EXTRACTING LINKED HYPERNYMS FROM FREE TEXT OF WIKIPEDIA ARTICLES

COMBINING MACHINE LEARNING WITH LEXICO-SYNTACTIC RULES

TOMÁŠ KLIEGR, ONDŘEJ ZAMAZAL, VÁCLAV ZEMAN

DEPARTMENT OF INFORMATION AND KNOWLEDGE ENGINEERING
FACULTY OF INFORMATICS AND STATISTICS
UNIVERSITY OF ECONOMICS PRAGUE, CZECH REPUBLIC

KEG
Dec 4, 2019

Selected projects in encyclopaedic linked data

DBpedia type extraction

en.wikipedia.org/wiki/Karel_%C4%8Capek

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Karel Čapek

From Wikipedia, the free encyclopedia

Karel Čapek (Czech: [ˈkarɛl ˈtʃapɛk] (listen); 9 January 1890 – 25 December 1938) was a Czech writer of the early 20th century. He had multiple roles throughout his career, including playwright, dramatist, essayist, publisher, literary reviewer, photographer and art critic. Nonetheless, he is best known for his science fiction including his novel *War with the Newts* and the play *R.U.R.* (*Rossum's Universal Robots*), which introduced the word *robot*.[1][2] He also wrote many politically charged works dealing with the social turmoil of his time. Largely influenced by American pragmatic liberalism,[3] he campaigned in favor of free expression and utterly despised the rise of both fascism and communism in Europe.[4][5]

Čapek was nominated for the Nobel Prize in Literature seven times,[6] but he never won. However, several awards are named after him,[7][8] such as the Karel Čapek Prize, which is awarded every other year by Czech PEN Club for literary work that contributes to reinforcing or maintaining democratic and humanist values in the society.[9] He was also a key figure in the creation of the Czechoslovak PEN Club as a part of the International PEN.[10] He died on the brink of World War II as a result of lifelong medical condition,[11] but his legacy as a literary figure has been well established after the war.[4]

Contents [hide]

Karel Čapek

Born	9 January 1890 Malé Svatoňovice, Austria-Hungary (today Czech Republic)
Died	25 December 1938 (aged 48) Prague, Czechoslovakia
Pen name	K. Č., B. Č.
Occupation	Novelist, dramatist, journalist
Nationality	Czech
Alma mater	Charles University in Prague

Infobox

Our approach to type extraction

en.wikipedia.org/wiki/Karel_%C4%8Capek

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Karel Čapek

From Wikipedia, the free encyclopedia

Karel Čapek (Czech: [ˈkarɛl ˈtʃapɛk] (listen); 9 January 1890 – 25 December 1938) was a Czech writer of the early 20th century. He had multiple roles throughout his career, including playwright, dramatist, essayist, publisher, literary reviewer, photographer and art critic. Nonetheless, he is best known for his science fiction including his novel *War with the Newts* and the play *R.U.R.* (*Rossum's Universal Robots*), which introduced the word *robot*.[1][2] He also wrote many politically charged works dealing with the social turmoil of his time. Largely influenced by American pragmatic liberalism,[3] he campaigned in favor of free expression and utterly despised the rise of both fascism and communism in Europe.[4][5]

Čapek was nominated for the Nobel Prize in Literature seven times,[6] but he never won. However, several awards are named after him,[7][8] such as the Karel Čapek Prize, which is awarded every other year by Czech PEN Club for literary work that contributes to reinforcing or maintaining democratic and humanist values in the society.[9] He was also a key figure in the creation of the Czechoslovak PEN Club as a part of the International PEN.[10] He died on the brink of World War II as a result of lifelong medical condition,[11] but his legacy as a literary figure has been well established after the war.[4]

Free text

Contents [hide]

Karel Čapek

Born	9 January 1890 Malé Svatoňovice, Austria-Hungary (today Czech Republic)
Died	25 December 1938 (aged 48) Prague, Czechoslovakia
Pen name	K. Č., B. Č.
Occupation	Novelist, dramatist, journalist
Nationality	Czech
Alma mater	Charles University in Prague

Linked Hypernyms Dataset

Algorithms

- Hand-crafted lexico-syntactic patterns (JAPE grammar)
- Type co-occurrence analysis across knowledge graphs
- Hierarchical SVM

Objective

- Complete missing types in DBpedia
- Get more specific types than in DBpedia (or DBpedia ontology)

dataset	description	English	German	Dutch
Inference	2016-04 DBpedia release	3,8 million	1,1 million	1,1 million

Dataset size

Hearst patterns

- Input text: Wikipedia article
- Question: Who was Karel Čapek?

