



# Web based Multilingual Question Answering

Günter Neumann

LT-lab, DFKI &  
Computational Linguistics Department,  
Saarland University



# Machine Learning for Web-based QA

- Our interest:
  - Developing ML-based strategies for **complete end-to-end question answering** for different type of questions
    - Exact answers
    - Open-domain
    - Multilingual
- Our vision:
  - Complex QA system existing of a community of collaborative (smaller) ML-based QA-agents
  - QA as a basic functionality for larger systems, e.g., intelligent services, interactive Web, robots or androids



# Machine Learning for Web-based QA

- QA at Trec and Clef evaluation forums have created reasonable amount of freely available corpora
  - Question-Answer pairs
  - Multilingual and different types of questions
  - Contextual information: sentences (mainly news articles)
- Enables
  - Training, evaluating ML algorithms and
  - Comparisons with other approaches.



# Machine Learning for Web-based QA

- Our initial goals:
  - Extract **exact answers** for different types of questions **only** from **web snippets**
  - Use strong **data-driven** strategies
- Our current results:
  - ML-based strategies for **factoid**, **definition** and **list** questions
  - Mainly **unsupervised** statistical-based methods
  - **Language poor**: Stop-word lists and simplistic patterns as main language specific resources
  - Promising performance on Trec/Clef data (~ 0.55 MRR)

**F:** When was Madonna born?

**D:** What is Ubuntu?

**L:** What movies did James Dean appear in?



# Machine Learning for Web-based QA

- Current SOA approaches:
  - Large corpora of full text documents (**fetching problem**)
  - Recognition of utterances by aligning surface patterns with sentences within full documents (**selection problem**)
  - Exploitation of additional external concept resources such as encyclopedias, dictionaries (**wrapping problem**)
  - Do not provide clusters of potential senses (**disambiguation problem**)
- Our idea:
  - Extract from Web Snippets only (**avoid first three problems**)
  - Unsupervised sense disambiguation for clustering (**handle fourth problem**)
  - Language independent, e.g., English, German, Spanish



## Why Snippets only?

- Avoid downloading of full documents
- Snippets are automatically “anchored” around questions terms → Q-A proximity
- Considering N-best snippets → redundancy via implicit multi-document approach
- Via IR query formulation, search engines can be biased to favor snippets from specialized data providers (e.g., Wikipedia) → no specialized wrappers needed



## Core components of our webQA approach

- Generic seed patterns
  - Automatic generation of web search queries
  - Automatic generation of answer extraction patterns
- Word-pair-distance statistics
  - Identification of statistical regularities for word sequences
  - Extraction of answer context for factoid questions
- Semantic kernels
  - Clustering for definition and list based questions



# Example output: When was Madonna born?

DFKI - Experimental Question Answering System - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://experimental-qaetal.dfkf.de/experimental-qaetal/foo.do;jsessionid=31C887337A4C68FD79BE9C7F887783F4

machine learning web question answering

Meistbesuchte Seiten SPIEGEL ONLINE - Na... LEO Deutsch-Englisch... User-Profil: Günter Ne... Open Source Web De... LTAG-spinal: Treeban... Allegro CL 8.1 Docum... Main Page - Clairlib Open Directory RDF D... Text Analytics Wiki: S... www.grammatical-evo...

Seiten-Ladefehler Groupware - Günter Neumann - OPEN... DFKI - Experimental Question An...

## DFKI - MULTI LINGUAL WEB QUESTION ANSWERING SYSTEM

Enter Your Question:

### WHEN WAS MADONNA BORN?

[1000] 16 AUGUST 1958

- [52.0871] MADONNA IS ONE OF THE HOTTEST SINGER EVER ! MADONNA WAS BORN MADONNA CICCONE ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [50.8193] MADONNA WAS BORN ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [49.6008] ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ) IS AN AMERICAN MULTI-GRAMMY AWARD ...
- [47.4438] BORN 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [43.6032] WHERE WAS MADONNA BORN ? ANSWER MADONNA WAS BORN MADONNA CICCONE AT BAY CITY MICHIGAN ON 16 AUGUST 1958 ...
- [27.7543] 16 AUGUST 1958 BAY CITY MICHIGAN USA ...

[356.1016] 1958

- [52.0871] MADONNA IS ONE OF THE HOTTEST SINGER EVER ! MADONNA WAS BORN MADONNA CICCONE ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [50.8193] MADONNA WAS BORN ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [46.0091] BORN MADONNA LOUISE VERONICA CICCONE IN AUGUST 1958 IN BAY CITY MICHIGAN ...
- [49.6008] ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ) IS AN AMERICAN MULTI-GRAMMY AWARD ...
- [47.4438] BORN 16 AUGUST 1958 IN BAY CITY MICHIGAN ...
- [43.6032] WHERE WAS MADONNA BORN ? ANSWER MADONNA WAS BORN MADONNA CICCONE AT BAY CITY MICHIGAN ON 16 AUGUST 1958 ...
- [27.7543] 16 AUGUST 1958 BAY CITY MICHIGAN USA ...

