Topic 11: Wikipedia: Surface form extraction (HARD!)

Philipp Heinisch
Ricardo Usbeck

final presentation

February 1, 2018

Table of Contents

1. Task
   - Task description
   - Theory (basics)
2. My approach
3. Evaluation
   - Evaluation results
   - Discussion of evaluation results
4. Code

Knowledge extraction from text

Blaise Pascal (19 June 1623 to 19 August 1662) was a French mathematician, physicist, inventor, writer and Catholic theologian[...]

Pascal was an important mathematician[...]

Belongs *Blaise Pascal* and/ or *Pascal* to the URI `http://dbpedia.org/resource/Blaise_Pascal`?

Knowledge extraction from text

Blaise Pascal (19 June 1623 to 19 August 1662) was a French mathematician, physicist, inventor, writer and Catholic theologian[...]

Pascal was an important mathematician[...]

Belongs *Blaise Pascal* and/ or *Pascal* to the URI `http://dbpedia.org/resource/Blaise_Pascal`?

Task: Knowledge bases contain labels for resources (`rdfs:label`). [...] The goal of this task is to detect candidates for labels using Wikipedia. [...]

Knowledge extraction from text

> *Blaise Pascal (19 June 1623 to 19 August 1662) was a French mathematician, physicist, inventor, writer and Catholic theologian[...]*
>
> *Pascal was an important mathematician[...]*

Belongs *Blaise Pascal* and/ or *Pascal* to the URI `http://dbpedia.org/resource/Blaise_Pascal`?

Task: Knowledge bases contain labels for resources (`rdfs:label`). [...] The goal of this task is to detect candidates for labels using Wikipedia. [...]

Solve this task with the KATANA algorithm

- $\exists$ list of candidates (+ knowledge about these)
- $\exists$ list of extracted labels with extracted knowledge from text

KATANA!

Match the labels to the candidates = calculate the score for each candidate-label $\Rightarrow$ the highest score wins!

Solve this task with the KATANA algorithm

- ∃ list of candidates (+ knowledge about these)
- ∃ list of extracted labels with extracted knowledge from text

KATANA!

Match the labels to the candidates = calculate the score for each candidate-label ⇒ the highest score wins!

How good is the KATANA algorithm?

Solve this task with the KATANA algorithm

- $\exists$ list of candidates (+ knowledge about these)
- $\exists$ list of extracted labels with extracted knowledge from text

KATANA!

Match the labels to the candidates = calculate the score for each candidate-label $\Rightarrow$ the highest score wins!

How good is the KATANA algorithm?

KATANA Algorithm/ Formulas

Given

Knowledge base KB (s, p, o), our extracted triples *ext* from natural text with labels, find out the matching URI-candidate c_s from $\{c_1, ..., c_n\}$

Determine the ambiguity of a fact to a given subject s

$$\psi(p, o) = 1 - \frac{1}{|\{s|(s,p,o) \in KB\}|}$$

KATANA Algorithm/ Formulas

Given

Knowledge base KB (s, p, o), our extracted triples ext from natural text with labels, find out the matching URI-candidate c_s from $\{c_1, ..., c_n\}$

Determine the ambiguity of a fact to a given subject s

$$\psi(p, o) = 1 - \frac{1}{|\{s|(s,p,o) \in KB\}|}$$

Determine the score (matching grade) candidate → label $(\rightarrow \lambda)$

for a certain candidate c:
$M(c, s) = \{(p, o)|(\lambda, p, o) \in ext \cap (c, p, o) \in KB\}$

$$score(c, s) = \begin{cases} 0 & M(c, s) = \emptyset \\ 1 - \prod_{(p,o) \in M(c,s)} \psi(p, o) & M(c, s) \neq \emptyset \end{cases}$$

KATANA Algorithm/ Formulas

Given

Knowledge base KB (s, p, o), our extracted triples ext from natural text with labels, find out the matching URI-candidate c_s from $\{c_1, ..., c_n\}$

