Bridging NLP and Semantic Web to enhance user interactions with the Web of Data

Elena Cabrio

INRIA - Sophia Antipolis, France (Wimmics Team)
Motivations

The web of data continues to grow (e.g., Linked Data initiatives).

How can we support end users in querying and exploring this novel, massive and heterogeneous structured information space? [Lopez et al, 2011]

Development of methods for a flexible mapping between NL expressions and concepts and relations in a structured KB.
Motivations
The web of data continues to grow (e.g. Linked Data initiatives)

How can we support end users in querying and exploring this novel, massive and heterogeneous structured information space? [Lopez et al, 2011]

Development of methods for a flexible mapping between NL expressions and concepts and relations in a structured KB
Outline

1. Question Answering over Linked Data
   - Our approach: QA using WikiFramework patterns
   - QAKiS system description
   - Demo
   - Participation to the QALD-2 challenge
   - Future work

2. Combining NLP and Argumentation Theory for supporting online debates interaction
Question Answering over Linked Data

• **Crucial issue:** how to convert a NL question into a corresponding query in a formal language? (e.g. SPARQL)

• **Goal of our research:** allow an end user to submit a query to an RDF triple store in her mother tongue and get results in the same language
Question Answering over Linked Data

LIMITATIONS of most of the current approaches:

- limited support for expressing queries (keyword-based approach, e.g. Swoogle, Watson, Sindice)
- a narrow search scope (in particular, closed-domain approaches, e.g. Orakel, Gino, Panto)
- limited ability to cope with ambiguity (users are asked to disambiguate their queries, e.g. SearchWebDB)

STATE-OF-THE-ART QA systems over the SW:

- PowerAqua (Lopez et al. 2009)
- Freya (Bizer, Lehmann et al. 2009)
Our proposal: **Question Answering** wiKiframework-based **System**

- **Task:** QA over structured knowledge bases, extracted from NL corpora (e.g. DBpedia)
- **Question interpretation:** relation-based match, i.e. fragments of the questions are matched to relational textual patterns automatically collected from Wikipedia (WikiFramework repository)

E. Cabrio, *Bridging NLP and Semantic Web to enhance user interactions with the Web of Data* July 19, 2012.
WikiFramework

• **Relational patterns:** lexicalization of ontological relations

  birthDate(Person, Date)

  [Person was born in Date]
  [Person(Date)]
  [Person, whose date of birth is Date]

• We provide a robust **methodology** to collect relational patterns in several languages for DBpedia relations
WikiFramework pattern collection

<http://dbpedia.org/resource/Golden_Gate_Bridge>
<http://dbpedia.org/ontology/crosses>
<http://dbpedia.org/resource/Golden_Gate>
WikiFramework pattern collection

• **Subject** and **object** of the **relation** are substituted with the corresponding DBpedia ontology classes:

\[
\text{crosses(Bridge, River)} \\
\Rightarrow \text{[Bridge is a suspension bridge spanning the River]}
\]

• Different matching strategies:
  - exact string match
  - string match after cleaning page disambiguation
  - match of tokenized/lemmatized strings (Stanford Core NLP)
  - match of normalized value for numeric or time ranges
WikiFramework pattern repository

<table>
<thead>
<tr>
<th>Relation</th>
<th>Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>spouse</td>
<td>Person wife Person</td>
</tr>
<tr>
<td></td>
<td>Person married Person</td>
</tr>
<tr>
<td></td>
<td>Person husband Person</td>
</tr>
<tr>
<td>crosses</td>
<td>Bridge spanning River</td>
</tr>
<tr>
<td></td>
<td>Bridge bridge River</td>
</tr>
<tr>
<td></td>
<td>Bridge crossing River</td>
</tr>
</tbody>
</table>

- Patterns are clustered, and ranked according to frequency
- 3 sets of 20 keywords are created for each relation (most frequent tokens, lemmas and stems)
- the CamelCase name of the relation is added to the set of keywords
QAKIS architecture

QA system

Query Generator

- Question file parser
- Query generator
- Typed Question generator
- Property identifier
- Query selector

Pattern Matching

- Pattern matcher
- Pattern Base

Which river does the Brooklyn Bridge cross?

EAT identifier
NE identifier

DBpedia SPARQL endpoint

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012.
QAKIS architecture

QA system

Query Generator

- Question file parser
- Query generator
- Typed Question generator
- Property identifier
- Query selector

Pattern Matching

- Pattern matcher
- Pattern Base

Which river does the Brooklyn Bridge cross?