### OTHER RESULTS

- [52.0871] MADONNA IS ONE OF THE HOTTEST SINGER EVER ! MADONNA WAS BORN MADONNA CICCONE ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ....
- [50.8193] MADONNA WAS BORN ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ....
- [49.6008] ON 16 AUGUST 1958 IN BAY CITY MICHIGAN ) IS AN AMERICAN MULTI-GRAMMY AWARD ....
- [47.4438] BORN 16 AUGUST 1958 IN BAY CITY MICHIGAN ....





# Example Output: What is epilepsy ?

DFKI - Experimental Question Answering System - Mozilla Firefox

http://www2.dfkf.de:8080/experimental-qaetal/foo.do

DFKI - MULTI LINGUAL WEB QUESTION ANSWERING SYSTEM

Enter Your Question:

**DEFINE:EPILEPSY**  
POTENTIAL SENSE: YEARS

1. EPILEPSY *HAS BEEN A RECOGNIZED UNIQUE DISORDER FOR THOUSANDS OF YEARS. ...*  
POTENTIAL SENSE: USUALLY

1. ABSENCE EPILEPSY *THAT CHARACTERIZED BY ABSENCE SEIZURES, USUALLY HAVING ITS ONSET IN CHILDHOOD OR ADOLESCENCE. ...*  
POTENTIAL SENSE: UNPROVOKED

1. AND, EPILEPSY *IS A CHRONIC DISORDER, THE HALLMARK OF WHICH IS RECURRENT, UNPROVOKED SEIZURES. ...*  
POTENTIAL SENSE: TREATMENT

1. TREATMENT FOR EPILEPSY, *A DISORDER THAT CAUSES SEIZURES; INCLUDES. ...*  
POTENTIAL SENSE: RECURRING SEIZURES

1. EPILEPSY *IS A BRAIN DISORDER THAT CAUSES PEOPLE TO HAVE RECURRING SEIZURES. ...*  
POTENTIAL SENSE: NERVE

1. EPILEPSY *IS A DISORDER IN WHICH THERE IS EXCESSIVE ELECTRICAL ACTIVITY IN THE NERVE CELLS OF THE BRAIN, WHICH RESULTS IN INVOLUNTARY MOVEMENT OR CHANGES IN AWARENESS. ...*  
2. EPILEPSY *IS A BRAIN DISORDER IN WHICH CLUSTERS OF NERVE CELLS, OR NEURONS, IN THE BRAIN SOMETIMES SIGNAL ABNORMALLY. ...*  
POTENTIAL SENSE: GROUP

1. IMITATORS OF EPILEPSY *ARE A DIVERSE GROUP THAT INVOLVE CONSIDERATION OF MANY AREAS OF INTERNAL MEDICINE, NEUROLOGY, AND PSYCHIATRY. ...*  
POTENTIAL SENSE: FLICKERING

Suchen: kann

Fertig



# Example Output: What is epilepsy ?

- Our system's answer in terms of clustered senses:

----- **Cluster STRANGE** -----

0<->In epilepsy, the normal pattern of neuronal activity becomes disturbed, causing strange...

----- **Cluster SEIZURES** -----

0<->Epilepsy, which is found in the Alaskan malamute, is the occurrence of repeated seizures.

1<->Epilepsy is a disorder characterized by recurring seizures, which are caused by electrical disturbances in the nerve cells in a section of the brain.

2<->Temporal lobe epilepsy is a form of epilepsy, a chronic neurological condition characterized by recurrent seizures.

----- **Cluster ORGANIZATION** -----

0<->The Epilepsy Foundation is a national, charitable organization, founded in 1968 as the Epilepsy Foundation of America.

----- **Cluster NERVOUS** -----

0<->Epilepsy is an ongoing disorder of the nervous system that produces sudden, intense bursts of electrical activity in the brain.

