Ontology-based Information Extraction and Question Answering – Coming Together

Günter Neumann

LT lab, DFKI, Saarbrücken
 Ontology-based information extraction

- Ontology defines target knowledge structures
  - i.e., type of entities, relations, templates
- IE for identifying and extracting instances
- Merging of partial instances by means of reasoning
Question answering from text and Web

- Answering questions about who, what, whom, when, where or why

- Question analysis:
  - “Human carries ontology”
  - Identifies the partially instantiated relation expressed in a Wh-question
  - Identification of the “expected answer type”

- Answer extraction
  - The „information extraction“ part of QA
  - Also here: RTE for validating answer candidates (cf. Clef 2007/2008)

Who is Prime Minister of Canada?
-> PM_of(person:X,country:Canada)
-> EAT=person

Stephen Harper was sworn in as Canada’s 22nd Prime Minister on February 6, 2006. (Source: http://pm.gc.ca/eng/pm.asp)
Two Possible Approaches of OBIES+QA

☆ Entailment-based QA

- Domain ontology as interface between NL and DB
- Bijective mapping between NL patterns and DB patterns
- Textual entailment for mastering the mapping/reasoning
- EU project QALL ME

☆ Web-based ontology learning using QA

- Unsupervised methods for extracting answers for factoid, list and definition based question
- Basis for large-scale, web-based bottom-up knowledge extraction and ontology population
- BMBF project Hylap
LT lab

Architectures of QA Systems

DB-QA

Text-QA

Hybrid-QA

NL Question

NL Question

NL Question
LT lab

Architectures of QA Systems

**DB-QA**

- NL Question
- NL2DB Interface
- SQL Query
- DB System
- Answer: facts

**Text-QA**

- NL Question

**Hybrid-QA**

- NL Question

German Research Center for Artificial Intelligence

OBIES 2008 • Sept. 2008

Mittwoch, 17. März 2010
**LT lab**

**Architectures of QA Systems**

**DB-QA**
- NL Question
- NL2DB Interface
- SQL Query
- Answer: facts

**Text-QA**
- NL Question
- NL2IR Interface
- Keywords
- Answer Extraction
- Answer: Text fragments

**Hybrid-QA**
- NL Question
- NL Interface
- NL2DB Interface
- SQL Query
- Keywords
- Answer Extraction
- Answer Integration
- Answer: facts

---

German Research Center for Artificial Intelligence

Mittwoch, 17. März 2010
Architectures of QA Systems

**DB-QA**

1. **NL Question**
2. **NL2DB Interface**
3. **SQL Query**
4. **Answer: facts**

**Text-QA**

1. **NL Question**
2. **NL2IR Interface**
3. **Keywords**
4. **Answer Extraction**
5. **Answer: Text fragments**

**Hybrid-QA**

1. **NL Question**
2. **NL Interface**
3. **NL2DB Interface**
4. **SQL Query**
5. **Keywords**
6. **Answer Extraction**
7. **Answer: facts**
☆ Hybrid QA:

- Increase of semantic structure (Semantic Web, Web 2.0) ⇒ Fusion of ontology-based DBMS and information extraction from text
- Dynamics and interactivity of Web requests for additional new complexity of the NL interface.

„Who wrote the script of Saw III?"

Complex linguistic & knowledge-based reasoning

SELECT DISTINCT ?writerName WHERE
{ ?movie name "Saw III"^^string . ?movie hasWriter ?writer . ?writer name ?writerName . }

„Who is the author of the script of the movie Saw III?"
Possible approaches

☆ Full computation (inference)
  - ⇒ AI complete; especially, if incomplete/wrong queries are allowed

☆ Controlled sublanguage
  - A user may only express questions using a constrained grammar and with unambiguous meaning
  - ⇒ cognitive burden is not acceptable

☆ Controlled mapping
  - One-to-one mapping between NL patterns and DB-query patterns
  - Flexible use of NL possible through methods of textual inference
Textual Inference

☆ Motivation: textual variability of semantic expressions

☆ Idea: for two text expressions T & H:
  – Does text T justify an inference of hypothesis H?
  – Is H semantically entailed in T?

