Interactive Text Exploration

Günter Neumann,
DFKI, Saarbrücken, Germany
Joined work with Sven Schmeier, DFKI, Berlin.

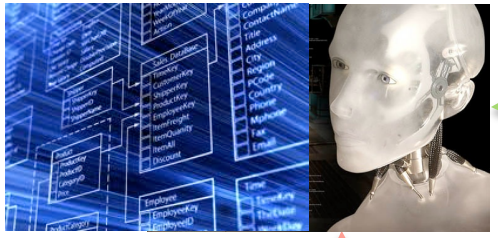
Overview of my talk

- Motivation and Background
- Interactive exploratory search
- Methods and technology
- Where we are, where we want to go

"The Big Idea"

- The **extraction, classification**, and **talking about information** from large-scale unstructured noisy multi-lingual text sources.

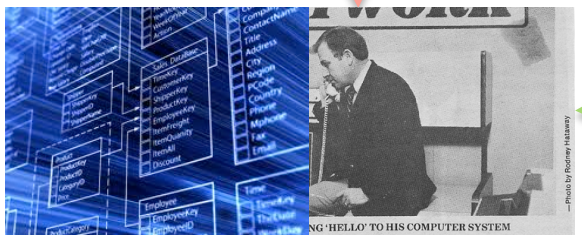
Topic of Interest



Private KB

Text as interface

Open NL-KB

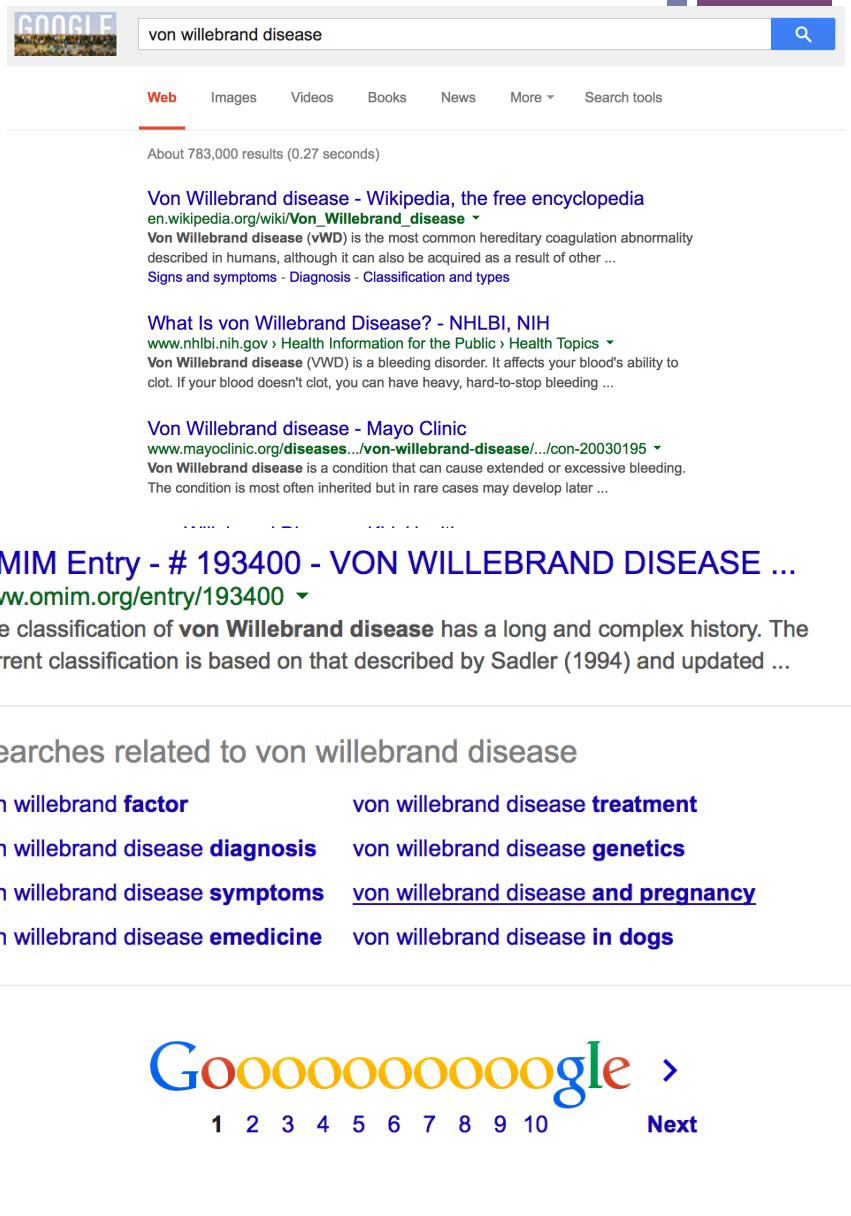


Private KB

„Reading text and talking about it"

Motivation

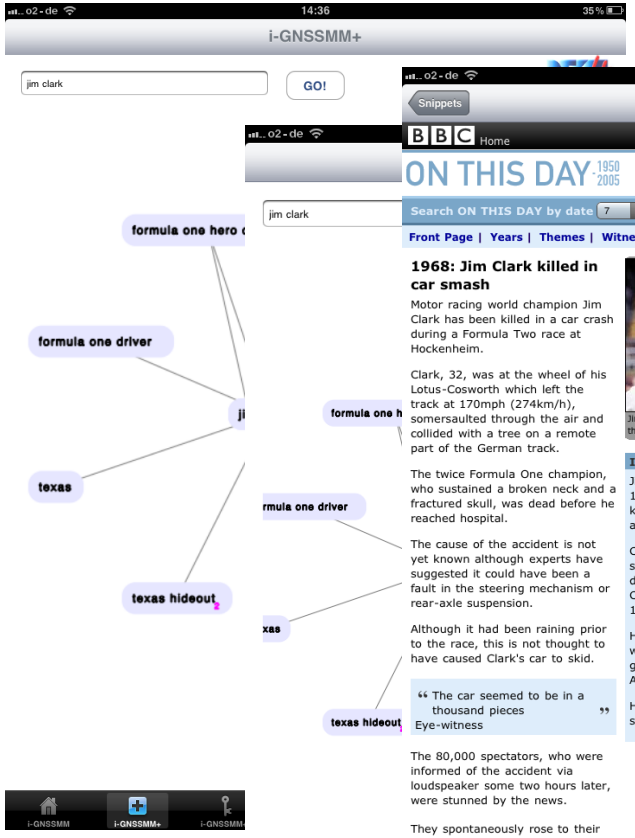
- Today's Web search is still dominated by one-shot-search:
  - Users basically have to know what they are looking for.
  - The documents serve as answers to user queries.
  - Each document in the ranked list is considered independently.
- Restricted assistance in content-oriented interaction

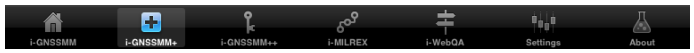


Exploratory Search

- We consider **a user query** as a specification of a **topic** that the user wants to know and learn more about. Hence, the **search result** is basically a **graphical structure of the topic and associated topics** that are found.
- The **user can interactively explore this topic graph** using a simple and intuitive (touchable) user interface in order to either learn more about the content of a topic or to interactively expand a topic with newly computed related topics.