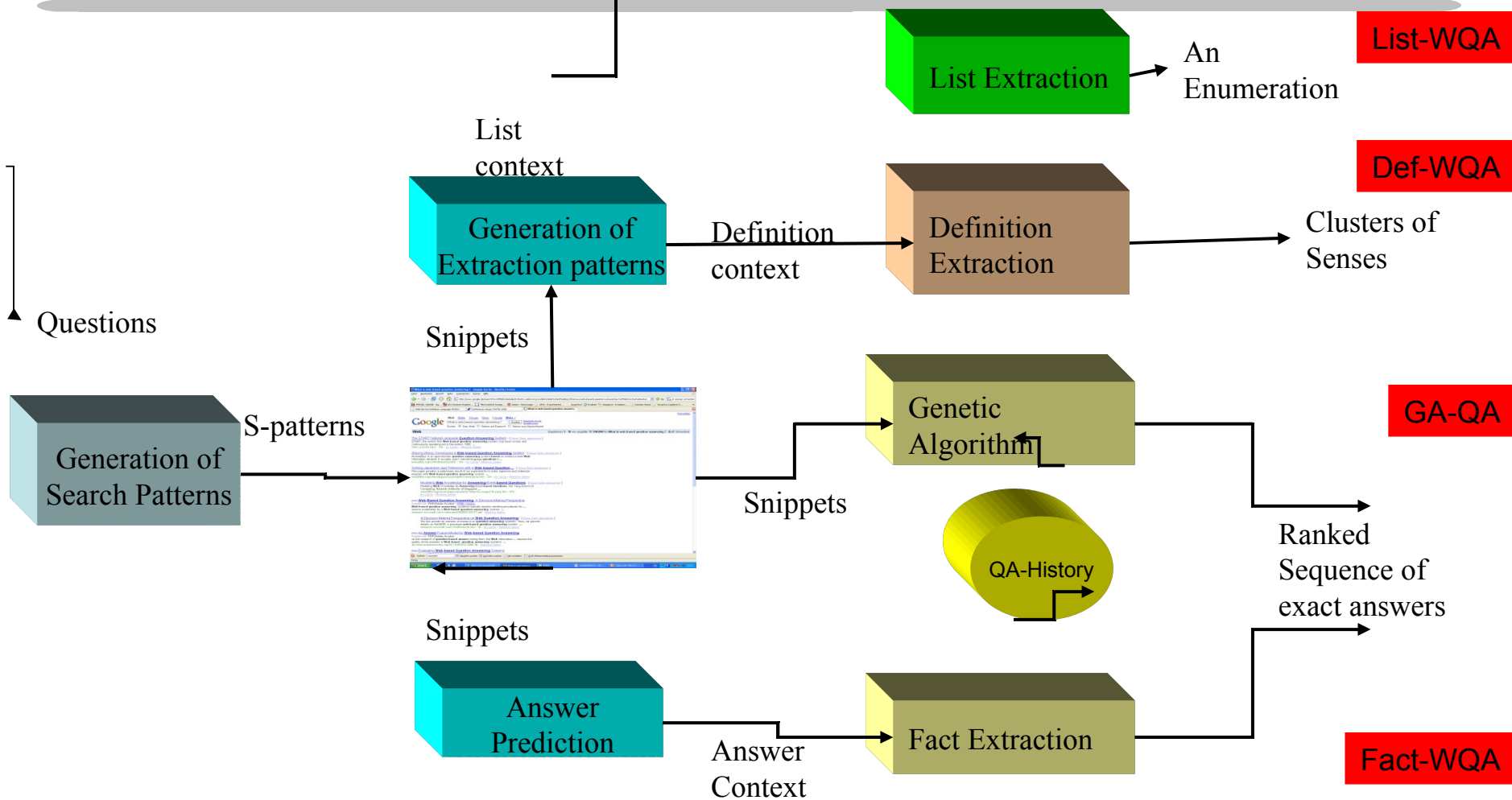


## EXample output: Novels written by John Updike?

- Answers: [In the Beauty, Lilies, National Book Award, Poorhouse, Poorhouse Fair, Rabbit, Rabbit Angstrom, Rabbit At Rest, Rabbit Is Rich, Rabbit Redux, Rabbit Run, Roger, Terrorist, The Centaur, The Coup, The Poorhouse Fair, The Witches of Eastwick, Villages, YOUR SHOES TOO BIG TO KICKBOX GOD]
- **###0###**The SALON Interview: John Updike. **#####** THE SALON INTERVIEW: JOHN UPDIKE "As close as you can get to the stars" . **#####** ... novel "In the Beauty of the Lilies" -- a vigorous and expansive book that tracks four generations in a single American family -- as well as a career that has spanned some 40 books, including 17 novels ... **#####** **#####**
- **###1###**John Updike, Writer. **#####** John Updike 1932 - Novels. **#####** Updike, John, -- The Poorhouse Fair, 1959. -- Rabbit, Run, 1960. -- The Centaur, 1963. -- Of the Farm, 1965. -- Couples, 1968. -- Rabbit Redux, 1971. **#####**
- **###2###**PAL: John Updike (1932- ). **#####** Comic Morality in The Centaur and the Rabbit Novels. **#####** NY: Peter Lang, 2005. **#####** Luscher, Robert M. John Updike: a study of the short fiction. **#####** NY: Twayne, 1993. **#####**
- **###3###**John Updike - Wikipedia, the free encyclopedia. **#####** is well known for his careful craftsmanship and prolific writing, having published 22 novels. **#####** The book's title is "**#####**
- **###4###**CRITICAL MASS: Reviewing 101: John Updike's rules. **#####** ago, in the introduction to "Picked Up Pieces," his second collection of assorted prose, John Updike. **#####** Was there a particular reason for this to be written using only male examples, or are we being old. **#####**