Determine the ambiguity of a fact to a given subject s

$$\psi(p, o) = 1 - \frac{1}{|\{s|(s,p,o) \in KB\}|}$$

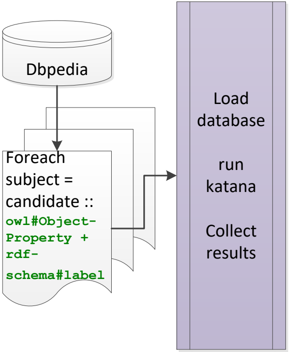
Determine the score (matching grade) candidate $\rightarrow$ label ($\rightarrow \lambda$)

for a certain candidate c:

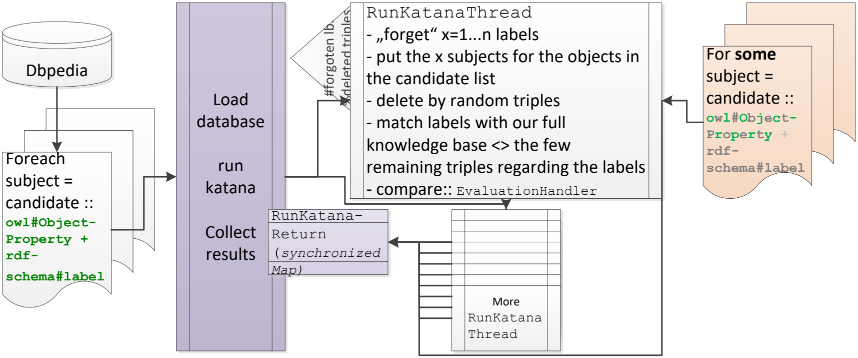
$M(c, s) = \{(p, o)|(\lambda, p, o) \in ext \cap (c, p, o) \in KB\}$

$$score(c, s) = \begin{cases} 0 & M(c, s) = \emptyset \\ 1 - \prod_{(p,o) \in M(c,s)} \psi(p, o) & M(c, s) \neq \emptyset \end{cases}$$

My application


Dbpedia
Foreach subject = candidate :: owl#Object-Property + rdf-schema#label
Load database
run katana
Collect results

My application



Dbpedia

Foreach
subject =
candidate ::
**owl#Object-
Property +
rdf-
schema#label**

Load
database

run
katana

Collect
results

RunKatana-
Return
(*synchronized
Map*)

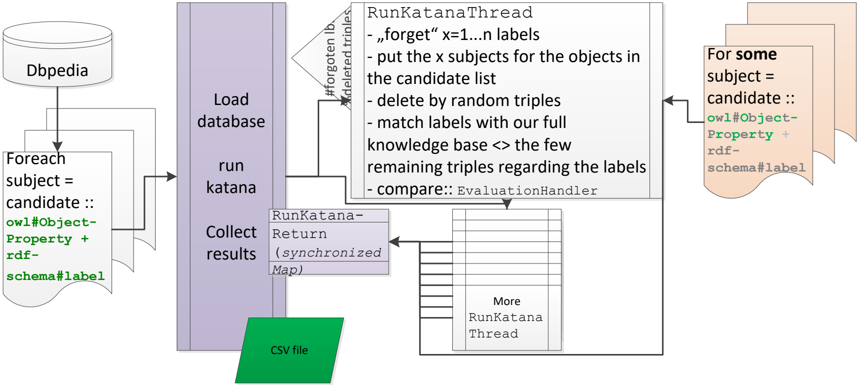
#forgoten lb.
deleted triples

RunKatanaThread
- „forget“ x=1…n labels
- put the x subjects for the objects in the candidate list
- delete by random triples
- match labels with our full knowledge base <> the few remaining triples regarding the labels
- compare:: EvaluationHandler

For **some**
subject =
candidate ::
**owl#Object-
Property +
rdf-
schema#label**

More
RunKatana
Thread

My application



Dbpedia

Foreach subject = candidate :: owl#Object-Property + rdf-schema#label

Load database

run katana

Collect results

RunKatana-Return (*synchronized Map*)

CSV file

RunKatanaThread
- „forget“ x=1...n labels
- put the x subjects for the objects in the candidate list
- delete by random triples
- match labels with our full knowledge base <> the few remaining triples regarding the labels
- compare:: EvaluationHandler

More RunKatana Thread

For **some** subject = candidate :: owl#Object-Property + rdf-schema#label

My application follows the command pattern

Available commands without their parameters

- Environment commands: `help`, `exit`
- Database commands: `load`, `edit`, `print`
- KATANA (evaluation) commands: `katana`

Demo in the end

... if there is time...

