/

- Tokenizer
- Named Entity Recognition (NER)
- Abbreviation handling
- Heuristics for ambiguity handling
- Lemmatization and Parsing
- Supervised Relation Extraction

Example

- Interaction of Munc-18-2 with syntaxin 3 controls the association of apical SNAREs in epithelial cells

- Named entities
- Processes
- Relations

Relation Extraction

- Input: pair of named entities in sentence
- Output: true or false (relation or not)
- Features (output of pipeline)
  - Syntactic and morphological features of NEs and their local context, distance between them, patterns between them
- 83% f-score when trained on BioInfer corpus

Handling Uncertainty

- Negation and Speculation

When U937 cells were infected with HIV-1, no [induction of NF-KB factor was detected], whereas high level of progeny virions was produced, suggesting [that this factor was not [required for viral replication]].

Negation and Speculation module

- Determines whether an extracted relation is in the scope of negation or speculation
- Input: parsed sentence
- Output:

  1 (extracted relation not negated or hedged)

  0 (extracted relation in the scope of speculation)

  -1 (extracted relation in the scope of negation, but no speculation)

Negation and Speculation

- Trained on BioScope corpus
- Two classifiers (TiMBL)
  - Cue detection (including multiword)
    - Features: lexical, syntactic, dictionary (cue lists)
  - Scope (classify words as being in the scope or not)
    - Features: lexical, syntactic, relative position
    - Postprocessing
      - Filter impossible scopes

Results

	Precision	Recall	F-score
Speculation: cue	79	75	77
Speculation: scope	60	55	57
Negation: cue	94	90	92
Negation: scope	66	65	65

R. Morante and W. Daelemans. A metalearning approach to processing the scope of negation. Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 21-29, Boulder, Colorado, June 2009. Association for Computational Linguistics.

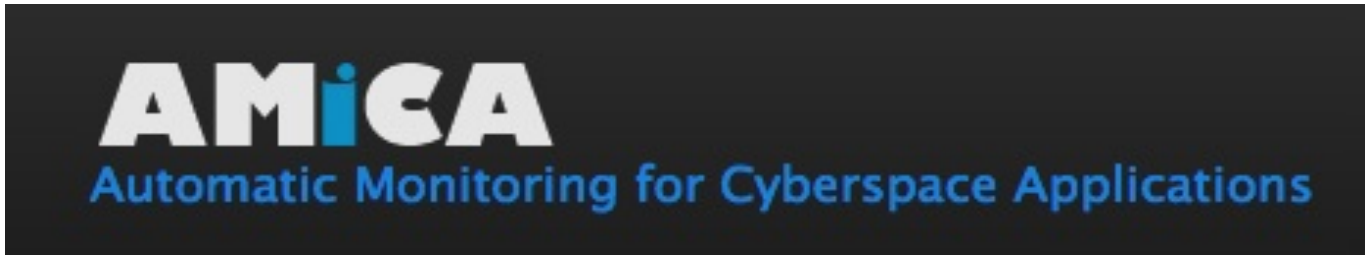
R. Morante, V. Van Asch, and W. Daelemans. Memory-based resolution of in-sentence scopes of hedge cues. Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, pages 40-47, Uppsala, Sweden, 2010. Association for Computational Linguistics.

Evaluation

- Benchmark dataset known gene – disease associations (Endeavour)
  - 627 genes known to cause 29 diseases
  - Remove direct links between genes and (related) diseases from the biograph
- Biograph mean AUC 93% vs. 87% Endeavour

Biograph Evaluation

- Results
  - Biograph (no text) AUC 80.8
  - Biograph (with text) AUC 80.9
  - Biograph (text only) AUC 69.4
- AUC doesn't give the complete picture
  - More qualitative analysis needed
  - Anecdotal evidence for interesting missing and new functional relations (as well as for nonsensical ones)
    - E.g. a link between ovary cancer and BRCA2 was not in the databases but found by the text mining