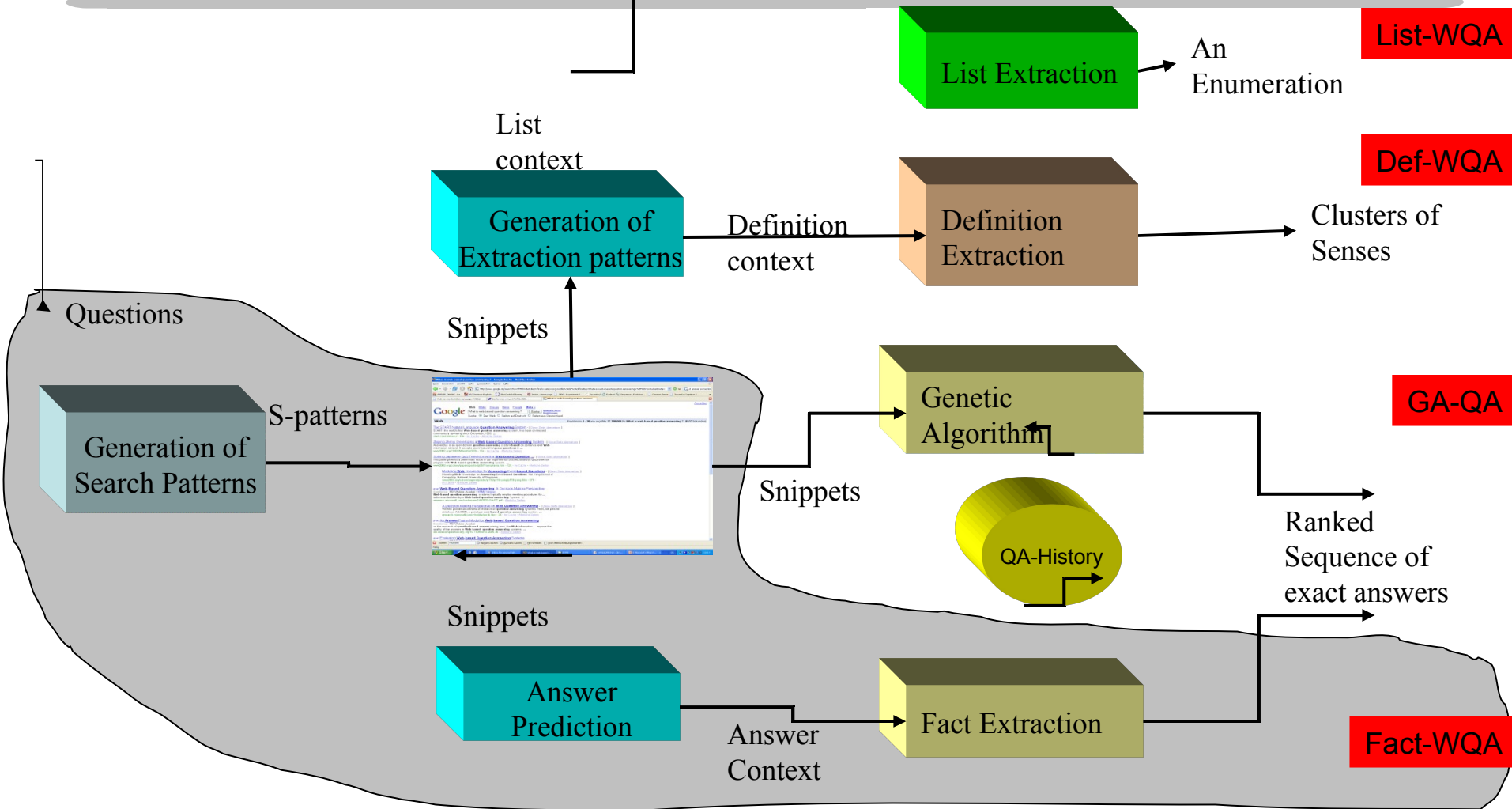


# Multilingual Web based QA: Overview





# Multilingual Web based QA: Overview





# Fact-WQA - Technology

<http://amasci.com/tesla/tradio.txt> TESLA INVENTED RADIO? ... He invented modern radio, but made such serious business mistakes that the recognition (to say ...