☆ PASCAL Recognizing Textual Entailment (RTE) Challenge
  – 2008: 4th RTE (at TAC), 26 groups (two subtasks)

☆ RTE is considered as a core technology for a number of text based applications:
  – QA, IE, semantic search, text summarization, …
☆ RTE successfully applied to answer validation

- Example
  - Q: „In which country was Edouard Balladur born?”, A: “France”
  - T: „Paris, Wednesday CONSERVATIVE Prime Minister Edouard Balladur, defeated in France’s presidential election, resigned today clearing the way for President-elect Jacques Chirac to form his own new government…”

- Entailed(Q+A, T) ⇒ YES/NO ?

- Clef 2008, AVE task ⇒ DFKI best results for English and German

☆ New: RTE for semantic search

- Does question X entail an (already answered) question Y ?
**Current Control Flow**

- **NL Question**
  - Linguistic Analysis
  - Textual Entailment

- **Domain ontology**
  - Bijective mapping between NL-patterns and SPARQL-patterns
  - DBMS: RDF expressions
  - Answers: values

**Domain ontology**
- attr:val
- attr:val
- attr:val
- attr:val

**Answers:**
- values

**DBMS: RDF expressions**
- attr:val
- attr:val
- attr:val
LT lab

Current Control Flow

NL Question

Linguistic Analysis

Textual Entailment

Domain ontology

Bijective mapping between NL-patterns and SPARQL-patterns

DBMS: RDF expressions

Answers: values

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann man [MOVIE] sehen?</td>
<td>SELECT ?cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT ?location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT ?director ...</td>
</tr>
</tbody>
</table>
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Textual Entailment

Domain ontology

Bijective mapping between NL-patterns and SPARQL-patterns

DBMS: RDF expressions

Answers: values

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann man [MOVIE] sehen?</td>
<td>SELECT ?cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT ?location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT ?director ...</td>
</tr>
</tbody>
</table>

German Research C

OBIES 2008 • Sept. 2008

Mittwoch, 17. März 2010
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment

Bijective mapping between NL-patterns and SPARQL-patterns

Domain ontology

DBMS: RDF expressions

Answers: values

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann [MOVIE] sehen?</td>
<td>SELECT ?cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT ?location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT ?director ...</td>
</tr>
</tbody>
</table>
**LT lab**

**Current Control Flow**

**Wo läuft Dreamgirls?**

**NL Question**

**Linguistic Analysis**

**Wo läuft [movie]?**

**Textual Entailment**

![Diagram showing the flow of information from NL question to SPARQL-patterns and RDF expressions, leading to answers.](image)

- **Domain ontology**
- **Bijective mapping between NL-patterns and SPARQL-patterns**
- **DBMS: RDF expressions**
- **Answers: values**

**Table:**

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann man [MOVIE] sehen?</td>
<td>SELECT %Cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT %Location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT %Director ...</td>
</tr>
</tbody>
</table>

**German Research C**

**OBIES 2008 • Sept. 2008**

**Mittwoch, 17. März 2010**
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment

"SELECT ?cinema ... WHERE ?movie name Dreamgirls ..."

Bijective mapping between
NL-patterns and SPARQL-patterns

DBMS: RDF expressions

Domain ontology

Answers: values

Table:

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann man [MOVIE] sehen?</td>
<td>SELECT ?cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT ?location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT ?Director ...</td>
</tr>
</tbody>
</table>
Wo läuft Dreamgirls?

Wo läuft [movie]?

NL Question

Linguistic Analysis

Textual Entailment

Domain ontology

DBMS: RDF expressions

Bijective mapping between NL-patterns and SPARQL-patterns

"SELECT ?cinema ... WHERE ?movie name Dreamgirls ..."

Answers: values

Xanadu
Advantages

✩ Inference remains on the linguistic level

✩ RTE method are by definition robust ⇒ supports processing of underspecified/illspecified requests

✩ Good interplay with ontology-based DB

✩ Opens up possibility to automatically learn mappings via ontology-based information extraction
LT lab  Ontology-based Information Extraction

☆ Extraction of relevant information from textual sources (Web pages)

☆ Integration of the extracted data into current DB

☆ Domain ontology as starting point:
  - Relevance
  - Normalization
  - Mapping
LT lab  Possible approaches

☆ Bootstrapping an ontology
  − Basic components for handling IE-specific subtasks expressed as Wh-questions
  − Unsupervised, language-independent approaches
  − Populating/extending domain ontology

☆ Interactive dynamic information extraction
  − Topic-based web crawling
  − IE system mines for all possible relevant entities and relations
  − See talk on Eichler et al., Friday, 13:30
Unsupervised Web-based Question Answering for ontology bootstrapping

☆ Our goal:

– Development of ML-based strategies for complete end-to-end answer extraction for different types of questions and the open domain.