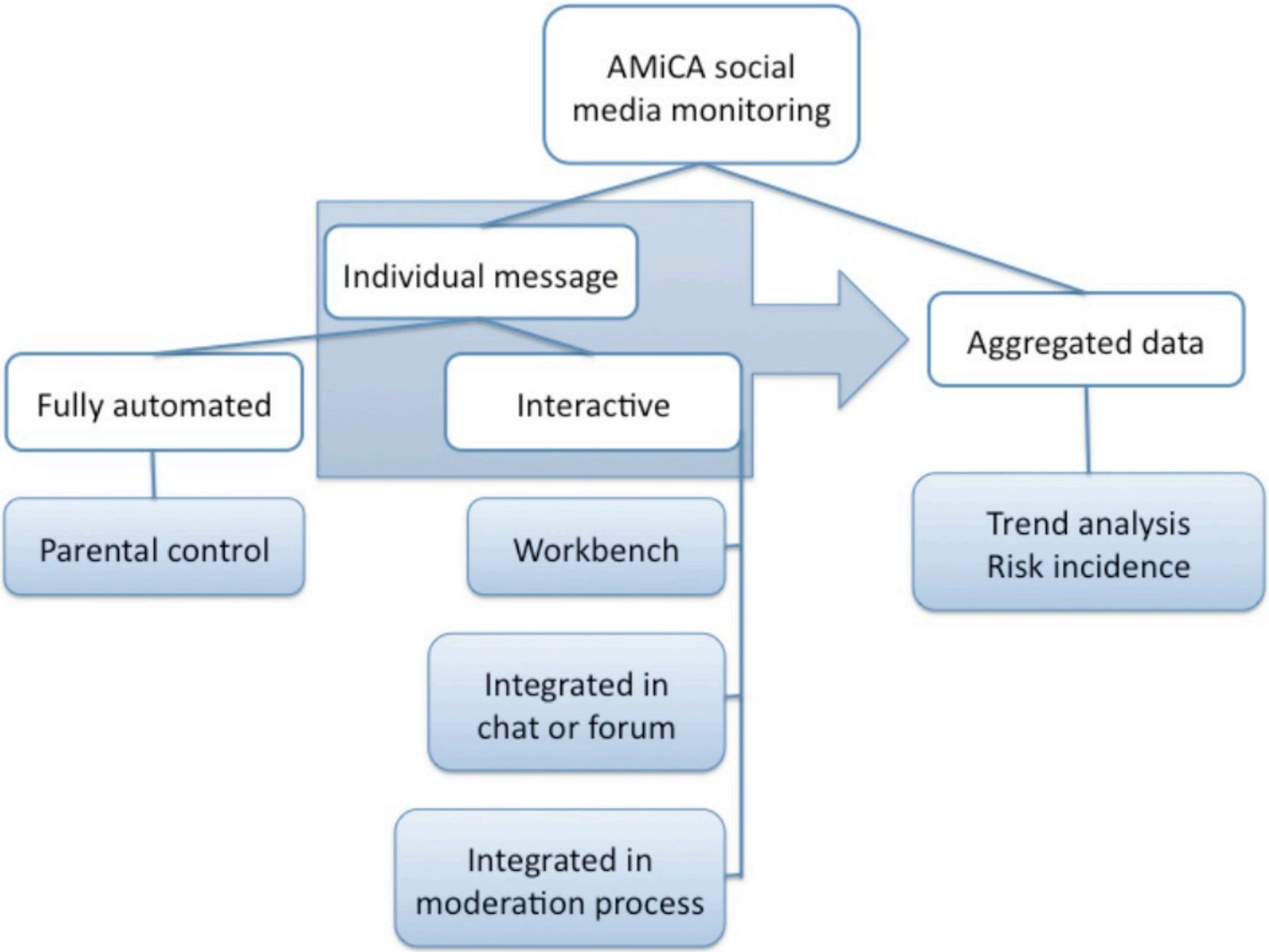


Toward event and script extraction

www.amicaproject.be

AMiCA
Automatic Monitoring for Cyberspace Applications

- IWT project coordinated by CLiPS (text mining) with MIOS (sociology), LT3 (text mining), IBCN (Software), and VISICS (image processing)
- Goals
  - Detect situations that are harmful or threatening to young people in social networks
    - Cyberbullying
    - Sexually transgressive behaviour (for example grooming by paedophiles)
    - Depression and suicide announcement
  - Efficient action by moderators, police, parents, peer group, social services, ...
  - Objective measurement, monitoring, trend analysis, …