- Consult only snippets
  - Submit NL question string (no query refinement, expansion, reformulation, ...)
- Goal
  - Identify **smallest possible phrases** from snippets that contain exact answers (AP phrases)
  - Do not make use of any smoothing technology or pre-specified window sizes or length of phrases
- Answer extraction
  - Use only very **trivial patterns** for extracting exact answers from AP phrases
  - Only Wh-keywords, distinguish type of tokens, punctuation symbols for sentence splitting

The prime minister Tony Blair said that

The prime minister Tony Blair said that

Who → Person; When → Time



# Factoid-WQA – Technical Details

- Snippet-Document:

$$D = \{ \langle \omega_i, \omega_j, \epsilon, \text{freq}(\omega_i, \omega_j, \epsilon) \rangle \}$$

radio \* \* \* Tesla; 3  
 Tesla \* \* radio; 6  
 Tesla \* \* \* \* \* radio; 1

„The president of France went on  
 Holidays yesterday“

„The president of France \* on  
 Holidays \*“

„The president of France“, „on  
 Holidays“

- QA-specific ranking of sentences

$$M_{ij}(S_s) = \begin{cases} \text{freq}(\omega_i, \omega_j, \epsilon) & \text{if } i < j \\ \text{freq}(\omega_j, \omega_i, \epsilon) & \text{if } j > i \\ 0 & \text{otherwise} \end{cases}$$

$$\text{rank}(S_s) = \lambda_{\max}(M(S_s))$$

- Define: Sequences of pairs of words which occur with a high frequency in M (i.e., in a sentence) are *chains of related words* (AP phrases)
- Words with no strong relation with any other word in  $S_s$  are replaced with \*  
 → defines cutting points for sentences

# Factoid-V

Difference wrt. a ranking based on *n-grams* (e.g., AskMSR)

- We do not have any dependency on lengths
- We do not need to estimate back-off probabilities
- Long sentences will tend to have a lower rank than small sentences

## • Snippet-Document

$$D = \{ \langle \omega_i, \omega_j, \epsilon, \text{freq}(\omega_i, \omega_j, \epsilon) \rangle \}$$

radio \* \* \* Tesla; 3  
 Tesla \* \* radio; 6  
 Tesla \* \* \* \* \* radio; 1

$$M_{ij}(S_s) = \begin{cases} \text{freq}(\omega_j, \omega_i, \epsilon) & \text{if } i < j \\ \text{freq}(\omega_i, \omega_j, \epsilon) & \text{if } j > i \\ 0 & \text{otherwise} \end{cases}$$

„The president of France went on  
 Holidays yesterday“

„The president of France \* on  
 Holidays \*“

„The president of France“, „on  
 Holidays“

$$\text{rank}(S_s) = \lambda_{\max}(M(S_s))$$

- **Define:** Sequences of pairs of words which occur with a high frequency in  $M$  (i.e., in a sentence) are *chains of related words* (AP phrases)
- Words with no strong relation with any other word in  $S_s$  are replaced with \*  
 → defines cutting points for sentences





# Factoid-WQA – Experiments

- Pattern-based Answer extraction
  - Simplistic extraction patterns
  - Open-domain fact questions (889 from Clef 2004)
  - Answers from Web (DE,EN,ES,P)
  - **0.52 MRR**

CA	Total	MRR	NAG(%)	WAG(%)	NAF(%)	1(%)	2(%)	3(%)
WHEN	218	0.60	25.11	10.96	21.46	35.16	5.02	1.8
WHERE	232	0.57	10.77	24.14	20.68	30.60	9.91	3.87
WHO	439	0.98	11.39	27.56	32.57	18.90	6.83	2.73