Exploratory Search on Mobile Devices





Our Approach – On-demand Interactive Open Information Extraction

- Topic-driven Text Exploration
  - Search engines as API to text fragment extraction (snippets)
- Dynamic construction of topic graphs
  - Empirical distance-aware phrase collocation
  - Open relation extraction
- Interaction with topic graphs
  - Inspection of node content (snippets and documents)
  - Query expansion and eventually additional search
  - Guided exploratory search for handling topic ambiguity

Sear

von Willebrand

The von Willeb

Type 2 von Wil

Von Willebranc

Intracellular st

Acquired von V

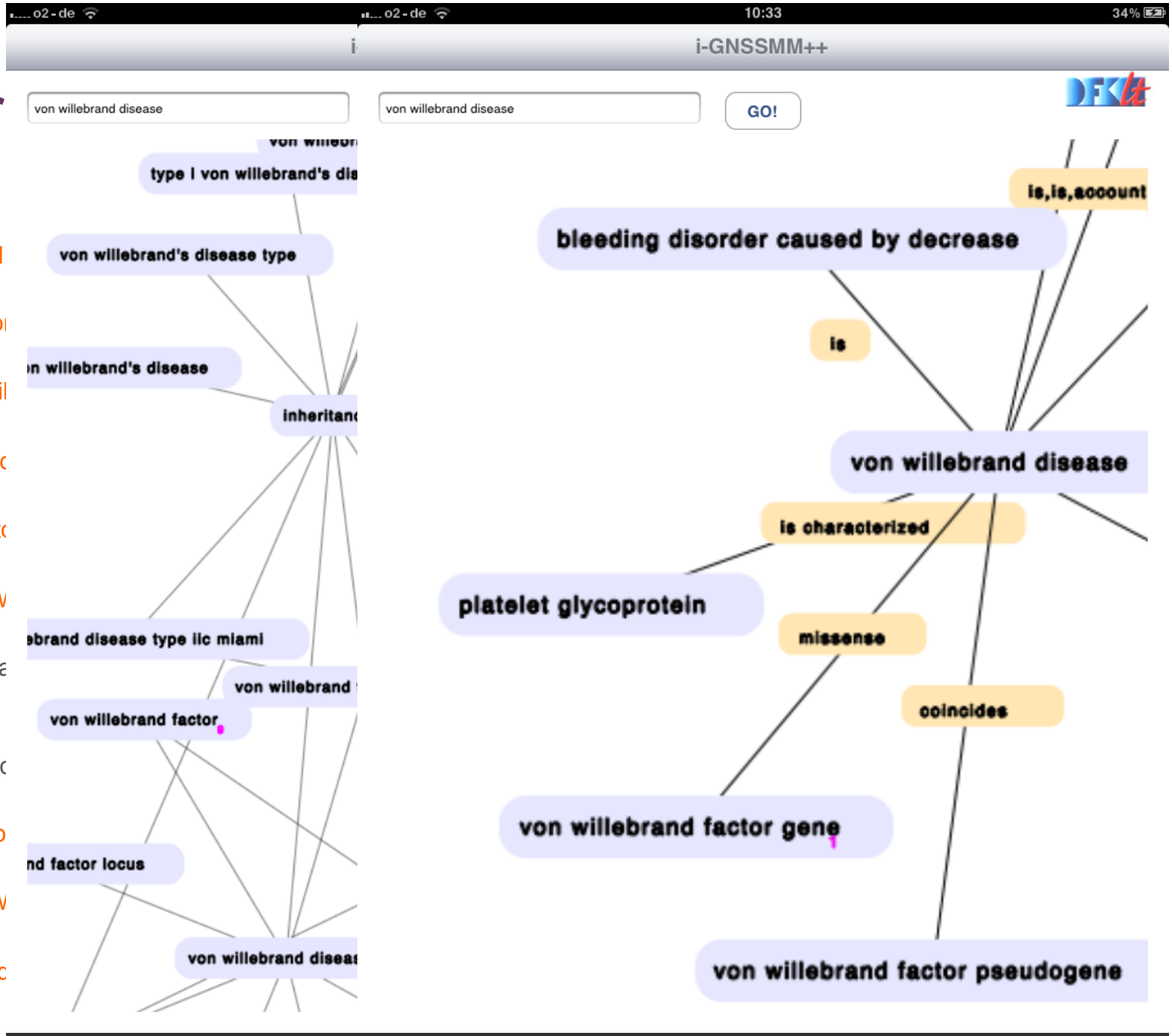
Porcine and ca
studies.