AMiCA social media monitoring
Individual message
Aggregated data
Fully automated
Interactive
Parental control
Workbench
Trend analysis
Risk incidence
Integrated in chat or forum
Integrated in moderation process

Approach

- Combine text analysis, image and video analysis, and machine learning
- Computational Stylometry
  - Based on specific language variation, predict author characteristics
    - mental health, personality, deception, native speaker, age, gender, region, educational level, …

Computational Stylometry

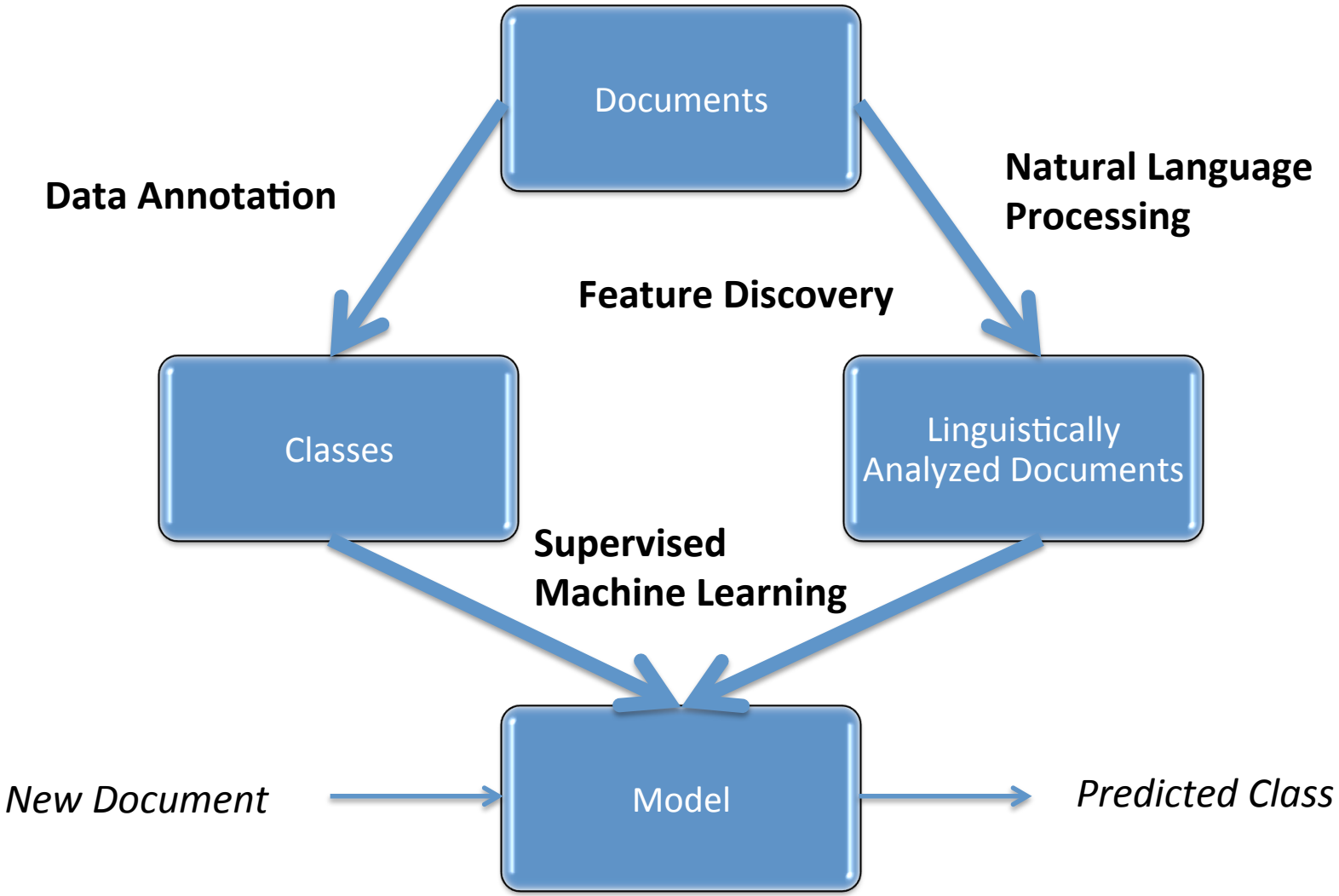


Figure 3. Words, phrases, and topics most highly distinguishing females and males.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0073791

Landmark: gender from text

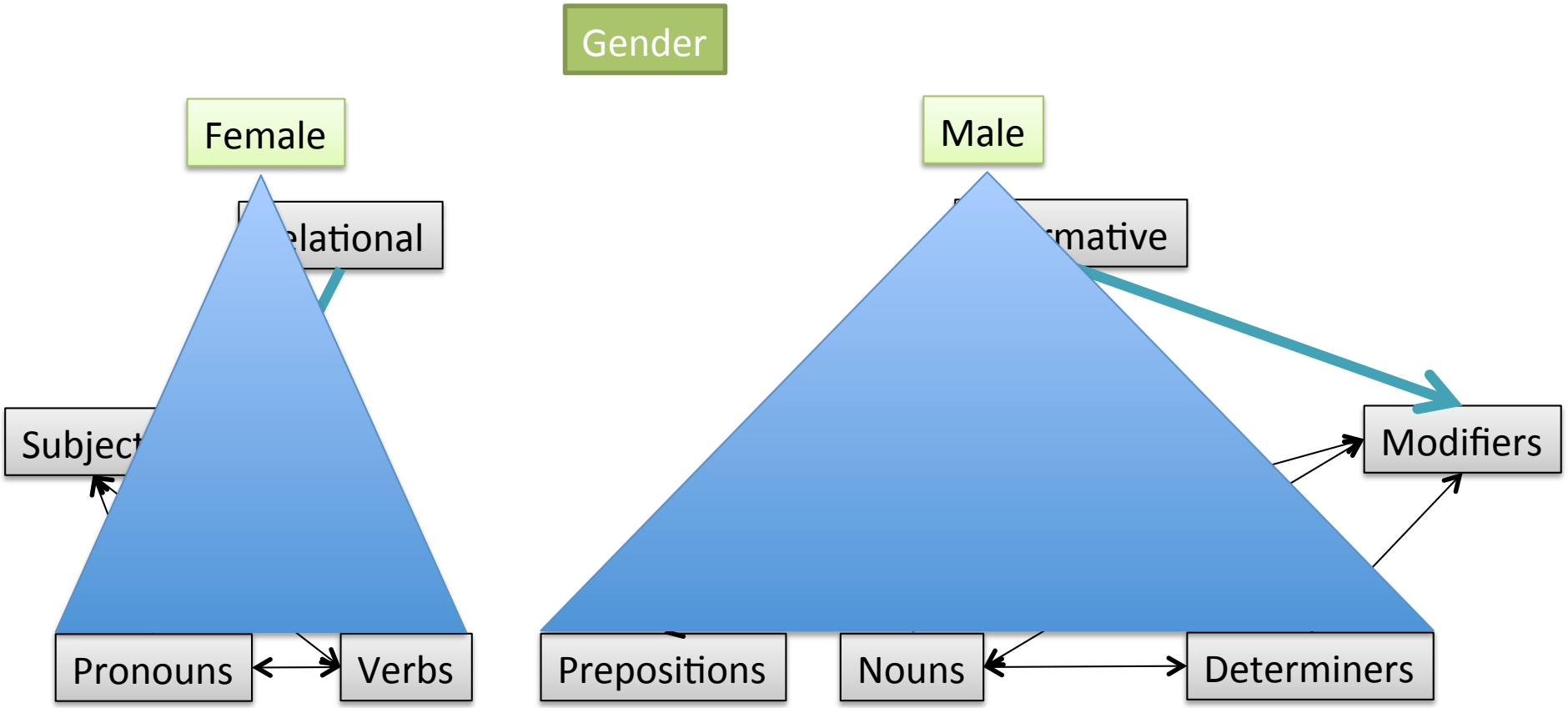
- Shlomo Argamon, Moshe Koppel et al. (from 2002)
- Documents: British National Corpus (fiction and non-fiction)
- Class: gender of author
- Feature construction:
  - lexical (Function Words)
  - POS (Function Words)
- Supervised learning: linear separator
- Results: gender ~ 80% predictable from text