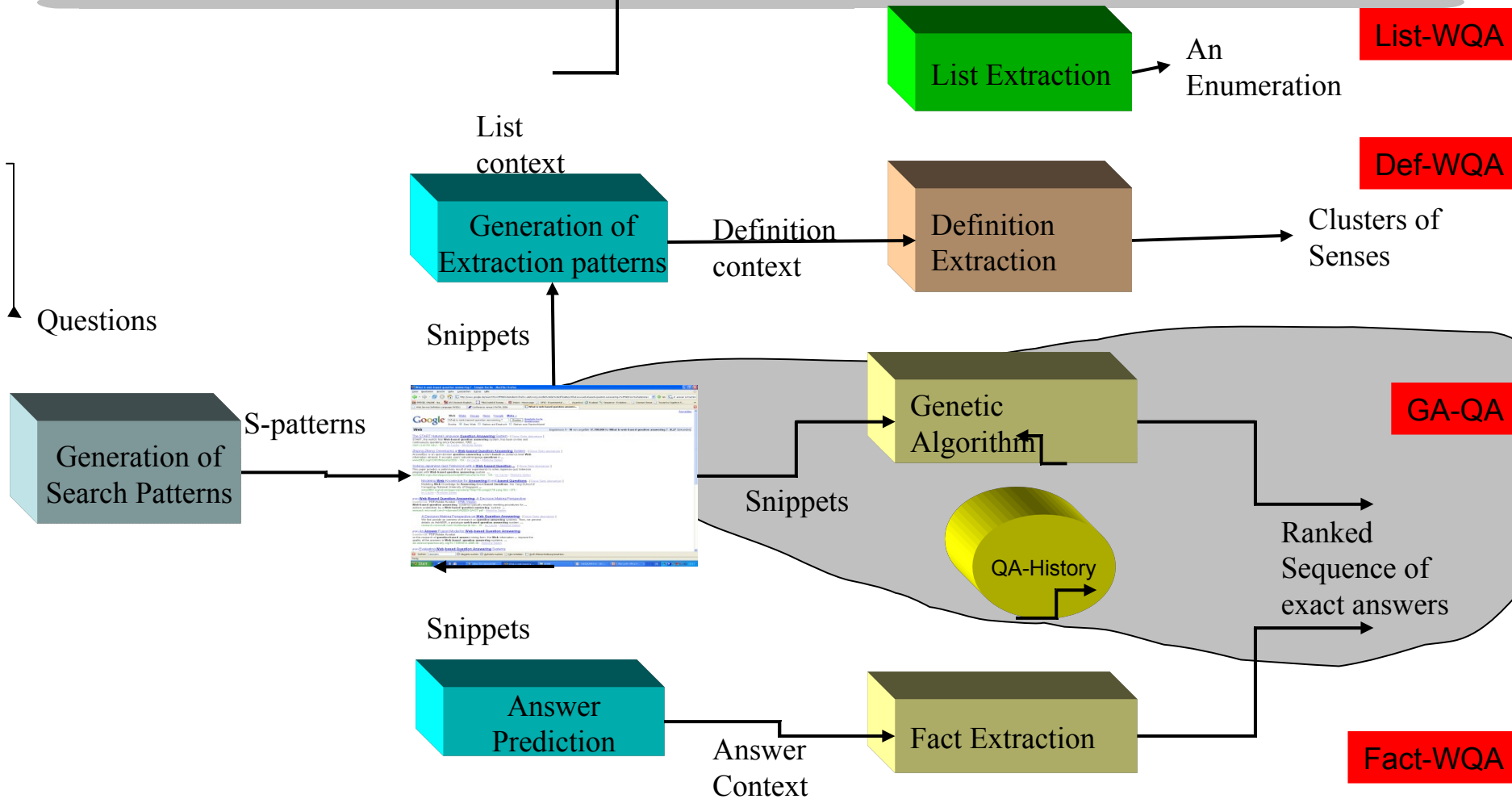
- Two types of answers:
  - Exact Answer:
    - Exact matching with the answer provided by CLEF.
  - Inexact Answer:
    - Are not exact answers, but they are very close answers:
      - **WHERE**: not only city name, country name is also correct.
      - **WHO**: variants like „G. Bush“, „George W. Bush“.
      - **WHEN**: „6 1945“, „1945“.

Lita&Carbonell:2004:  
MRR=0.447

for 296 English temporal questions for exact answer matching in TREC data



# Multilingual Web based QA: Overview

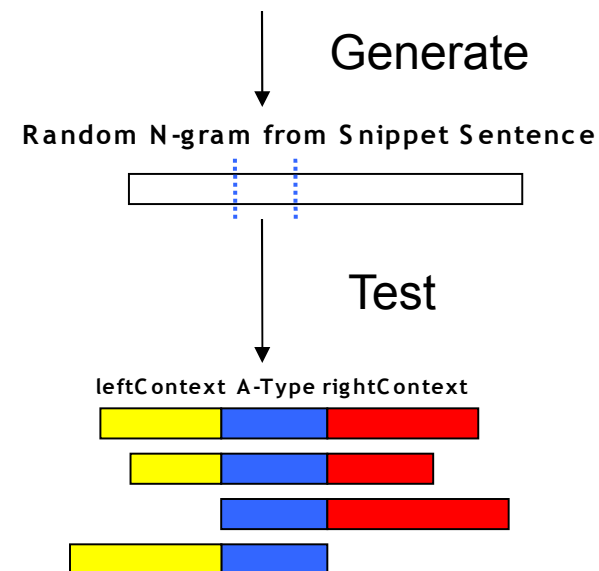




# GA-QA - Technology

- Goal:
  - Manually encoding of patterns for answer extraction is at least difficult, because snippets do have “un-predictive structure”
  - Compute answer candidates AC **via random search**
  - Validate/adapt AC on basis of past results computed by Factoid-WQA (**QA-store**)
- Answer extraction
  - QA-history oriented alignment of context of answer candidates AC
  - Word-pair-distance statistics for A-type compatible elements from QA-store (left/right model)
- Genetic algorithm for:
  - Identification of AC
  - Stretching/shrinking of context and AC
  - Specific operations for crossover and mutation
- Figueroa & Neumann, Evolutionary Computing Journal, 2008