☆ Our perspective:

– Extract exact answers for different types of questions only from web snippets
– Use strong data-driven strategies
– Evaluate them with Trec/Clef Q-A pairs

☆ Our current results:

– ML-based strategies for open domain factoid, definition and list questions
– Question type specific query expansion for controlling web search
– Unsupervised learning for answer extraction
– Promising performance ( ~ 0.5 MRR on Trec/Clef data)
Our goal:

- Development of ML-based strategies for complete end-to-end answer extraction for different types of questions and the open domain.

Our perspective:

- Extract exact answers for different types of questions only from web snippets
- Use strong data-driven strategies
- Evaluate them with Trec/Clef Q-A pairs

Our current results:

- ML-based strategies for open domain factoid, definition and list questions
- Question type specific query expansion for controlling web search
- Unsupervised learning for answer extraction
- Promising performance (~ 0.5 MRR on Trec/Clef data)
Current ML-based Web-QA System

- **NL-Question**
  - Lexico-syntactic patterns
  - NL-string(s)
  - Snippets

- **Surface E-patterns**
  - Definition context
  - Snippets

- **List Extraction**
  - List context
  - Snippets

- **Definition Extraction**
  - Definition context
  - Snippets

- **Genetic Algorithms**
  - Exact answ 1
  - Exact answ 2
  - …

- **QA-History**
  - Extraction via Trivial patterns

- **List-WQA**
  - Clusters of Potential senses
  - …

- **Def-WQA**
  - …

- **GA-QA**
  - …

- **Factoid-WQA**
  - …

**German Research Center for Artificial Intelligence**

Mittwoch, 17. März 2010
☆ 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

http://amasci.com/tesla/tradio.txt TESLA INVENTED RADIO? ... He invented modern radio, but made such serious business mistakes that the recognition (to say ...
Results for each question type over all languages.

<table>
<thead>
<tr>
<th>CA</th>
<th>Total</th>
<th>MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEN</td>
<td>218</td>
<td>0.60</td>
<td>25.11</td>
<td>10.96</td>
<td>21.46</td>
<td>35.16</td>
<td>5.02</td>
<td>1.8</td>
</tr>
<tr>
<td>WHERE</td>
<td>232</td>
<td>0.57</td>
<td>10.77</td>
<td>24.14</td>
<td>20.68</td>
<td>30.60</td>
<td>9.91</td>
<td>3.87</td>
</tr>
<tr>
<td>WHO</td>
<td>439</td>
<td>0.38</td>
<td>11.39</td>
<td>27.56</td>
<td>32.57</td>
<td>18.90</td>
<td>6.83</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Distribution of answer candidates (all languages).

<table>
<thead>
<tr>
<th>CA</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEN</td>
<td>33.82</td>
<td>55.42</td>
<td>7.91</td>
<td>2.84</td>
</tr>
<tr>
<td>WHERE</td>
<td>31.86</td>
<td>47.00</td>
<td>15.23</td>
<td>5.95</td>
</tr>
<tr>
<td>WHO</td>
<td>53.37</td>
<td>30.97</td>
<td>11.19</td>
<td>4.47</td>
</tr>
</tbody>
</table>
The results for the individual languages.