Pregnancy and

Investigation o

Multiple von W

Oligosaccharic



Topic Graphs

- Main data structure
  - A graphical summary of relevant text fragments in form of a graph
  - Nodes and edges are text fragments
    - Nodes: entities phrases
    - Edges: relation phrases
    - Content of a node: set of snippets it has been extracted from, and the documents retrievable via the snippets' web links.
- Properties
  - Open domain
  - Dynamic index structure
  - Weight-based filtering/construction

Construction of Topic graphs

- Identification of relevant text fragments
  - A document consisting of topic-query related text fragments
- Identification of nodes and edges
  - Distance-aware collocation
  - Clustering-based labels for filtering
- Technology
  - **Shallow** Open relation Extraction (ORE) for snippets
  - **Deeper** ORE for more regular text

Chunk-pair distance model

Topic pair weighting

Topic graph visualization

For each chunk c_i do:

$(c_i,c_{i+1},d_{i(i+1)})$, $(c_i,c_{i+2},d_{i(i+2)})$, ...

$(c_i , c_j , \#c_i , \#c_j , D_{ij})$ with

$D_{ij} = \{(freq_1,dist_1),(freq_2,dist_2),...\}$

$$PMI(cpd) = log_2((p(c_i,c_j)/(p(c_i) * p(c_j)))$$
$$= log_2(p(c_i,c_j)) - log_2(p(c_i) * p(c_j))$$

The visualized topic graph TG is then computed from a subset $CPD'_M \subset CPD_M$ using the m highest ranked cpd for fixed c_i. In other words, we restrict the complexity of a TG by restricting the number of edges connected to a node.

Evaluation of Mobile Touchable User Interface

- 20 testers
  - 7 from our lab
  - 13 "normal" people
- 10 topic queries
  - Definitions: EEUU, NLF
  - Person names: Bieber, David Beckham, Pete Best, Clark Kent, Wendy Carlos
  - General: Brisbane, Balancity, Adidas.
- Average answer time for a query: ~0.5 seconds

Table 1: Google

#Question	v.good	good	avg.	poor
results first sight	55%	40%	15%	-
query answered	71%	29%	-	-
interesting facts	33%	33%	33%	-
suprising facts	33%	-	-	66%
overall feeling	33%	50%	17%	4%

Table 2: i-GNSSMM

#Question	v.good	good	avg.	poor
results first sight	43%	38%	20%	-
query answered	65%	20%	15%	-
interesting facts	62%	24%	10%	4%
suprising facts	66%	15%	13%	6%
overall feeling	54%	28%	14%	4%

Guided Exploratory Search

- Problem: a topic graph might merge information from different topics/concepts
- Solution:
  - Guided exploratory search
  - Using an external KB (e.g., Wikipedia)
- Strategy
  - Compute topic graph TD_q for query q
  - Ask KB (Wikipedia or any other KB) if q is ambiguous
  - Let user select reading r, and use selected Wikipedia article for expanding q to q'
  - Compute new topic graph TD_q'


jim clark's 1967 italian grand prix
grand prix legends
jim clark rally photos
25 grand prix races
1
german grand prix
jim clark
business
business solution delivery
family
business data base
3
clark
5
employee jim clark
jim & chris clark
14:45
Lädt nicht
i-GNSSMM
GO!
jochen rindt
mike spence
1
im clark
graham hill
2
mike hawthorn
2
race
criminalRobert Big Bob Brady and
i-GNSSMM
i-GNSSMM+
i-GNSSMM++
i-GNSSMM
i-GNSSMM+
i-GNSSMM++
i-Wiki
i-WebQA
Settings
About

Evaluation

List of celebrity guest stars in Sesame Street:

209 different queries

List of film and television directors:

229 different queries

Evaluation

- Goal:
  - We want to analyze whether our approach helps building topic graphs which **express a preference** for the selected reading.
- Automatic evaluation:
  - Method
    - For each reading article r, compute topic graph TD_r using expanded query
    - Compare TD_r with all readings and check whether best reading equals r
  - Advantage: No manual checking necessary
  - Disadvantage: Correctness of TD_R needs to be proven
- Manual evaluation:
  - Double-check the results of the automatic evaluation
  - Prove the results at least for the examples used in evaluation

Results

set	#queries	good	bad	acc
Sesame + Colloc.	209	375	54	87.41 %
Sesame + Colloc.+ SemLabel	209	378	51	88.11 %
Hollywood + Colloc.+ SemLabel	229	472	28	94.40 %
Hollywood + Colloc.+ SemLabel	229	481	19	96.20 %

	1st task	2nd task			
set	guidance	associated topics	good	bad	accuracy
Sesame	ca. 95 %	167	132	35	79.04 %
Hollywood	ca. 95 %	145	129	16	89.00 %
Sesame	> 97 %	167	108	59	64.67 %
Hollywood	> 97 %	145	105	40	72.41 %

Automatic

- Colloc. – empirical collocations for topic graph computation
- SemLabel – Filtering of nodes using semantic labels computed via SVD (Carrot2)

Manual

- 2 test persons
- 20 randomly chosen celebrities and 20 randomly chosen directors
- 1st task: Exploratory search and personal judgments of the Guidance by the system
- 2nd task: Check all associated nodes after choosing a meaning in the list

Summary and Discussion

- Interactive topic graph exploration
  - Unsupervised open information extraction
  - On-demand computation of topic graphs
  - Strategies for guided exploratory search
  - Effective for Web snippet like text fragments
  - Implemented for EN and DE on mobile touchable device
- Drawback
  - Problems in processing text fragments from large-scale text directly
  - Especially Open Relation Extraction for German is challenging
- Solution:
  - Nemex - A new multilingual Open Relation Extraction approach

Nemex – A Multilingual Open Relation Extraction Approach

- Uniform multilingual core ORE
  - N-ary extraction
  - Clause-level
- Multi-lingual
  - Very few language-specific constraints over dependency trees
  - Current: English and German
- Efficiency
  - Complete pipeline (form sentence splitting, to POS-tagging, to NER, to dependency parsing, to relation extraction)
  - About 800 sentences/sec
  - Streaming based – small memory footprint

German ORE is Challenging

- Challenging properties of German
  - Morphology/Compounding*
  - No strict word ordering (especially between phrases)
  - Discontinuous elements, e.g., verb groups
- Simple, pattern-based ORE approach difficult to realize (
- Deep sentence analysis helpful
  - Current multilingual dependency parsers provide very good robustness!
  - DFKI's MDParser is very efficient: 1000sentences/second (but see also Chen&Manning, 2014)
- Challenge:
  - Can we design a core uniform ORE approach for English, German, … ?

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
"the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef"

Multilingual ORE – Our Approach

- Multi-lingual open relation extraction
  - Only few Language-specific constraints necessary (constraints over direct dependency relations (head, label, modifier))
  - Few language-independent constraints in case of uniform dependency annotations, e.g., McDonald et al., 2013
- Processing strategy
  - Head-Driven Phrase Extraction
  - Top-down head-driven traversal of dependency tree

Example: English

Mammalian NMD was mostly studied in cultured cells so far and there was no direct evidence yet that NMD could operate in the brain .