Karel Čapek was a Czech writer of the early 20th century. He made…

Karel [NNP] Čapek [NNP] was VBN a Czech JJ writer NN, …

Karel Čapek was a Czech writer of the early 20th century. He made…

ANNIE ENGLISH TOKENIZER

↓

SENTENCE SPLITTER

↓

PART OF SPEECH TAGGER

↓

NOUN PHRASE EXTRACTION

↓

GRAMMAR INTERPRETER

Extraction grammar → Regular expressions over annotations

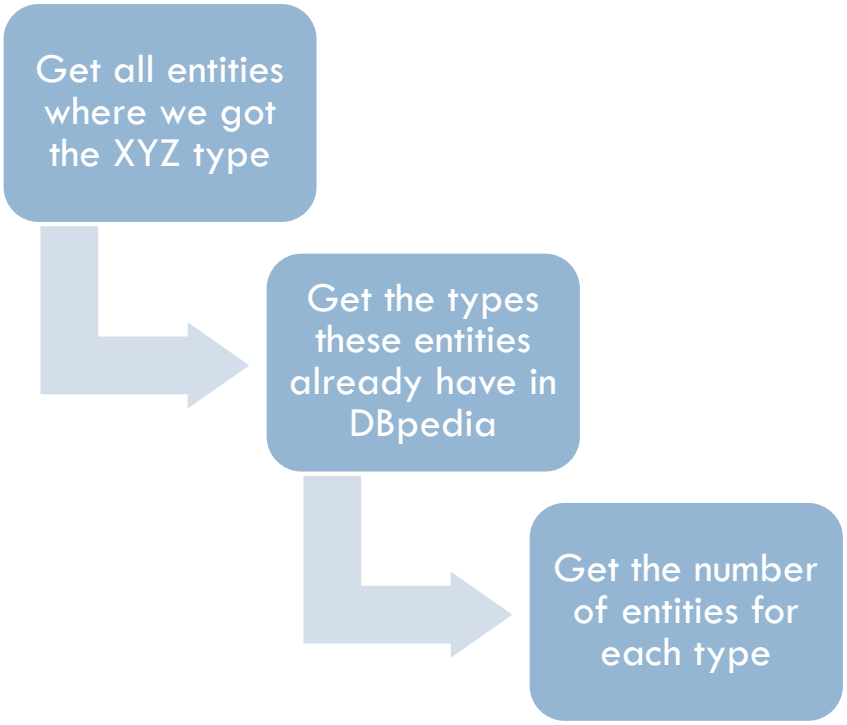
↓

Answer: writer

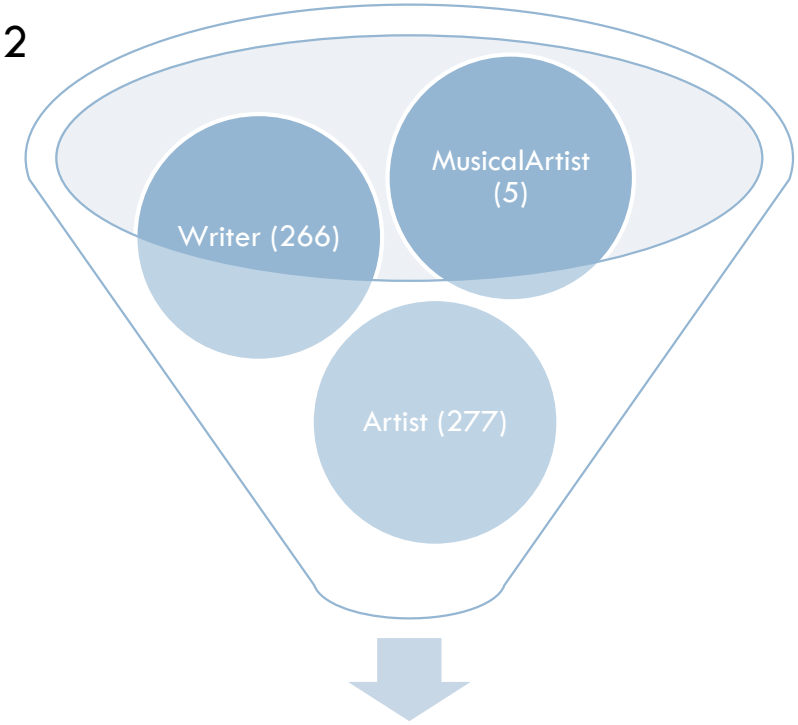
general architecture GATE for text engineering

… when the hypernym is a word not in DBpedia Ontology => Instance based ontology alignment

Step 1



Step 2



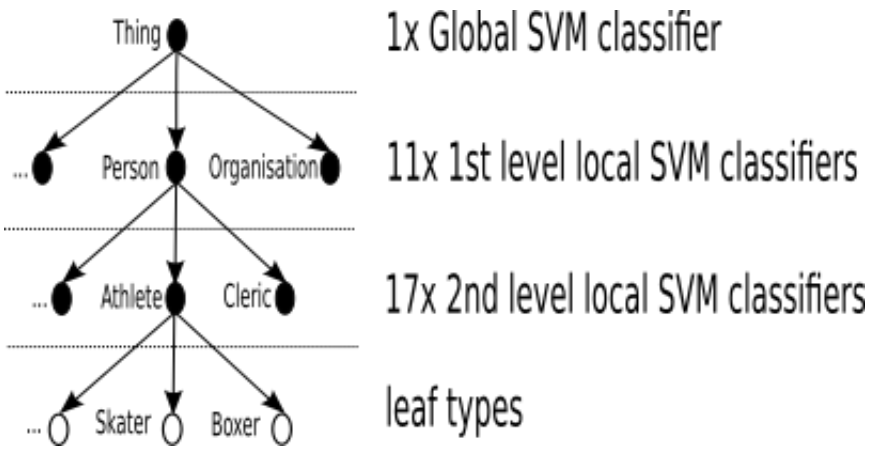
Type with best balance of specificity and support

Kliegr, Tomáš, and Ondřej Zamazal. "LHD 2.0: A text mining approach to typing entities in knowledge graphs." Web Semantics: Science, Services and Agents on the World Wide Web 39 (2016): 47-61.