DBpedia

SPARQL endpoint

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 13
Typed question generation

• **QAKiS target:** questions containing a NE related to the answer through **one property** of the ontology, i.e. requiring only one pattern to be matched

• **Typed question:**

   \[
   \text{Who is the husband of Amanda Palmer?}
   \]

   \[
   \begin{align*}
   \text{EAT: } & \text{[Person] or [Organisation] ([owl:Thing])} \\
   \text{NE: } & \text{Amanda Palmer [MusicalArtist] ([Artist], [owl:Thing])} \\
   \Rightarrow & \text{e.g. [Person] is the husband of [MusicalArtist]?}
   \end{align*}
   \]
WikiFramework pattern matching

- Tokens, lemmas and stems are extracted from typed questions (stopwords filtered)
- Word Overlap algorithm for typed question-pattern matching
- The relations represented by the patterns with the highest similarity score are selected
- 5-best relations sorted by decreasing matching score
QAKIS architecture

QA system

Query Generator

- Question file parser
- Query generator
- Typed Question generator
- EAT identifier
- NE identifier
- Property identifier
- Query selector

Pattern Matching

- Pattern matcher
- Pattern Base

DBpedia SPARQL endpoint

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 16
SPARQL queries generation

• For each relation, SPARQL queries are generated (according to the compatibility between NE/ EAT types and property’s domain and range):
  - select ?s where{?s <property> <NE>}
  - select ?s where{<NE> <property> ?s}

• queries sent to BDpedia SPARQL endpoint
• if no result is obtained, the next best pattern is considered
QAKiS demo

http://dbpedia.inria.fr/qakis/
Participation to the QALD-2 challenge

• Open challenge in Question Answering over Linked Data (@ESWC 2012, Interacting with Linked Data Workshop)
• Datasets (100 questions for training and 100 for test):
  ■ DBpedia 3.7
  ■ MusicBrainz
• SPARQL endpoint (for both datasets)
• Evaluation: Precision, Recall, F-measure
• The following attributes are specified for each question:
  ■ answertype: resource, string, number, date, boolean
  ■ aggregation (e.g., counting, filters, ordering)
  ■ onlydbo, i.e. only for DBpedia questions
# Systems results on QALD-2 data

<table>
<thead>
<tr>
<th></th>
<th>SemSek</th>
<th>Alexandria</th>
<th>MHE</th>
<th>QAKiS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>0.44</td>
<td>0.43</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.48</td>
<td>0.46</td>
<td>0.4</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>0.46</td>
<td>0.45</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td># answered questions</td>
<td>80/100</td>
<td>25/100</td>
<td>97/100</td>
<td>35/100</td>
</tr>
<tr>
<td># right answers</td>
<td>32</td>
<td>5</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td># partially right answers</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

## QAKiS on DBpedia

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>0.476</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.479</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>0.477</td>
<td>0.38</td>
</tr>
<tr>
<td># answered questions</td>
<td>40/100</td>
<td>35/100</td>
</tr>
<tr>
<td># right answers</td>
<td>17/40</td>
<td>11/35</td>
</tr>
<tr>
<td># partially right answers</td>
<td>4/40</td>
<td>4/35</td>
</tr>
</tbody>
</table>
### Systems results on QALD-2 data

<table>
<thead>
<tr>
<th></th>
<th>DBpedia</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SemSek</td>
<td>Alexandria</td>
<td>MHE</td>
<td>QAKiS</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.44</td>
<td>0.43</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.48</td>
<td>0.46</td>
<td>0.4</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>0.46</td>
<td>0.45</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td><strong># answered questions</strong></td>
<td>80/100</td>
<td>25/100</td>
<td>97/100</td>
<td>35/100</td>
</tr>
<tr>
<td><strong># right answers</strong></td>
<td>32</td>
<td>5</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td><strong># partially right answers</strong></td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>QAKiS on DBpedia</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set</td>
<td>Test set</td>
<td></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.476</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.479</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>0.477</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td><strong># answered questions</strong></td>
<td>40/100</td>
<td>35/100</td>
<td></td>
</tr>
<tr>
<td><strong># right answers</strong></td>
<td>17/40</td>
<td>11/35</td>
<td></td>
</tr>
<tr>
<td><strong># partially right answers</strong></td>
<td>4/40</td>
<td>4/35</td>
<td></td>
</tr>
</tbody>
</table>
Error analysis

- **Pattern matching:**
  - introduce more sophisticated algorithms than WO (Textual Entailment?)
  - cluster of relations with several patterns in common