My application follows the command pattern

Available commands without their parameters

- Environment commands: `help`, `exit`
- Database commands: `load`, `edit`, `print`
- KATANA (evaluation) commands: `katana`

Demo in the end

... if there is time...

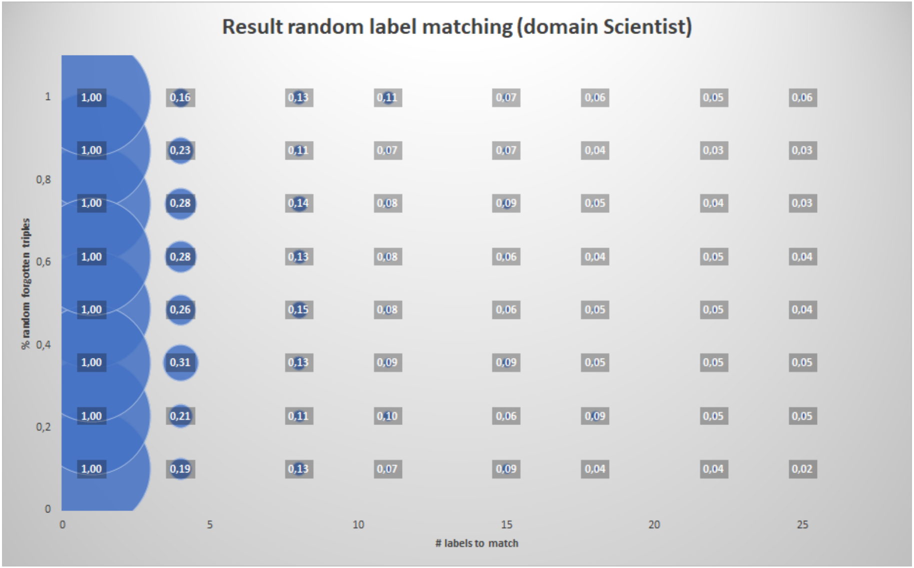
Evaluation

Result measurement function (`EvaluationHandler::calculateAccuracy()`)

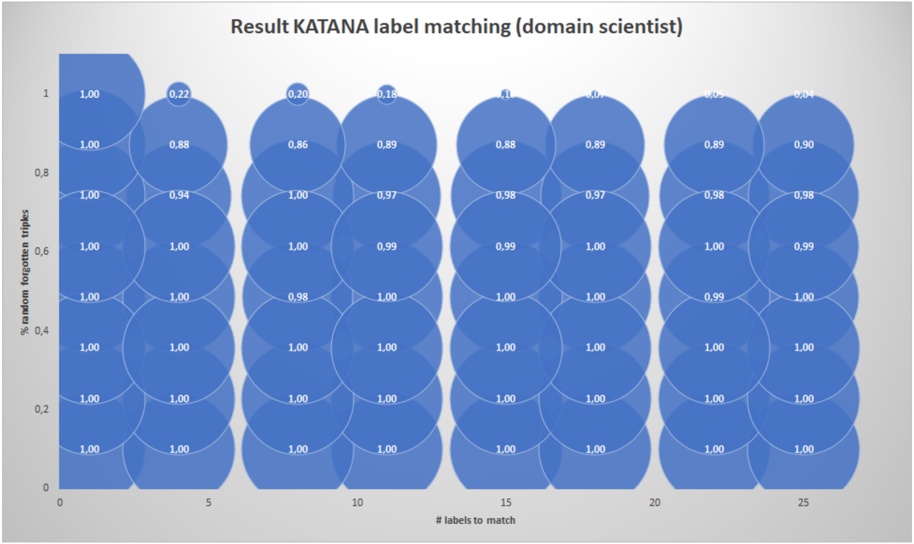
L is set of guessed labels, $L' \subseteq L$ set of wrong guessed labels, C set of correct labels and $C' \subseteq C$ set of labels that doesn't appear in L:

Size of bubbles $s = \frac{2 - \frac{|L'|}{|L|} - \frac{|C'|}{|C|}}{2}$, $0 \leq s \leq 1$

