Question Answering and the development of the *Hajen* System

Pierre Nugues

Joint work with Marcus Klang, Rebecka Weegar, Peter Exner, Juri Pyykkö

Background: IBM Watson

- IBM Watson: A system that can answer questions better than any human
- Video: https://www.youtube.com/watch?v=WFR3lOm_xhE

- IBM Watson builds on the extraction of knowledge from masses of texts: Wikipedia, archive of the New York Times, etc.
- Bottom line: Text is the repository of human knowledge

LUND UNIVERSITY

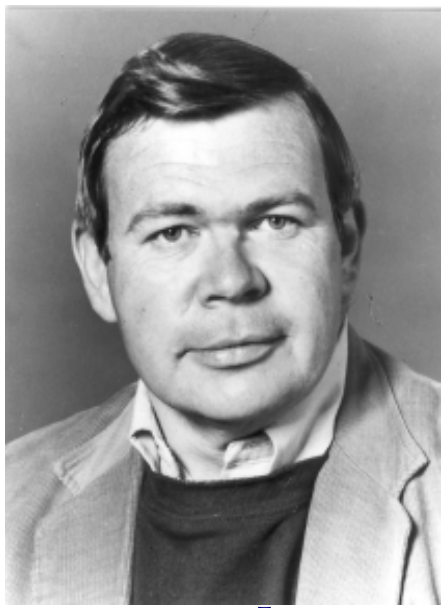
Goals of *Hajen*

- Build a system that can answer questions in a way similar to IBM Watson
- Use semantic knowledge extracted from Swedish sources: Wikipedia, encyclopedias, *Sydsvenskan,* others?
- Beyond Swedish:
  - Implement a generic, open-source, modular, question answering platform
  - Make it easy to integrate multilingual components
  - Use of full-text indexing, graph-based document models, map-reduce Hadoop ecosystem, dependency injection techniques

LUND UNIVERSITY

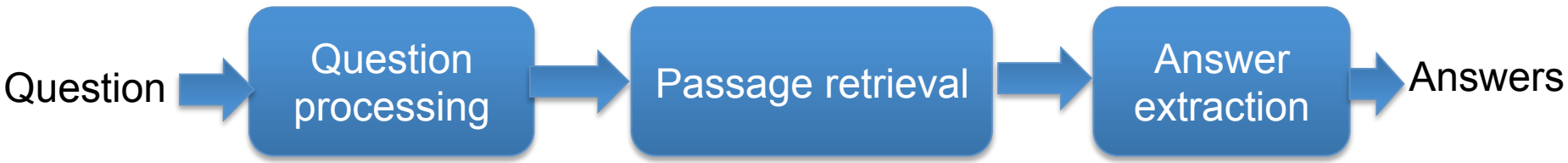
Hajen?

- Part of Vetenskapsrådets frame funding **Det digitaliserade samhället**
- Semantic processing project: *Mot kunskapsbaserad storskalig kunskapsutvinning ur svenskt text*
- Cooperation with Gothenburg University and Chalmers
- The name is a tribute to Ulf Hannerz, the legendary winner of *Kvitt eller dubbelt* in 1957.
- http://www.svtplay.se/klipp/297574/10-000-kronorsfragan-kvitt-eller-dubbelt



LUND UNIVERSITY

Question-Answering Architecture (simplified from IBM Watson)



Question parsing and classification: *Syntactic parsing, entity recognition, answer classification*

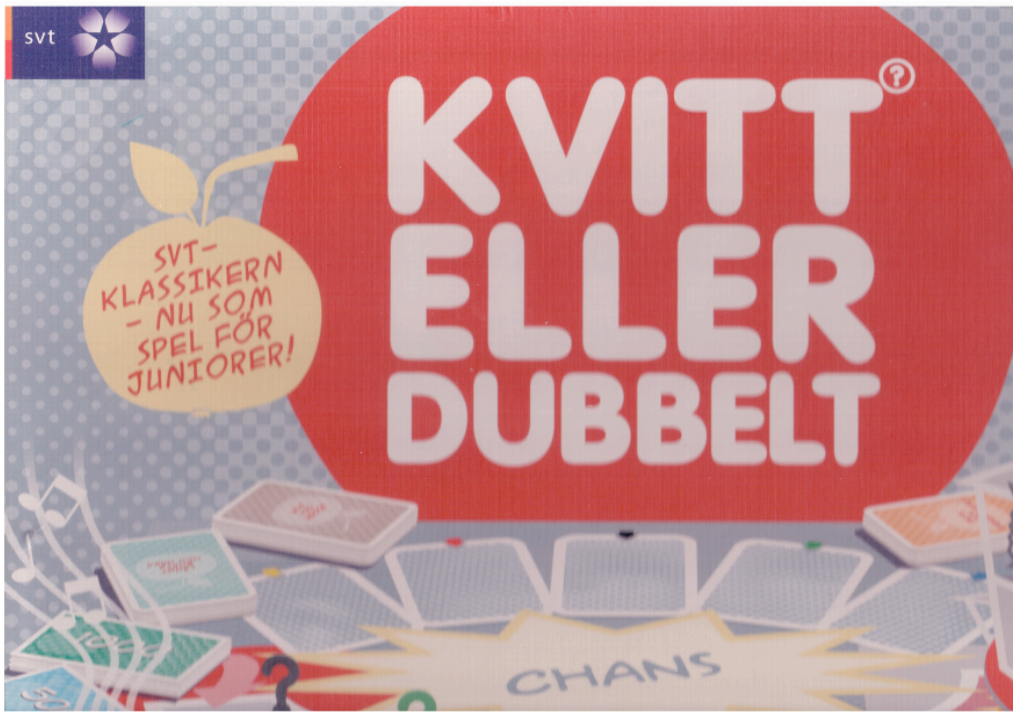
Document retrieval. Extraction and ranking of passages: *Indexing, vector space model.*