Gender Differences

- Use of pronouns (more by women) and some types of noun modification (more by men)
  - *"Male" words: a, the, that, these, one, two, more, some*
  - *"Female" words: I, you, she, her, their, myself, yourself, herself*
- More "relational" language use by women and more "informative" (descriptive) language use by men
- Even in formal language use!
- Strong correlation between male language use and non-fiction, and female language use and fiction

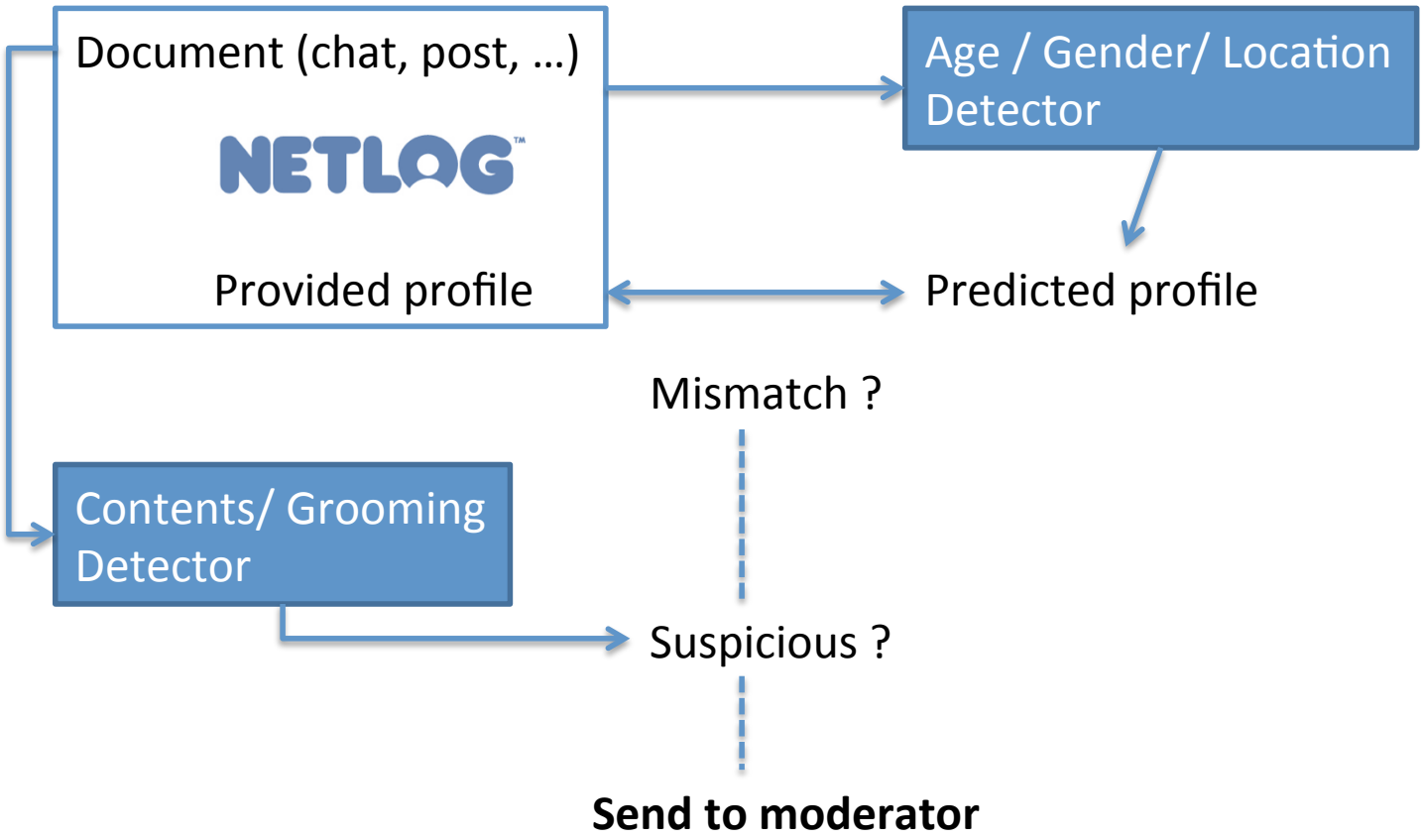
- LIWC categories (in Blogs):
  - Men talk more about jobs, money, sports, tv
  - Women talk more about sex, family, eating, friends, sleep, emotions