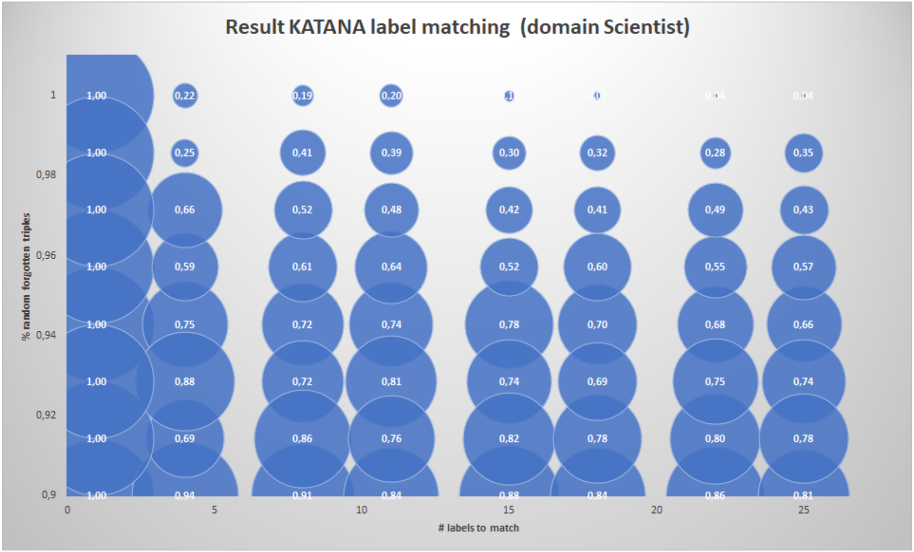
Evaluation


Result random label matching (domain Scientist)
% random forgotten triples
1
0,8
0,6
0,4
0,2
0
1,00
0,16
0,13
0,11
0,07
0,06
0,05
0,06
1,00
0,23
0,11
0,07
0,07
0,04
0,03
0,03
1,00
0,28
0,14
0,08
0,09
0,05
0,04
0,03
1,00
0,28
0,13
0,08
0,06
0,04
0,05
0,04
1,00
0,26
0,15
0,08
0,06
0,05
0,05
0,04
1,00
0,31
0,13
0,09
0,09
0,05
0,05
0,05
1,00
0,21
0,11
0,10
0,06
0,09
0,05
0,05
1,00
0,19
0,13
0,07
0,09
0,04
0,04
0,02
0
5
10
15
20
25
# labels to match

Evaluation


Result KATANA label matching (domain scientist)
% random forgotten triples
1
0,8
0,6
0,4
0,2
0
0
5
10
15
20
25
# labels to match
1,00
0,22
0,20
0,18
0,10
0,07
0,05
0,04
1,00
0,88
0,86
0,89
0,88
0,89
0,89
0,90
1,00
0,94
1,00
0,97
0,98
0,97
0,98
0,98
1,00
1,00
1,00
0,99
0,99
1,00
1,00
0,99
1,00
1,00
0,98
1,00
1,00
1,00
0,99
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00
1,00

Evaluation



Result KATANA label matching (domain Scientist)
% random forgotten triples
1
0,98
0,96
0,94
0,92
0,9
1,00 0,22 0,19 0,20 0,10 0,07 0,04 0,04
1,00 0,25 0,41 0,39 0,30 0,32 0,28 0,35
1,00 0,66 0,52 0,48 0,42 0,41 0,49 0,43
1,00 0,59 0,61 0,64 0,52 0,60 0,55 0,57
1,00 0,75 0,72 0,74 0,78 0,70 0,68 0,66
1,00 0,88 0,72 0,81 0,74 0,69 0,75 0,74
1,00 0,69 0,86 0,76 0,82 0,78 0,80 0,78
1,00 0,94 0,91 0,84 0,88 0,84 0,86 0,81
0
5
10
15
20
25
# labels to match

Discussion of evaluation results

1. KATANA leads to much more better results than the random matching
2. KATANA is very precise ($\geq 80\%$) until 90% data-loss [in a big data set, too]
   - lots of properties are nearly unambiguously, e.g.: birth date, spouse, (wikiPageExternalLink), ...
   - to match 1 label in the real world, you would have ≥ 1 candidates (selective range)

Link to repository (Code)

https://github.com/dice-group/KATANA
branch: Philipp_Heinisch_master