Extraction and ranking of answers: Answer parsing, entity recognition

LUND UNIVERSITY

Corpus: *Kvitt eller dubbelt (KED)*

- In language processing, everything starts with a corpus
- *SVT-klassikern*, a simplified *Jeopardy!* in Sweden



LUND UNIVERSITY

Kvitt eller dubbelt: The Questions

LUND UNIVERSITY

Representing the Questions

- We transcribed the 2300 questions and we represented them as an RDF graph.
- A question from the category: “Får jag lov”

  Q: Vad heter den argentinska dans som Petra Nielsen dansade i melodifestivalen 2004? **A: Tango**

- Represented as:

  [kvitt:line 1; kvitt:value 250; kvitt:text "Vad heter den argentinska dans som Petra Nielsen dansade i melodifestivalen 2004?"; kvitt:answer "Tango”]

- Using SPARQL, we can extract easily data from the graph and carry out tests.

LUND UNIVERSITY

Passage Retrieval: Textual Resources

- Question answering needs a knowledge source in the form of a collection of documents.
- Wikipedia has a large and growing coverage of topics.
- It is easy to download from dumps.wikimedia.org.
- Is the Wikipedia suitable to answer *Kvitt eller dubbelt* questions?

LUND UNIVERSITY

Passage Retrieval: Indexing

- We segmented the Wikipedia articles into paragraphs: the passages.
- We indexed the passages using Apache Lucene
- Given a question, like:

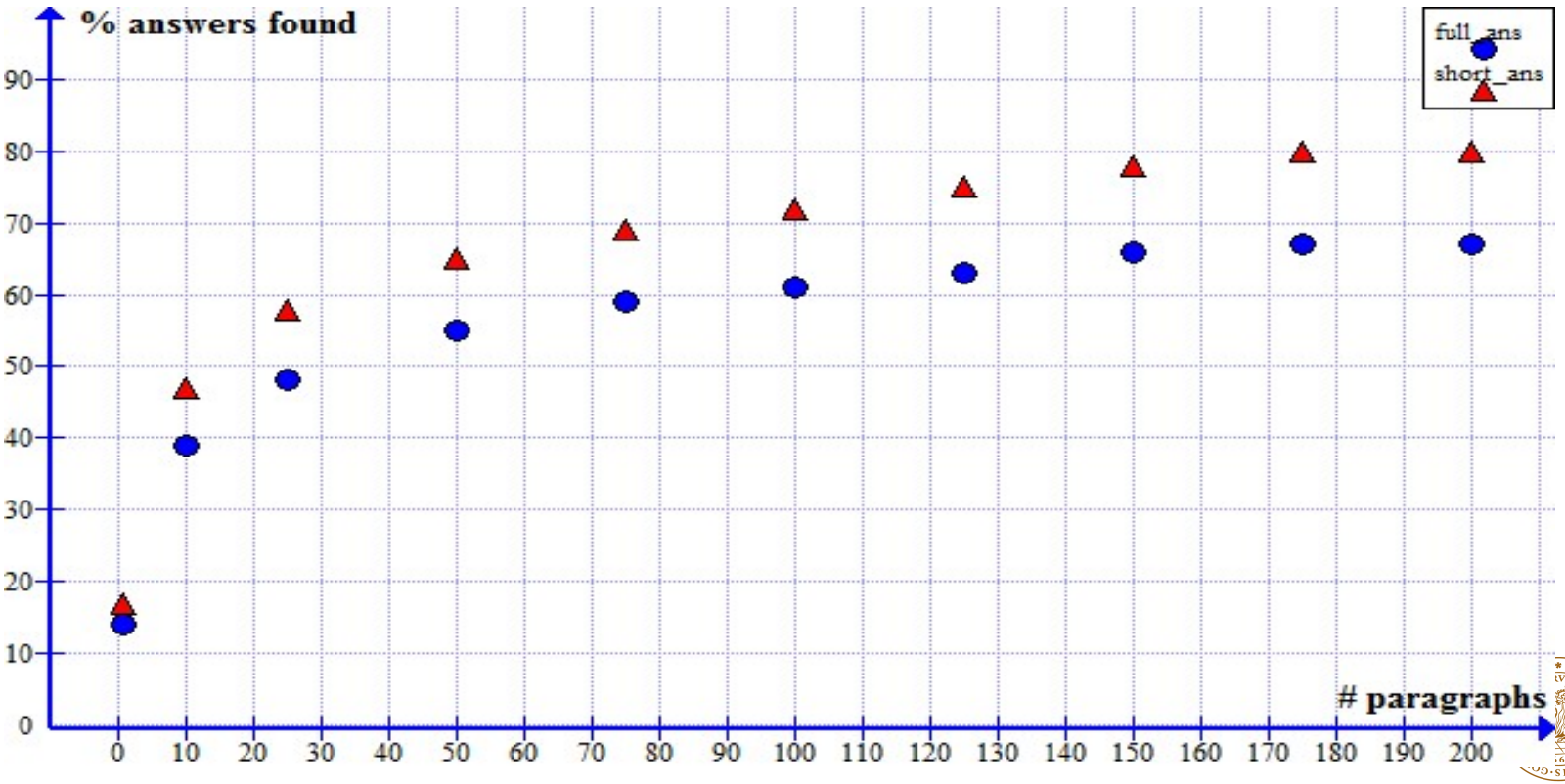
  Vad heter den argentinska dans som Petra Nielsen dansade i melodifestivalen 2004?

