Can we use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expressions?

Prof. Michel Gagnon, Ecole Polytechnique de Montréal, Canada
Prof. Amal Zouaq, Royal Military College of Canada, Canada
Dr. Ludovic Jean-louis, Ecole Polytechnique de Montréal, Canada

Introduction

- What is Semantic Annotation on the LOD?
  - Candidate Selection
  - Disambiguation
- Traditionally focused on named entity annotation

Semantic Annotators

- LOD-based Semantic Annotators
  - Most used annotators
  - REST API / Web Services
  - Default configuration

DBPedia Spotlight

Zemanta

Yahoo

Lupedia

Alchemy (entities + keywords)

Wikimeta

OpenCalais

Annotation Evaluation

- Few comparative studies have been published on the performance of linked data Semantic Annotators
- The available studies generally focus on “traditional” named entity annotation evaluation (ORG, Person, etc.)

NEs versus domain topics

- Semantic Annotators start to go beyond these traditional NEs
  - Predefined classes such as PERSON, ORG
  - Expansion of possible classes (NERD Taxonomy : sport events, operating systems, political events, …)
  - RDF (NE? Domain topic ?)
  - Semantic Web (NE? Domain topic ?)
- Domain Topics: important “concepts” for the domain of interest

NE versus Topics Annotation

Annotator	Detects	Pos. in text
Alchemy	NE/Top	Only for NE
Spotlight	NE/Top	yes
Wikimeta	Mainly NE	yes
Lupedia	NE	yes
Open Calais	NE/Top	yes
Yahoo	NE/Top	yes
Zemanta	Top	no

Table 1: Characteristics of semantic annotators (NE = named entites, Top = Topics)

Research Question 1

- Previous studies did not draw any conclusion about the relevance of the annotated entities
- But relevance might prove to be a crucial point for tasks such as information seeking or query answering

- RQ1: Can we use linked data-based semantic annotators to accurately identify Domain-relevant expressions?

Research Methodology

- We relied on a corpus of 8 texts taken from Wikipedia, all related to the artificial intelligence domain. Together, these texts represent 10570 words.
- We obtained 2023 expression occurrences among which 1151 were distinct.

Building the Gold Standard

- We asked a human evaluator to analyze each expression and make the following decisions:
  - Does the annotated expression represent an **understandable** named entity or topic?
  - Is the expression a **relevant keyword** according to the document?
  - Is the expression a **relevant named entity** according to the document?

Distribution of Expressions

Type of expression	Quantity	Nb. Relevant
Total detected	1151	
Total understandable	639	507
Topics	516	477
Named entities	123	30

Table 2: Distribution of expressions in our Gold Standard. We consider only distinct expressions.

Note that only 639 out of 1151 detected expressions are understandable (56%). Thus a substantial number of detected expressions represents noise.

Frequency of detected Expressions

N	Number of expressions
1	872
2	183
3	57
4	24
5	8
6	5
7	2
Total	1151

(326 relevant)
(117 relevant)
(40 relevant)
(15 relevant)
(4 relevant)
(4 relevant)
(1 relevant)

- Most of the expressions were detected by only one (76%) or two (16%) annotators.
- Semantic annotators seemed complementary.

Evaluation

- Standard precision, recall and F-score
- “Partial” recall, since our Gold Standard considers only the expressions detected by at least one annotator, instead of considering all possible expressions in the corpus.

All Expressions

All expressions						
Method	Det	Top	NE	Und.	Rel	P/R/F
Alchemy	619	564	55	428	347	0.56/0.69/0.62
Spotlight	356	338	18	149	108	0.3/0.21/0.25
Wikimeta	243	149	94	130	96	0.39/0.19/0.25
Lupedia	42	21	21	28	12	0.27/0.024/0.043
Open Calais	147	107	40	127	91	0.62/0.18/0.28
Yahoo	105	93	12	85	74	0.7/0.15/0.24
Zemanta	77	69	8	69	61	0.77/0.12/0.21

Named Entities

Only named entities						
Alchemy	55	0	55	50	18	0.37/0.62/$\boxed{0.44}$
Spotlight	18	0	18	17	8	0.44/0.27/0.32
Wikimeta	94	0	94	51	24	0.25/$\boxed{0.78}$/0.38
Lupedia	21	0	21	19	6	0.31/0.22/0.24
Open Calais	40	0	40	39	12	0.33/0.39/0.35
Yahoo	12	0	12	11	5	0.28/0.15/0.2
Zemanta	8	0	8	8	5	$\boxed{0.47}$/0.17/0.25