<table>
<thead>
<tr>
<th>CA(EN)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>when</td>
<td>69</td>
<td>0.69</td>
<td>15.69</td>
<td>15.69</td>
<td>17.65</td>
<td>45.10</td>
<td>3.92</td>
</tr>
<tr>
<td>where</td>
<td>64</td>
<td>0.74</td>
<td>7.81</td>
<td>12.5</td>
<td>15.62</td>
<td>53.12</td>
<td>10.93</td>
</tr>
<tr>
<td>who</td>
<td>148</td>
<td>0.50</td>
<td>7.43</td>
<td>12.83</td>
<td>32.43</td>
<td>33.78</td>
<td>10.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CA(DE)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wann</td>
<td>58</td>
<td>0.45</td>
<td>36.20</td>
<td>12.07</td>
<td>27.59</td>
<td>22.03</td>
<td>1.17</td>
</tr>
<tr>
<td>Wo</td>
<td>58</td>
<td>0.46</td>
<td>9.37</td>
<td>18.75</td>
<td>23.43</td>
<td>20.31</td>
<td>12.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CA(ES)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuándo</td>
<td>59</td>
<td>0.55</td>
<td>16.64</td>
<td>11.86</td>
<td>23.73</td>
<td>32.20</td>
<td>10.17</td>
</tr>
<tr>
<td>Dónde</td>
<td>63</td>
<td>0.59</td>
<td>10.93</td>
<td>31.25</td>
<td>15.62</td>
<td>26.56</td>
<td>10.93</td>
</tr>
<tr>
<td>Quién</td>
<td>86</td>
<td>0.27</td>
<td>9.65</td>
<td>40.68</td>
<td>28.96</td>
<td>11.72</td>
<td>6.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CA(PT)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quando</td>
<td>56</td>
<td>0.04</td>
<td>30.76</td>
<td>12.30</td>
<td>42.45</td>
<td>3.08</td>
<td>1.54</td>
</tr>
<tr>
<td>Onde</td>
<td>47</td>
<td>0.18</td>
<td>10.93</td>
<td>25</td>
<td>20.31</td>
<td>10.93</td>
<td>1.56</td>
</tr>
<tr>
<td>Quem</td>
<td>146</td>
<td>0.14</td>
<td>17.12</td>
<td>29.45</td>
<td>36.30</td>
<td>10.95</td>
<td>4.11</td>
</tr>
</tbody>
</table>
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

Exploit *only* web snippets:

- Avoid processing and downloading a wealth of documents.
- Avoid specialized wrappers (for dictionaries and encyclopedias)
- Snippets are automatically “anchored” around questions terms $\rightarrow$ Q-A proximity
- Considering N-best snippets $\rightarrow$ redundancy via implicit multi-document approach
- Extend the coverage by boosting the number of sources through simple surface patterns (also here: KB poor approach).
Determining descriptive phrases from snippets
Surface patterns, e.g., “What is the DFKI?”

- “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”
★ Surface patterns, e.g., “What is the DFKI?”

- “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 ...”
Determining descriptive phrases from snippets

☆ Surface patterns, e.g., “What is the DFKI?”
  - “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”

☆ 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. ...
  - ...
Example Output: What is epilepsy?

-Star 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.

...
Def-WQA: Results

<table>
<thead>
<tr>
<th>Corpus</th>
<th># Questions</th>
<th># Answered</th>
<th># nuggets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC 2003</td>
<td>50</td>
<td>50/38</td>
<td>14.14/7.7</td>
</tr>
<tr>
<td>CLEF 2006</td>
<td>152</td>
<td>136/102</td>
<td>13.13/5.43</td>
</tr>
<tr>
<td>CLEF 2005</td>
<td>185</td>
<td>173/160</td>
<td>13.86/11.08</td>
</tr>
<tr>
<td>TREC 2001</td>
<td>133</td>
<td>133/81</td>
<td>18.98/7.35</td>
</tr>
<tr>
<td>CLEF 2004</td>
<td>86</td>
<td>78/67</td>
<td>13.91/5.47</td>
</tr>
</tbody>
</table>

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)
- Def–WQA as a basis for more applications, e.g.,
  - list–based questions, web person identification, ontology learning
- Still missing: merging/splitting of partitions (evtl. using KBs and authority)
“What are 9 works written by Judith Wright?”

Search Query construction

Answer Candidate extraction

Answer Candidate selection

"What are 9 works written by Judith Wright?"

Search Query construction

Answer Candidate extraction

Answer Candidate selection

Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

“What are 9 works written by Judith Wright?”

Q1: (intitle:“Judith Wright”) AND (inbody:“works” OR inbody:“written”)

Q1: (intitle:“Judith Wright”) AND (inbody:“works” OR inbody:“written")

Max 80 snippets:
Most of Wright's poetry was written in the mountains of southern Queensland. ... Several of her early works such as 'Bullocky' and 'Woman to Man' became standard ...
“What are 9 works written by Judith Wright?”