  We return the most relevant paragraphs scored using the Lucene's built-in TF.IDF measure.
- [Demonstration: http://semantica.cs.lth.se:8888/sv/query/passage]
- How good is it?

LUND UNIVERSITY

Passage Retrieval: Some Results

We submitted questions from the corpus and we checked if the returned paragraphs contained the answer string.



Joint work with Juri Pyykkö and Rebecka Weegar

LUND UNIVERSITY

Answer Extraction

- Most KED answers are concepts or entities (real or imaginary unique things).
- Corresponds to common nouns or proper nouns in the passages.
- Extractible using a part-of-speech tagger.
- Proper nouns can be classified into categories such as persons, cities, countries, organizations, etc.
- Once extracted, candidates are ranked by increasing frequency.
- This is our baseline answer extraction

LUND UNIVERSITY

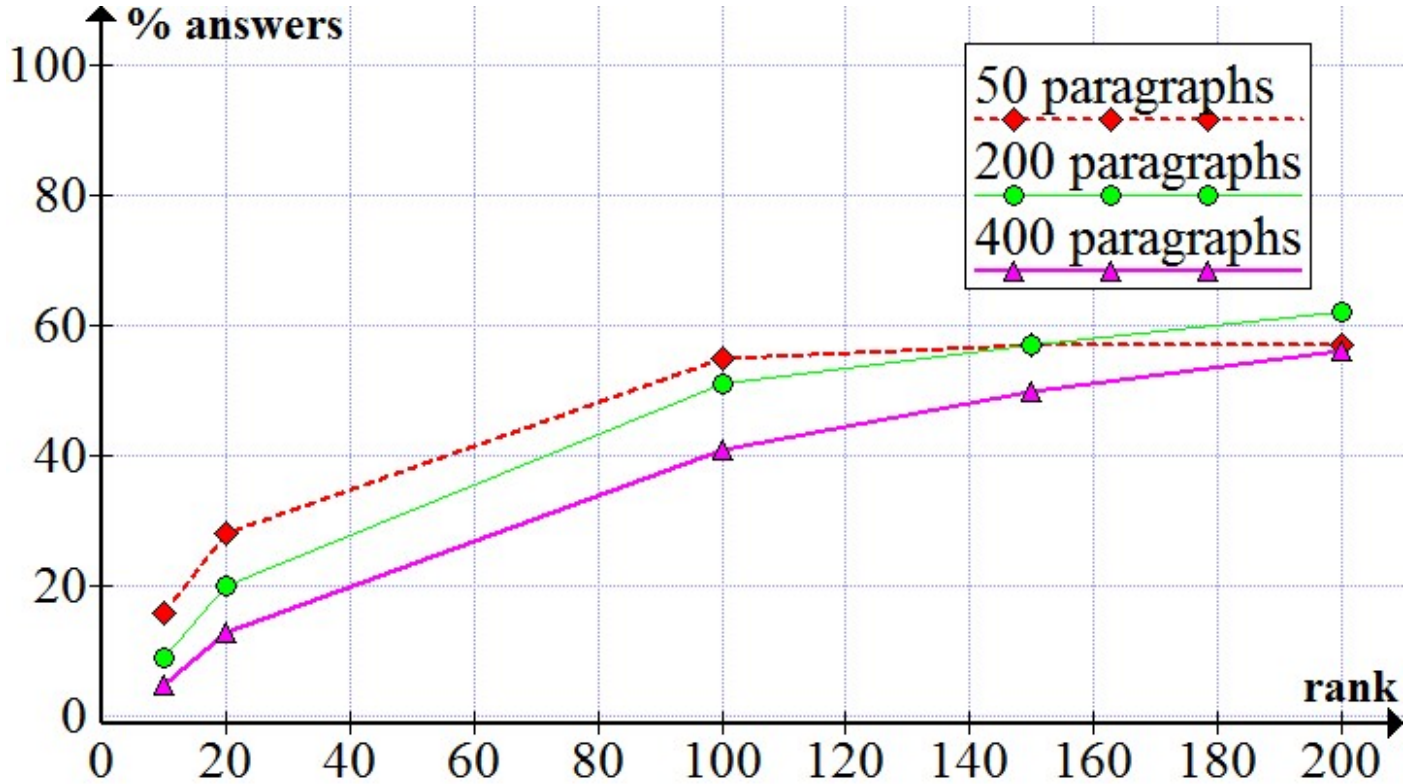
Part-of-Speech Tagging

1	Petra	Petra	PM	(person
2	Magdalena	Magdalena	PM	person
3	Nielsen	Nielsen	PM	person)
4	,	,	MID	_
5	ursprungligen	ursprungligen	AB	_
6	Inczèdy-Gombos	inczèdy-gombos	PM	_
7	,	,	MID	_
8	född	föda	PC	_
9		1	1RG	_
10	februari	februari	NN	_
11	1965	1965	RG	_
12	i	i	PP	_
13	Stockholm	Stockholm	PM	(place)
14	.	.	MAD	_

Part-of-Speech Tagging

1	Petra	Petra	PM	(person
2	Magdalena	Magdalena	PM	person
3	Nielsen	Nielsen	PM	person)
4	,	,	MID	_
5	ursprungligen	ursprungligen	AB	_
6	Inczèdy-Gombos	inczèdy-gombos	PM	_
7	,	,	MID	_
8	född	föda	PC	_
9		1	1RG	_
10	februari	februari	NN	_
11	1965	1965	RG	_
12	i	i	PP	_
13	Stockholm	Stockholm	PM	(place)
14	.	.	MAD	_