Domain Topics

Only topics						
Alchemy	564	564	0	378	329	0.59/0.69/0.63
Spotlight	338	338	0	132	100	0.3/0.21/0.25
Wikimeta	149	149	0	79	72	0.48/0.15/0.23
Lupedia	21	21	0	9	6	0.3/0.012/0.023
Open Calais	107	107	0	88	79	0.75/0.17/0.27
Yahoo	93	93	0	74	69	0.74/0.15/0.24
Zemanta	69	69	0	61	56	0.81/0.12/0.21

Few Highlights

- F-score is unsatisfactory for almost all annotators
  - Best F-score around 62-63 (All expressions, topics)
- Alchemy seems to stand out
  - It obtains a high value for recall when all expressions or only topics are considered
  - But the precision of Alchemy (59%) is not very high, compared to the results of Zemanta (81%), OpenCalais (75%) and Yahoo (74%)
- None of the annotators was good at detecting relevant named entities.
- DBpedia Spotlight annotated many expressions that are not considered understandable

Research Question 2

- Given the complementarity of Semantic Annotators, we decided to experiment with this second research question:

  Can we improve the overall performance of semantic annotators using voting methods and machine learning?

Voting and machine learning methods

- We experimented with the following methods:
  - Simple votes
  - Weighted votes where the weight is the precision of individual annotators on a training corpus
  - K-nearest-neighbours classifier
  - Naïve Bayes classifier
  - Decision tree (C4.5)
  - Rule induction (PART algorithm)

Experiment Setup

- We divided the set of annotated expressions into 5 balanced partitions
  - One partition for testing
  - Remaining partitions for training
- We ran 5 experiments for each of the following situations: all entities, only named entities and only topics

Baseline: Simple Votes

Threshold: Number of annotators

Threshold	P	R	F
All expressions considered			
1	0.44	1.0	0.61
2	0.65	0.36	0.46
3	0.67	0.13	0.21
4	0.6	0.047	0.086
5	0.55	0.018	0.034
6	0.4	0.0098	0.019
7	0.1	0.0019	0.0036
Only named entities considered			
1	0.24	1.0	0.39
2	0.38	0.62	0.47
3	0.45	0.5	0.46
4	0.34	0.26	0.29
5	0.3	0.095	0.14
6	0.4	0.095	0.15
7	0.2	0.029	0.05
Only topics considered			
1	0.46	1.0	0.63
2	0.72	0.34	0.46
3	0.83	0.1	0.18
4	0.9	0.036	0.069
5	0.8	0.012	0.024
6	0.2	0.0043	0.0083
7	0.0	0.0	0.0

All Expressions:
Best prec : 77% (Zemanta)

Best rec/F : 69% - 62% (Alchemy)

Best P: 81% (R/F 12%-21%) (Zemanta)

Best R/F: 69%-63% (Alchemy) P=59%

Weighted Votes

Threshold	P	R	F
Only topics considered			
0.0	0.46	1.0	0.63
0.05	0.46	1.0	0.63
0.1	0.56	0.91	0.69
0.15	0.66	0.53	0.56
0.2	0.7	0.37	0.48
0.25	0.78	0.29	0.42
0.3	0.8	0.23	0.35
0.35	0.88	0.12	0.21

Best P: 81% (Zemanta) (R/F 12%-21%)

Best R/F: 69%-63% (Alchemy) P=59%

Machine Learning

Best P: 81% (Zemanta)
(R/F 12%-21%)

Best R/F: 69%-63%
(Alchemy) P=59%

Method	Average P/R/F	Min P/R/F	Max P/R/F
Only topics considered			
Weighted vote (th = 0.15)	0.66/0.53/0.56	0.53/0.39/0.51	0.73/0.89/0.66
KNN (N=3)	0.59/0.73/0.64	0.52/0.43/0.52	0.66/0.84/0.7
Bayes	0.59/0.77/0.67	0.53/0.73/0.63	0.66/0.82/0.73
Dec. tree	0.59/0.83/0.69	0.54/0.73/0.66	0.64/0.9/0.75
PART	0.59/0.83/0.69	0.54/0.72/0.66	0.64/0.9/0.75

Highlights

- On the average, weighted vote displays the best precision (0.66) among combination methods if only topics are considered
  - Alchemy (0.59)
- Alchemy finds many relevant expressions that are not discovered by other annotators
- Machine learning methods achieve better recall and F-score on the average, especially if decision tree or rule induction (PART) is used

Conclusion

- Can we use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expressions?
  - Best overall performance (Recall/F-Score: Alchemy) (59%- 69%-63%)
  - Best precision: Zemanta (81%)
- Need of filtering!
  - Combination/machine learning methods
  - Increased recall with weighted voting scheme

Future Work

- Limitations
  - The size of the gold standard
  - A limited set of semantic annotators
- Increase the size of the GS
- Explore various combinations for annotators/features
- Compare Semantic Annotators with keyword extractors

QUESTIONS?

- Thank you for your attention!
- amal.zouaq@rmc.ca
- http://azouaq.athabascau.ca