  *Who is the owner of Universal Studios?* ⇒ owner
  *Who owns Aldi?* ⇒ keyPerson

- **Partially correct answers:** only one relation

  *Give me all people that were born in Vienna and died in Berlin.* ⇒ deathPlace ✓ X
  *Who is the daughter of Bill Clinton married to?* ⇒ spouse X
  (but Bill Clinton’s!)
Future work

- **Short-term** improvements on:
  - the WikiFramework pattern extraction algorithm
  - NE detection strategies
  - TE approaches investigation
  - boolean questions

- **Long-term** improvements on:
  - n-relation questions
  - natural language answer generation
NLP and Argumentation Theory for supporting online debates interaction

Possible application scenarios:

- verification of semantic editing in Wikipedia, and notification to the user
- community manager
- decision making

(Schneider et al. 2011)
First step: combining NLP and AT

Debate issue:
The use of cell-phones while driving is a public hazard.

Research shows that drivers speaking on a mobile phone have much slower reactions in braking tests than non-users...

Calling to say you will be late can reduce stress and make you less inclined to drive aggressively...

Argument A1 is rejected.
Argument A11 is accepted.

Decision Making

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 25
### Experimental setting: data set

<table>
<thead>
<tr>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic</strong></td>
<td><strong>#arg</strong></td>
</tr>
<tr>
<td>Violent games/aggressiveness</td>
<td>16</td>
</tr>
<tr>
<td>China one-child policy</td>
<td>11</td>
</tr>
<tr>
<td>Consider coca as a narcotic</td>
<td>15</td>
</tr>
<tr>
<td>Child beauty contests</td>
<td>12</td>
</tr>
<tr>
<td>Arming Libyan rebels</td>
<td>10</td>
</tr>
<tr>
<td>Random alcohol breath tests</td>
<td>8</td>
</tr>
<tr>
<td>Osama death photo</td>
<td>11</td>
</tr>
<tr>
<td>Privatizing social security</td>
<td>11</td>
</tr>
<tr>
<td>Internet access as a right</td>
<td>15</td>
</tr>
<tr>
<td>Vegetarianism</td>
<td>7</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>109</td>
</tr>
</tbody>
</table>

- **Debatepedia as case study** provides us with annotated arguments and casts our task as a yes/no entailment task
- **Test set pairs concern completely new topics**

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 26
Experimental set: evaluation

- **EDITS system** (Edit Distance Textual Entailment Suite) ([Kouylekov and Negri, 2010]), off-the-shelf system
  - Basic configuration: **cosine similarity** algorithm; distance calculated on lemmas; stopword list included

- **FIRST STEP: TEXTUAL ENTAILMENT**

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rel</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pr.</td>
<td>Rec.</td>
</tr>
<tr>
<td>EDITS</td>
<td>yes</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.66</td>
</tr>
<tr>
<td>WordOverl.</td>
<td>yes</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.56</td>
</tr>
</tbody>
</table>

- **SECOND STEP: TE+ARGUMENTATION THEORY**
  - Pr: 0.74, Rec: 0.76, Acc: 0.75
Experimental set: evaluation

- **EDITS system** *(Edit Distance Textual Entailment Suite)* *(Kouylekov and Negri, 2010)*, off-the-shelf system
  - Basic configuration: cosine similarity algorithm; distance calculated on lemmas; stopword list included

- **FIRST STEP: TEXTUAL ENTAILMENT**

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th></th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDITS</td>
<td>yes</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>WordOverl.</td>
<td>yes</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>0.56</td>
<td>0.55</td>
</tr>
</tbody>
</table>

- **SECOND STEP: TE+ARGUMENTATION THEORY**
  
  Pr: 0.74, Rec: 0.76, Acc.: 0.75
Research shows that drivers speaking on a mobile phone have much slower reactions in braking tests than non-users ...

Calling to say you will be late can reduce stress and make you less inclined to drive aggressively

Debate issue:
The use of cell-phones while driving is a public hazard.

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 29
Conclusions

- Two different ongoing research directions, same goal: **support users in querying and exploring the Semantic Web**

- **Joint works with:** Bernardo Magnini, Alberto Lavelli, Alessio Palmero Aprosio (*FBK, Trento, Italy*); Julien Cojan, Serena Villata, Fabien Gandon (*INRIA Sophia Antipolis*)
Thanks for your attention!

elena.cabrio@inria.fr

http://www-sop.inria.fr/members/Elena.Cabrio/

http://dbpedia.inria.fr/qakis/

E. Cabrio, Bridging NLP and Semantic Web to enhance user interactions with the Web of Data July 19, 2012. 31