Answer Extraction: Preliminary Results

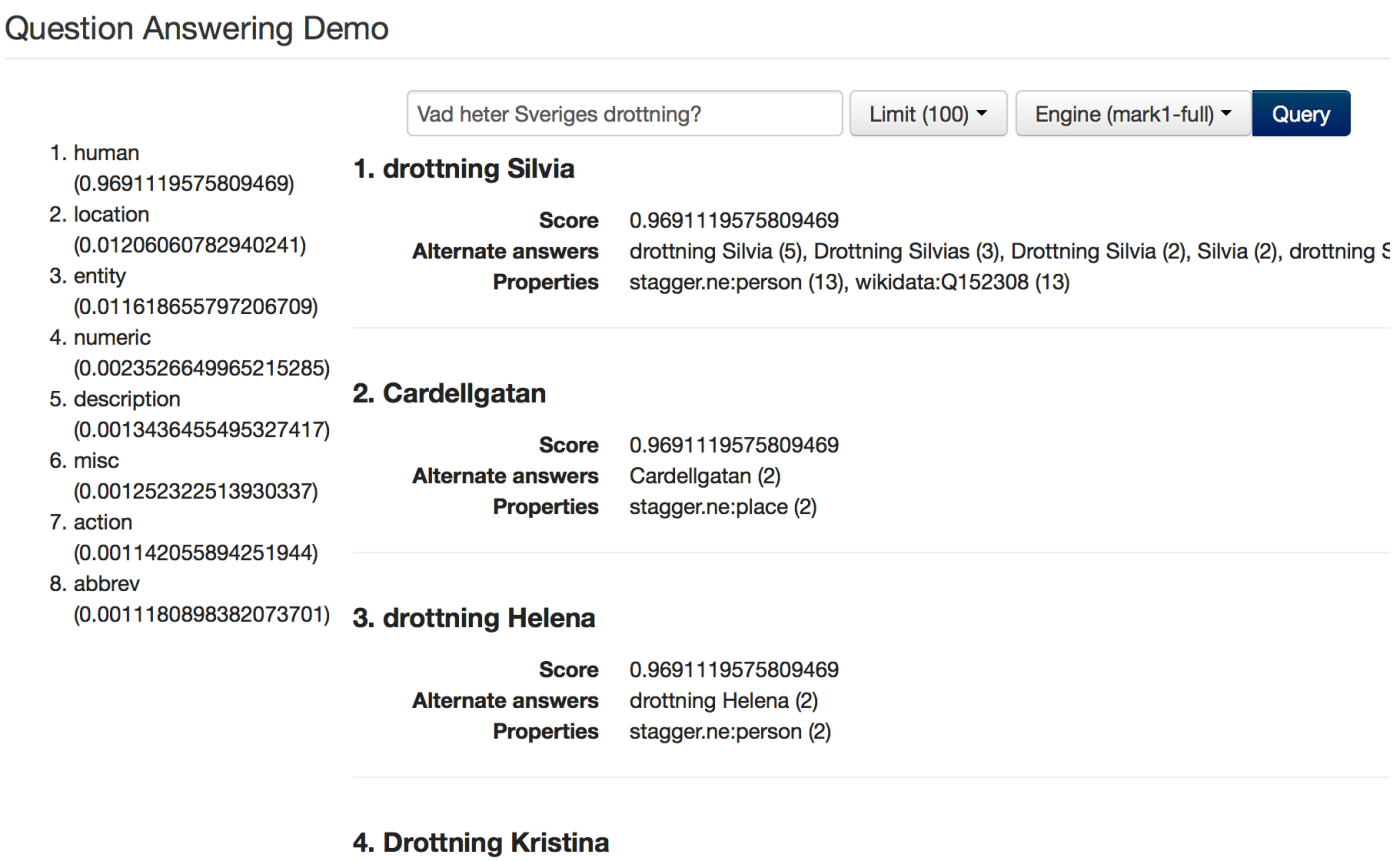


Results limited to answers classified as entities.

Joint work with Juri Pyykkö and Rebecka Weegar.

LUND UNIVERSITY

The Mark-1 Question-Answering Engine



Demonstration: http://semantica.cs.lth.se:8888/sv/query/qa

LUND UNIVERSITY

Answer Type Classification

- From the question text, we can often guess the type of the answer and discard unlikely candidates.
- *Q: Vilket är det enda nordiska land man inte kan köra bil från Sverige?* **A: Island**
- We annotated a part of the answers from KED and we trained an answer type classifier.
- We use logistic regression and we get probabilities for the answer types.