Q1: \((\text{intitle:} \text{“Judith Wright”}) \land (\text{inbody:} \text{“works”} \lor \text{inbody:} \text{“written”})\)

Max 80 snippets:
Most of Wright's poetry was written in the mountains of southern Queensland. ... Several of her early works such as 'Bullocky' and 'Woman to Man' became standard ...

Apply 8 patterns \(\pi_i\) (hyponym, possessive, copula, quoting, etc.)

“What are 9 works written by Judith Wright?”

Q1: (intitle:“Judith Wright”) AND (inbody:“works” OR inbody:“written”)

Max 80 snippets:
Most of Wright’s poetry was written in the mountains of southern Queensland. ... Several of her early works such as ‘Bullocky’ and ‘Woman to Man’ became standard ...

Apply 8 patterns $\pi$ (hyponym, possessive, copula, quoting, etc.)

$\pi 4$: entity is $w+$ qfocus $w^*$

*Chubby Hubby* is .... Ben and Jerry’s *ice cream* brand.

List-WQA – Overview

"What are 9 works written by Judith Wright?"

Search Query construction

Qfocus \rightarrow \text{inbody}
NPs \rightarrow \text{intitle}
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND
(inbody:"works" OR inbody:"written")

Max 80 snippets:
Most of Wright's poetry was written in the mountains of southern Queensland. ...
Several of her early works such as 'Bullocky' and 'Woman to Man' became standard ...

Apply 8 patterns $\pi_i$ (hyponym, possessive, copula, quoting, etc.)

$\pi_4$: entity is \w+ qfocus \w*
Chubby Hubby is .... Ben and Jerry's ice cream brand.

Use Semantic kernel & Google N-grams

The Moving Image, Woman to Man, The Gateway,
The Two Fires, Birds, The Other Half, City Sunrise,
The Flame three and Shadow.
Answer Selection:

- Two measures **Accuracy** and $F_1$ score.
- Two values
  - All questions
  - Only questions where at least one answer was found in the fetched snippets.
- Duplicate answers have also an impact on the performance. For instance:
  - “Maybelline” (also found as “Maybellene” and “Maybeline”).
  - John Updike’s novel “The Poorhouse Fair” was also found as “Poorhouse Fair”.

<table>
<thead>
<tr>
<th>Systems \ Trec</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>ListWebQA($F_1$)</td>
<td>0.35/0.46</td>
<td>0.34/0.37</td>
<td>0.22/0.28</td>
<td>0.30/0.40</td>
</tr>
<tr>
<td>ListWebQA(Acc)</td>
<td>0.5/0.65</td>
<td>0.58/0.63</td>
<td>0.43/0.55</td>
<td>0.47/0.58</td>
</tr>
<tr>
<td>Top one(Acc.)</td>
<td>0.76</td>
<td>0.65</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Top two(Acc.)</td>
<td>0.45</td>
<td>0.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Top three(Acc.)</td>
<td>0.34</td>
<td>0.11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Top one($F_1$)</td>
<td>-</td>
<td>-</td>
<td>0.396</td>
<td>0.622</td>
</tr>
<tr>
<td>Top two($F_1$)</td>
<td>-</td>
<td>-</td>
<td>0.319</td>
<td>0.486</td>
</tr>
<tr>
<td>Top three($F_1$)</td>
<td>-</td>
<td>-</td>
<td>0.134</td>
<td>0.258</td>
</tr>
<tr>
<td>Yang &amp; Chua 04 ($F_1$)</td>
<td>-</td>
<td>-</td>
<td>.464 ~ .469</td>
<td>-</td>
</tr>
</tbody>
</table>

**We conclude:**
Encouraging results, competes well with 2nd best;
Still creates too much noise;
☆ WebQA:
  - Combining generic lexico-syntactic patterns with unsupervised answer extraction from Snippets only
  - Language independent and multilingual
  - Our approach has a close relationship to the new approach of unsupervised IE, e.g., Etzioni et al., Weikum et al., Rosenfeld & Feldman

☆ Information extraction
  - WebQA as a generic tool for web-based bottom-up knowledge extraction and ontology population
  - Ontology-based clustering for unsupervised information extraction
    - Use ontology for generating QA requests -> ontology-driven active QA
    - Use web QA for populating and extending ontology
  - Interactive dynamic information extraction