Dependency Tree (uniform tag and label set; Conll format):

1:Mammalian:NOUN:compmod:2
2:NMD:NOUN:nsubjpass:5
3:was:VERB:auxpass:5
4:mostly:ADV:advmod:5
5:studied:VERB:ROOT:0
6:in:ADP:adpmod:5
7:cultured:ADJ:amod:8
8:cells:NOUN:adpobj:6
9:so:ADV:advmod:10
10:far:ADV:advmod:5
11:and:CONJ:cc:5
12:there:DET:expl:13
13:was:VERB:conj:5
14:no:DET:det:16
15:direct:ADJ:amod:16
16:evidence:NOUN:nsubj:13
17:yet:ADV:advmod:13
18:that:ADP:mark:21
19:NMD:NOUN:nsubj:21
20:could:VERB:aux:21
21:operate:VERB:advcl:13
22:in:ADP:adpmod:21
23:the:DET:det:24
24:brain:NOUN:adpobj:22
25:.:.:p:5

Example English – cont.

*

(Mammalian NMD, was mostly studied so far, in cultured cells)
(no direct evidence, was yet, there)
(NMD, could operate, in the brain)

**Annotated sentence:
[[[Arg11 Mammalian NMD Arg11]]] --->Rel1 was mostly studied [[[Arg13 in cultured cells Arg13]]] so far Rel1<--- and [[[Arg23 there Arg23]]] --->Rel2 was [[[Arg21 no direct evidence Arg21]]] yet Rel2<--- that [[[Arg31 NMD Arg31]]] --->Rel3 could operate Rel3<--- [[[Arg33 in the brain Arg33]]] .

*Details omitted
**Extension of the annotation scheme introduced by Mesquita et al., 2013

Example: German

Zuvor hatte Asmussen mitgeteilt, dass er sein Amt als EZB-Direktor in Kürze aufgeben will:

**Earlier had Asmussen informed, that he his position as EZB-director in the_near_future quit will:*

Earlier Asmussen has informed that he will quit his position as EZB-director in the_near_future:

Dependency Tree (uniform tag and label set; Conll format):

1:Zuvor:ADV:advmod:2
2:**hatte**:VERB:ROOT:0
3:Asmussen:NOUN:nsubj:2
4:**mitgeteilt**:VERB:aux:2
5:,:.:p:2
6:dass:CONJ:mark:14
7:**er**:PRON:nsubj:14
8:sein:PRON:poss:9
9:Amt:NOUN:dobj:14
10:als:ADP:adpmod:14
11:EZB-Direktor:NOUN:adpobj:10
12:in:ADP:adpmod:14
13:Kürze:NOUN:adpobj:12
14:**aufgeben**:VERB:NMOD:2
15:**will**:VERB:aux:14
16:::.:NMOD:2

Example German – Cont.

(Asmussen, Zuvor hatte mitgeteilt)
(er, aufgeben will, sein Amt, als EZB-Direktor, in Kürze)

Annotation:
--->Rel1 Zuvor hatte [[[Arg11 Asmussen Arg11]]] mitgeteilt Rel1<--- , dass [[[Arg21 er Arg21]]] [[[Arg22 sein Amt Arg22]]] [[[Arg23 als EZB-Direktor Arg23]]] [[[Arg24 in Kürze Arg24]]] --->Rel2 aufgeben will Rel2<--- :

Nemex – Current Status

- Properties
  - Efficient text stream for EN and DE implemented
  - Uniform POS and Dependency labels
  - Small set of uniform constraints over dependency relations
- Very fast & Domain independent
  - About 800 sentences per second for complete pipeline
- Current /near future work
  - Improve cross-clausal resolution
  - Extensive evaluation, intrinsic and extrinsic
  - Adaptation to other languages
    - Conll based dependency treebanks (uniform and specific)

Future action points

- Cross-sentence open information extraction
  - **Goal: co-reference resolution, integration of more fine-grained information to dependency parsers (morphology), text inference**
- Beyond isolated topic graphs
  - **Goal: share topic graphs, compare topic graphs, monitor topic graphs**
- Interactive text data mining and knowledge discovery
  - **Goal: support abstract interactions, e.g., "more like this", "less like this", "what is this", …**

DONE

Thank you for Your Attention !