- Not completely integrated yet.

1. location (0.9751840864733327)
2. entity (0.020735020889680874)
3. numeric (0.0015616698708187907)
4. human (6.548017611252503E-4)
5. misc (5.00469009714687E-4)
6. action (5.004389990210438E-4)
7. abbrev (4.964844691307781E-4)
8. description (3.670285271759621 6E-4)

Work started with Christopher Käck and Robin Leussier

LUND UNIVERSITY

Reranking

- A reranker is a module that takes a list of ranked candidates and reorder them using “global” features such the answer type.

Q: Vilket är det enda nordiska land man inte kan köra bil från Sverige?

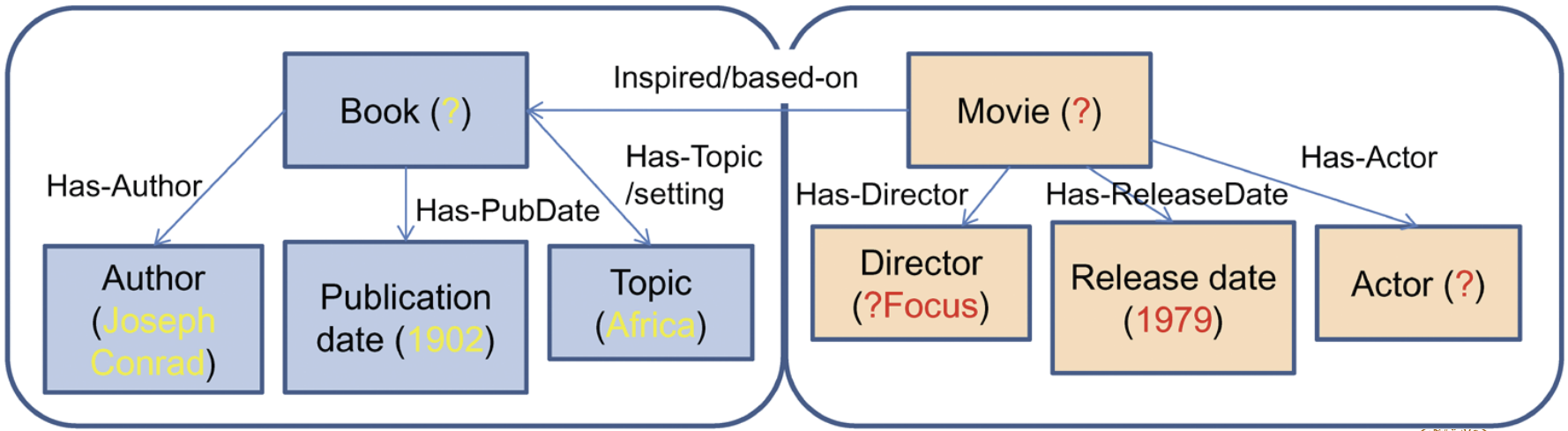
Rank	Answer	F/T
1	Europa	false
2	Patrik Budda Andersson	false
3	Sannas	false
…		
67	Island	true