Explanation in Stylometry



DAPHNE project

Defending Against Pedophiles in Heterogeneous Network Environments (IOF PhD project Claudia Peersman)



Properties of chat language

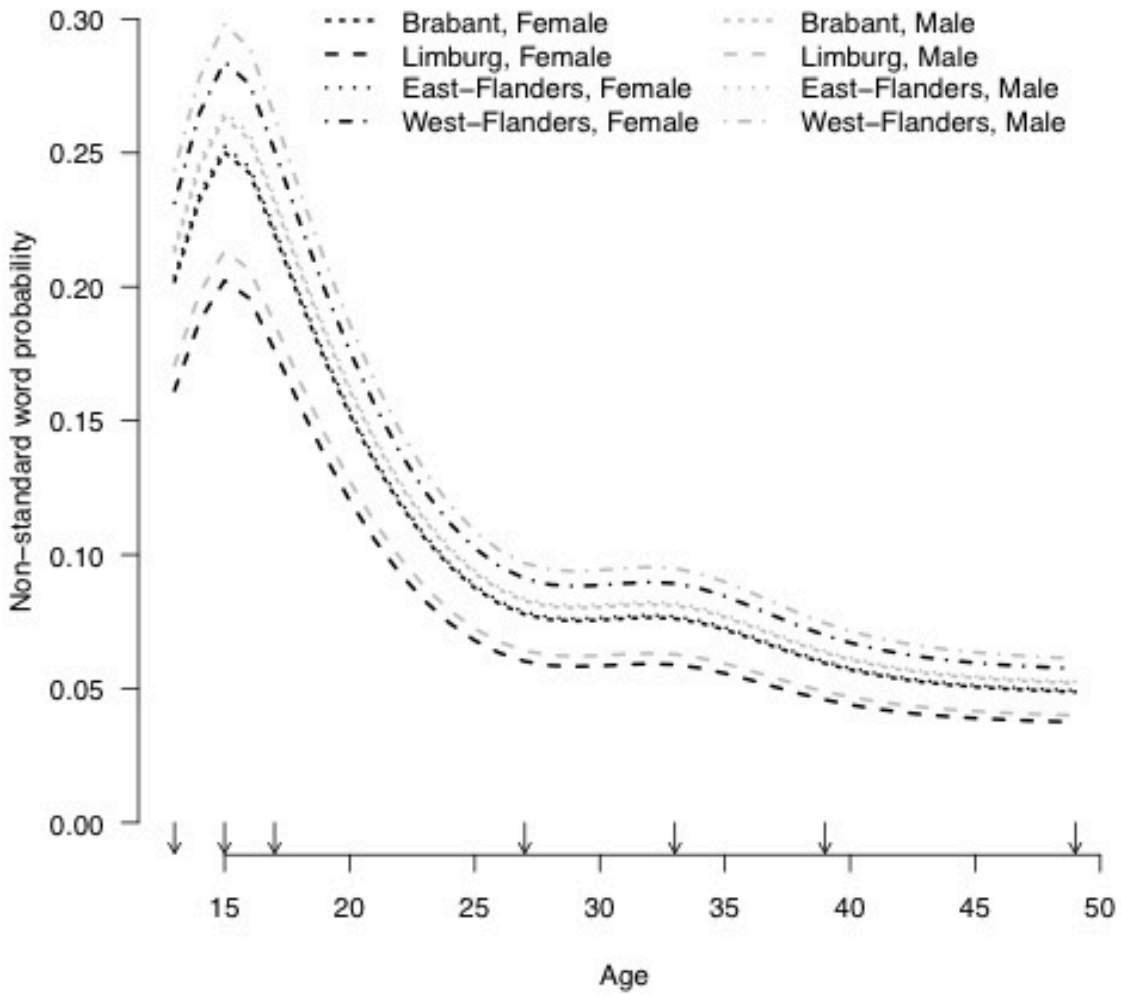
Variation type	Netlog example	Standard Dutch	English
Omission of letters or words	kbda nimr	Ik heb dat niet meer.	I don't have that anymore.
Abbreviations	wrm W8	waarom wacht	why wait
Acronyms	hjg	hou je goed	take care
Character flooding	keiii mooiii	heel mooi	very beautiful
Concatenation	IkKanOokNiiiZonderU!	Ik kan ook niet zonder jou!	I can't live without you either!