<http://amasci.com/tesla/tradio.txt>  
**TESLA INVENTED RADIO? ...**  
**He invented modern radio, but made such serious business mistakes that the recognition (to say ...**





# GA-QA – Experiments

- GA-based Answer extraction (GAQA)
  - Relation-open questions (Clef)
  - Relation-closed questions
    - E.g., X invented Y
- Baseline
  - Most frequent subsequences
  - TFIDF statistics

Performance of GA-QA relative to a baseline

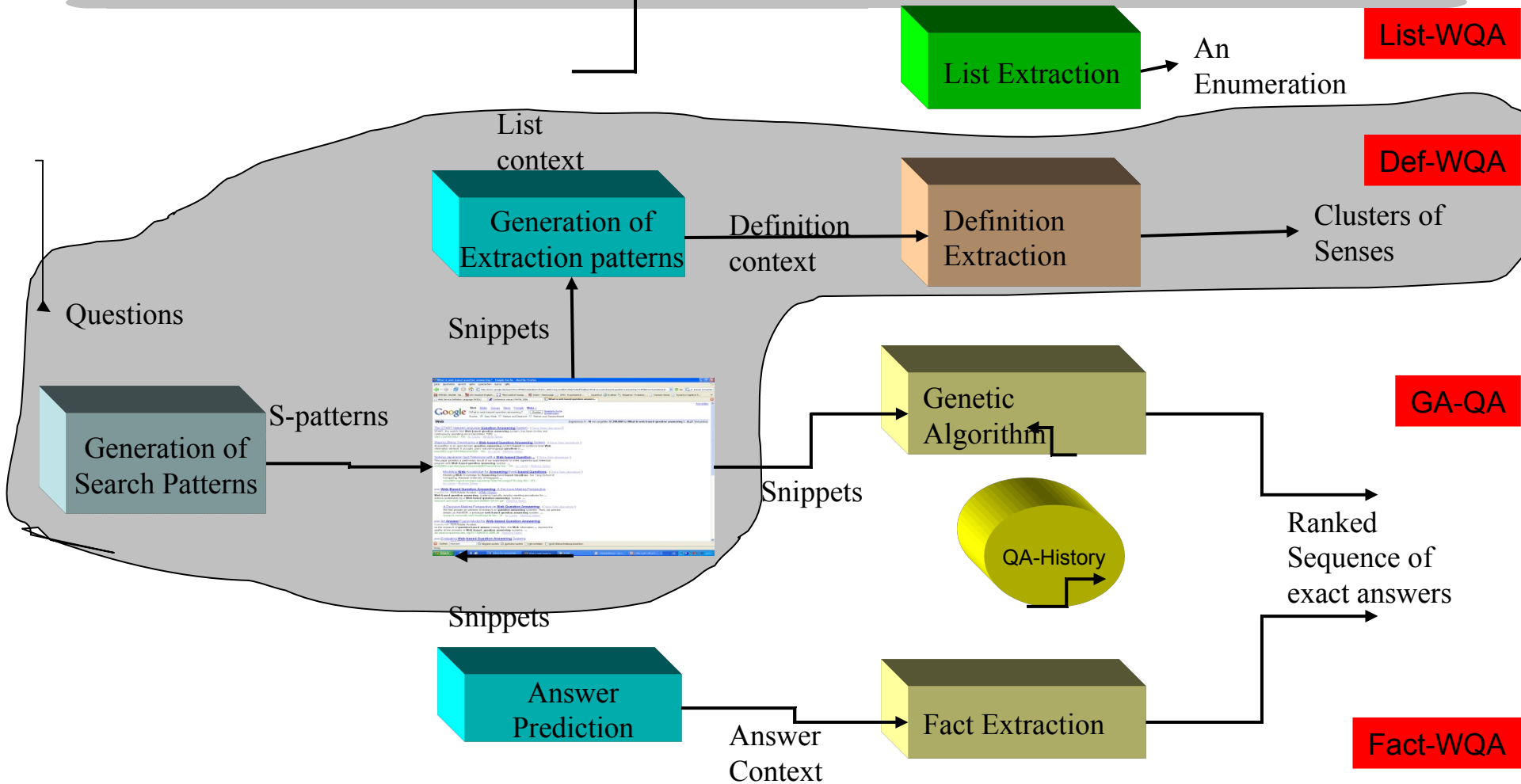
Strategy	MRR	Total	1	2	3	4	5	AA
Baseline	0.376	403	137	92	78	42	41	13
GAQA	0.497	401	242	78	38	31	12	14