- Using the corpus, it is possible to train a binary classifier with features such as the predicted answer type, word type
- Improves considerably the results (Not yet in the Mark-X line)

Work in cooperation with Jakob Grundström

Structured Data and Inference (From IBM Watson)

WAR MOVIES: ***A 1902 Joseph Conrad work set in Africa inspired this director to create a controversial 1979 war film.***



LUND UNIVERSITY

Named Entity Processing

- A first step to inference is to identify and disambiguate named entities.
- Two techniques: named entity recognition (NER) and named entity disambiguation (NED)

  Göran Persson *var statsminister mellan åren 1996 och 2006 samt var partiledare för* ***Socialdemokraterna***

- Google's motto *Things not Strings*

Named Entity Recognition

- Named entity recognition (NER) tags the words that correspond to named entities, possibly with their type.
- Common types include person, organization and location.
  - **Göran Persson** is a person.
  - **Socialdemokraterna** is an organization.
- Techniques to carry out named entity recognition use extensions of part-of-speech tagging

Named Entities are Ambiguous

Göran Persson kan syfta på

- Göran Persson (född 1949), socialdemokratisk partiledare och svensk statsminister 1996–2006
- Göran Persson (född 1960), socialdemokratisk politiker från Skåne
- Göran Persson (musiker), svensk proggmusiker
- Jöran Persson, svensk ämbetsman på 1500-talet

Named Entity Disambiguation

Named entity disambiguation (NED) links words to entities: *strings* to *things.*

Göran Persson var statsminister mellan åren 1996 och 2006 samt var partiledare för **Socialdemokraterna**.

LUND UNIVERSITY

Wikipedia and Wikidata

- Each page in Wikipedia (concepts and entities) has a unique identification number across all the languages, a sort of *personnummer*.
- The Q-number repository (wikidata) is a growing source of structured data.
  - Göran Persson (statsminister): Q53747
  - Göran Persson (riksdagsledamot): Q5626648

Wikidata Content

In other languages

français	Göran Persson
	homme politique suédois
svenska	Göran Persson
	Sveriges statsminister 1996-2006, Sveriges finansminister 1994-1996 samt Sveriges skolminister 1986-1991
suomi	Göran Persson
	enter a description in suomi

Statements

sex or gender	male
	3 references
LCNAF identifier	n98083850
	1 reference

In other languages

français	enter a label in français
	enter a description in français
svenska	Göran Persson (född 1960)
	enter a description in svenska
suomi	enter a label in suomi
	enter a description in suomi