Maxims of (Dutch) chat language

- Write as fast as you can to ensure a fluent interaction
- Write the way you speak to ensure the informal character of the conversation
  - (Vandekerckhove, 2010)

- Huge problem for automatic text analysis (even POS tagging is impossible)
  - Normalization
  - Special purpose tools
  - But: blessing in disguise for accurate prediction!

Language Variation in Netlog chat



Current Results

- Mismatch detection
  - Age ~ 80%, Gender ~ 70%, Location ~ 50%
  - Bag of words performs best ☹
    - Different age groups and genders use different intensifiers, emoticons, words in general …
- Grooming detection
  - Predator detection ~ 90% f-score
    - Based on positive data from perverted justice website
  - Suspicious posts ~ 30% f-score

Suspicious Utterances

- Behavioral analysis
  - Different stages in online grooming (Lanning, 2010)
  - Analysis of positive training data
- Dictionary-based filter
  - Terminology related to:
    - Sexual topic, reframing, approaching, requesting data, isolating from supervision, age-related references

Detecting complex events

- Cyberbullying
  - Multiple categorization tasks
    - Insult, help, personality, emotion, …
  - Temporal aspects
    - Repetition, reactions, forwards
  - Roles (network)
    - Bully, victim, bystanders
  - Multi-modal
    - Photoshopped pictures, text, chat
- Solution: event (aka script) extraction
  - Ensemble of classifiers with decision function

Deep Learning

- Corpus-based Distributed Representations
  - (Recursive) Neural Networks, unsupervised, layered
  - Fast take-up by companies (Google)
- Impressive results on benchmark NLP tasks
- Lexical and sentence semantics
- Solves the problem of inference?

  http://www.thisplusthat.me

  Paris – France + Italy = Rome

  Sushi – Japan + Germany = Bratwurst

  bigger – big + cold = colder

Conclusions

- How to get back from TM to NLU?
- Biograph & AMiCA style Text Mining
  - Gradually deepen the representations (negation, modality, …) while keeping the data-oriented context
  - Gradually move to more complex knowledge structures (script extraction) while keeping the text categorization framework
- Or:
  - It's third time lucky for Neural Networks ?