MRR for individual corpora

Corpus	# Quest.	NAS	Baseline	GAQA
CLEF-2004	75	24	0.309	0.387
Inventions	185	28	0.421	0.502
Presidents	89	1	0.524	0.571
Pr. Ministers	76	5	0.473	0.706
Composers	100	23	0.315	0.500
Locations	43	1	0.568	0.638
Dates	145	7	0.173	0.365



# Multilingual Web based QA: Overview





# ML for Definition Questions – Def-WQA

- Questions such as:
  - *What is a prism?*
  - *Who is Ben Hur?*
  - *What is the BMZ?*
- Answering consists in collecting as much descriptive information as possible (*nuggets*):
  - *The distinction of relevant information*
  - *Multiple sources*
  - *Redundancy*
- Def-WQA extracts descriptive sentences only from web snippets:
  - Avoid processing and downloading a wealth of documents.
  - Avoid specialized wrappers (for dictionaries and encyclopedias)
  - Extend the coverage by boosting the number of sources through simple surface patterns (also here: KB poor approach)
  - Due to the massive redundancy of web, chances of discriminating a paraphrase increase markedly.

Note: Our goal is on open domain question answering, i.e., no restrictions on the topic.



# Surface patterns for definition candidates

- Some surface patterns
  1. X (is|are|was|were) (a|the|an) Y.
    - “*Noam Chomsky is a writer and a critic...*”
  1. X , or Y. ↔ Y , or X.
    - “Myopia, or nearsightedness, can be ..”
  1. X (Y) ↔ X (Y).
    - “United Nations (UN)”
  1. X (become|became|becomes) Y.
    - “.... Althea Gibson became the first African American ...”
- We have manually defined a total of 8 patterns\*
- For example, “What is the DFKI?”, then surface patterns:
  - “DFKI is a” OR “DFKI is an” OR “DFKI is the” OR “DFKI are a”...
  - “DFKI, or ”.
  - “(DFKI)”
  - “DFKI becomes” OR “DFKI become” OR “DFKI became”
- Some fetched sentences:
  - “**DFKI is the** German Research Center for Artificial Intelligence”.
  - “**The DFKI is a** young and dynamic research consortium”
  - “**Our partner DFKI is an** example of excellence in this field.”
  - “**the DFKI, or** Deutsches Forschungszentrum für Künstliche ...”
  - “German Research Center for Artificial Intelligence (**DFKI GmbH**)”

\*cf. also Hildebrandt et al:2004, Miliaraki et al:2004



# Selecting and Clustering Definition Candidates

- Relaxed string matching for identifying possible paraphrases/mentionings of target in snippets
- Jaccard measure (cf. W. Cohen, 2003)
  - computes the ratio of common different words to all different words
  - $J(\text{"The **DFKI**"}, \text{"**DFKI**"}) = 0.5$
  - $J(\text{"Our partner **DFKI**"}, \text{"**DFKI**"}) = 0.333$
  - $J(\text{"**DFKI** GmbH"}, \text{"**DFKI**"}) = 0.5$
  - $J(\text{"His main field of work at **DFKI**"}, \text{"**DFKI**"}) = 0.1428$
- Avoids the need for additional specific syntax oriented patterns or chunk parsers
- LSA-based clustering into potential senses
  - Determine semantically similar words/substrings
  - Define different clusters/potential senses on basis of non-membership in sentences
- Ex: What is Question Answering ?
  - SEARCHING: Question Answering is a computer-based activity that involves searching large quantities of text and understanding both questions and textual passages to the degree necessary to. ...
  - INFORMATION: Question-answering is the well-known application that goes one step further than document retrieval and provides the specific information asked for in a natural language question. ...
  - ...





# Def-WQA: Results

Corpus	# Questions	# Answered Def-WQA/Baseline	# nuggets Def-WQA/Baseline
TREC 2003	50	50/38	14.14/7.7
CLEF 2006	152	136/102	13.13/5.43
CLEF 2005	185	173/160	13.86/11.08
TREC 2001	133	133/81	18.98/7.35
CLEF 2004	86	78/67	13.91/5.47

Corpus	F-score ( $\beta = 5$ )
Trec 2003	0.52

Trec 2003 best systems  
(on newspaper articles):  
0.5 – 0.56

## Notes:

- we prefer sentences instead of nuggets (readability)
- we need no predefined window size for nuggets (~ 125 characters)
- Still open: Merging of clusters (external knowledge sources needed !?)



## Summing up: Machine Learning based Web-QA

- Achievements
  - End-to-end ML QA learners for specific question types
  - Open-domain
  - Multilingual
- Next goals:
  - QA based Interactive information extraction
  - Crossing language barrier