Statements

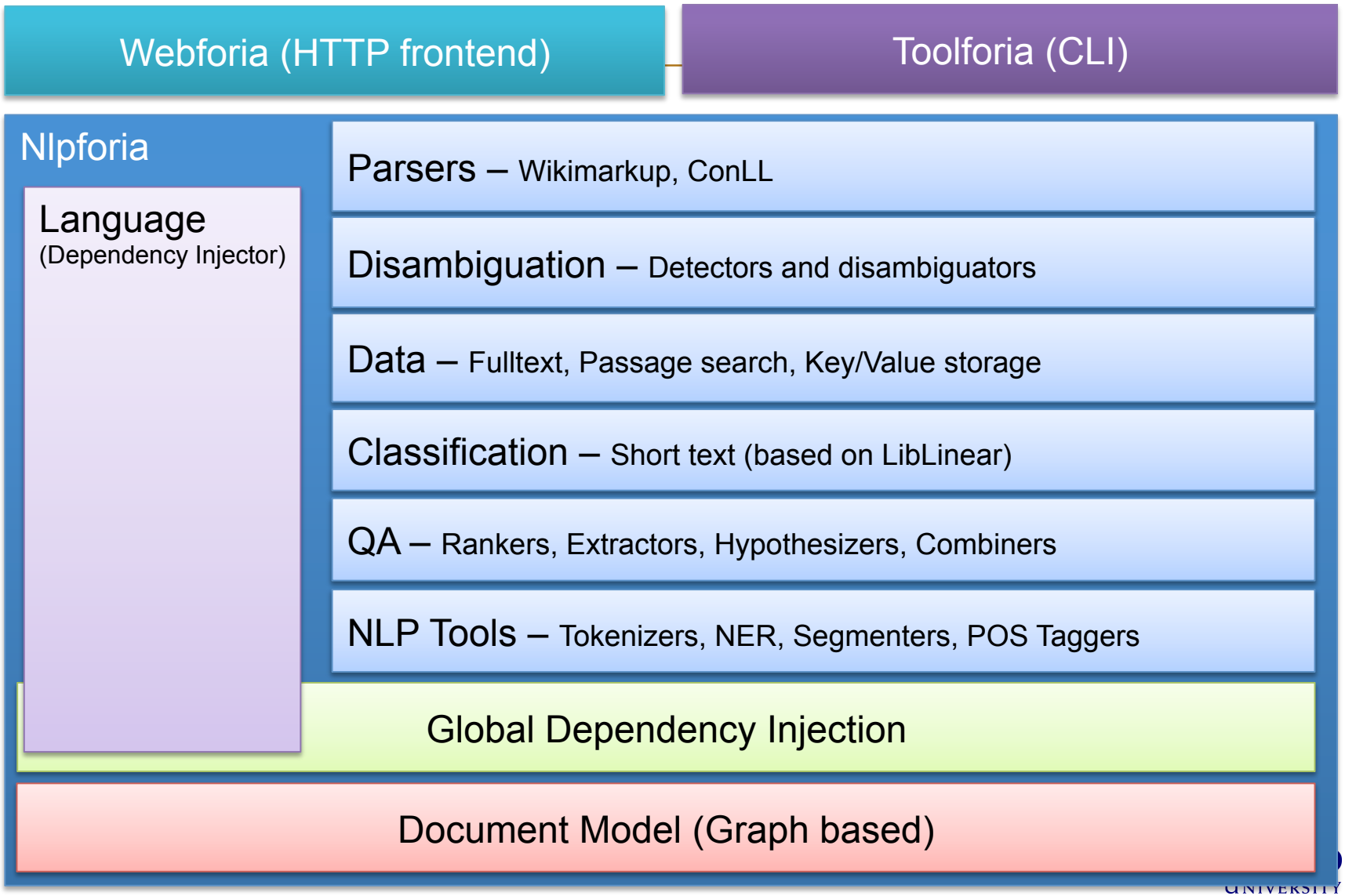
sex or gender	male
	1 reference
instance of	human
	1 reference

NEDforia: A Disambiguation Tool

Demonstration: http://semantica.cs.lth.se:8888/sv/ned/query

LUND UNIVERSITY

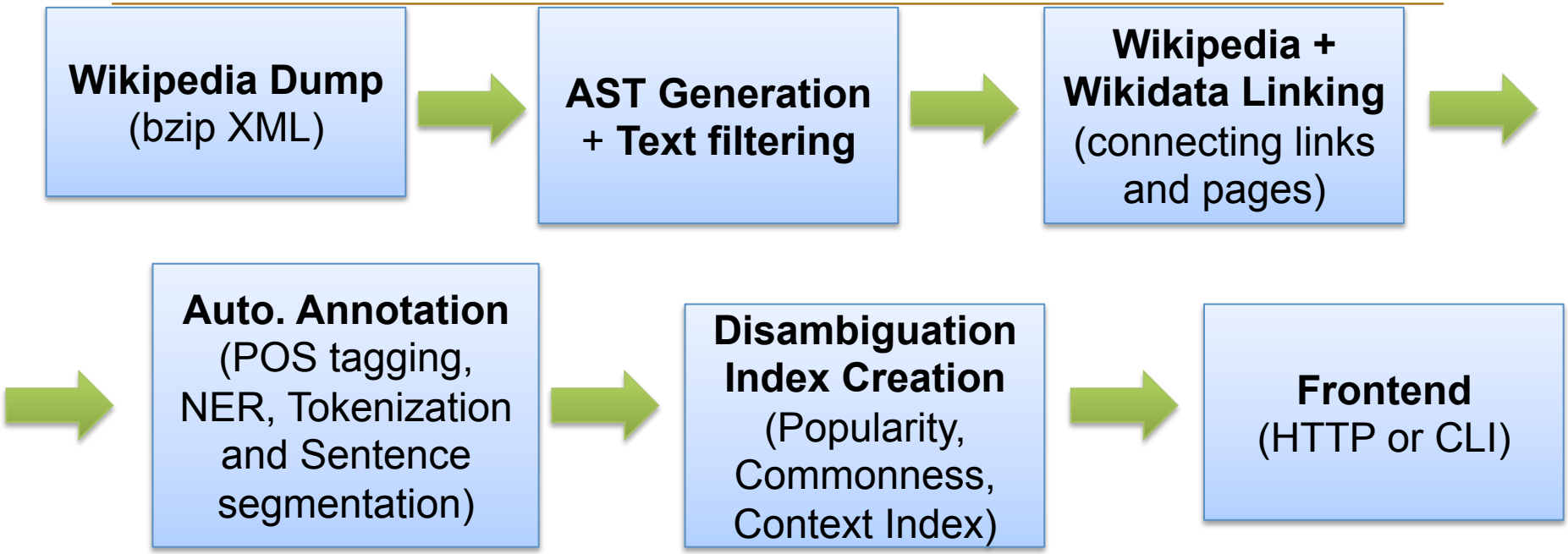
Mark 1 Architecture



Mark 1: Features

- **Document model**
  - Directed property graph
  - Built-in small query engine
  - Extends or re-uses ideas from KOSHIK (from Peter Exner)
- **Multilingual and modular**
  - **Idea:** A Language is nothing more than an assemblage of language specific components with models.
  - Constructor based dependency injection (Guice)
- **Designed for prototyping and experimentation**
- **Open source**

NEDforia: The NED Pipeline



Auto annotation was based on Stagger and Maltparser with respective models trained on the Stockholm-Umeå Corpus (SUC)

LUND UNIVERSITY

NEDforia: Disambiguation Results

- Small test corpus consisting of 10 news articles from mixed sources:
  Aftonbladet, Svd, DN, Nyteknik, HD, IDG and Dagens Industri
- We extracted and annotated 201 named entities, preliminary results.

Recall	Precision	F1
0.696	0.748	0.721

Cross-Language Extraction

English frame		
SBJ	Einstein	A0
VERB	received	receive.01
NMOD	the	
NMOD	1921	
OBJ	Nobel Prize	A1
TMP	in	
PMOD	Physics	

Swedish frame		
SS	Einstein	A0
VERB	fick	få.01
DT	1921	
DT	års	
OO	Nobelpris	A1
ET	i	
HD	fysik	

Joint work with Peter Exner

LUND UNIVERSITY

Koshik

- Koshik: End-to-end framework to process multilingual documents
- Based on Hadoop and our Crafoord cluster
- Koshik is a talking elephant (Korean)

LUND UNIVERSITY

Koshik Architecture

- Koshik uses the Avro binary format to serialize the documents.
- Avro is designed for Hadoop and allows other data warehousing tools to directly query the documents.
- Koshik can be queried directly through Hive, which offers an SQL-like query language called HiveQL.
- Available at https://github.com/peterexner/KOSHIK

Querying with HIVE

Number of articles:
> SELECT count(identifier) from koshikdocs;
Job 0: Map: 920 Reduce: 1   Cumulative CPU: 35243.42 sec   HDFS Read: 248123756147
HDFS Write: 8 SUCCESS
OK
4012291
Number of tokens:
> SELECT count(ann) FROM koshikdocs LATERAL VIEW explode(annotations.layer)
annTable as ann WHERE ann LIKE '%Token';
Job 0: Map: 920 Reduce: 1   Cumulative CPU: 35199.98 sec   HDFS Read: 248123756147
HDFS Write: 11 SUCCESS
OK
1485951256
Number of nouns:
> SELECT count(key) FROM (SELECT explode(ann) AS (key,value) FROM (SELECT ann
FROM koshikdocs LATERAL VIEW explode(annotations.features) annTable as ann) annmap)
decmap WHERE key='PPOS' AND value LIKE 'NN%';
Job 0: Map: 920 Reduce: 1   Cumulative CPU: 46694.33 sec   HDFS Read: 248123756147
HDFS Write: 10 SUCCESS
OK
476161210

LUND
UNIVERSITY

A Snapshot of the Future

When wireless is perfectly applied, the whole earth will be converted into a huge brain, which in fact it is, all things being particles of a real and rhythmic whole. We shall be able to communicate with one another instantly, irrespective of distance. Not only this, but through television and telephony we shall see and hear one another as perfectly as though we were face to face, despite intervening distances of thousands of miles; and the instruments through which we shall be able to do this will be amazingly simple compared with our present telephone. A man will be able to carry one in his vest pocket.

Nikola Tesla, *Colliers*, January 30, 1926

LUND UNIVERSITY