Hierarchical SVMs

Vaclav Havel […] was a Czech playwright, essayist, poet, dissident and politician. …

Amnesty International prisoners of conscience held by CzechoslovakiaCancer survivors; Charter 77 signatories;



Short abstracts
Categories

↓

Bag of words : tokenization, lower casing

↓

Train local classifier for all concepts in DBpedia

↓

Apply classifiers & combine results

↓

Selection of type

Evaluation with crowdsourcing

- Randomly selected entities from Wikipedia were assigned types by at least three annotators

- Used annotator agreement to establish groundtruth
- Gold standard with 2000 entity type assignments

Evaluation metrics

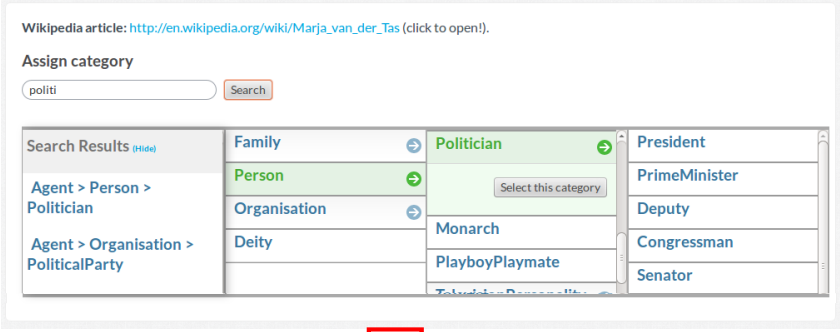
- Exact precision

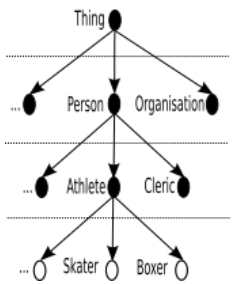
$$P_{exact} = \frac{\sum_i |P_i \cap T_i|}{\sum_i |P_i|},$$

- Hierarchical precision, recall and F-measure

$$hP = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{P}_i|}, \qquad hR = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{T}_i|}, \qquad hF = \frac{2 * hP * hR}{hP + hR},$$

where $\hat{P}_i$ is the set of the most specific type(s) predicted for test example i and all its (their) ancestor types and $\hat{T}_i$ is the set of the true most specific type(s) of test example i and all its (their) ancestor types.





1x Global SVM classifier

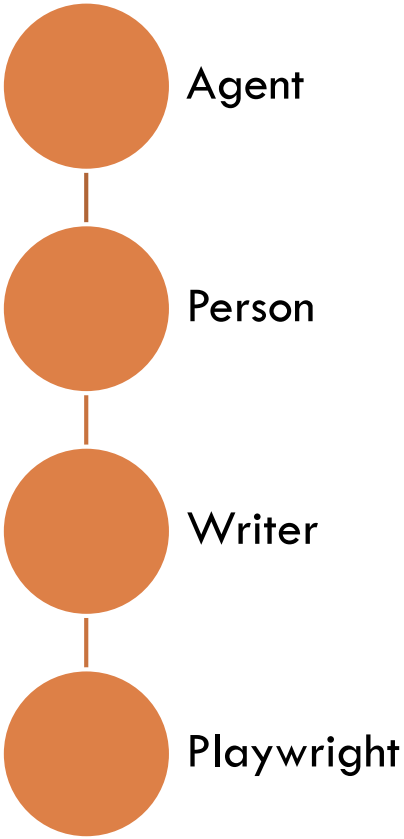
11x 1st level local SVM classifiers

17x 2nd level local SVM classifiers

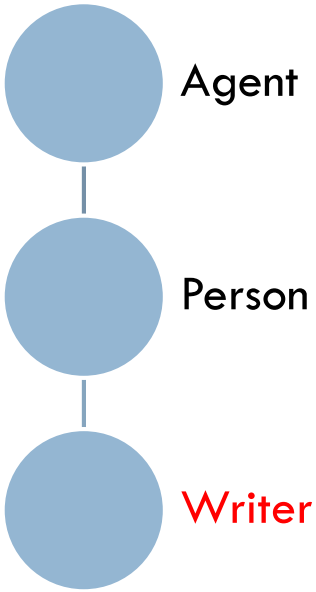
leaf types

general architecture for text engineering GATE

Extraction grammar



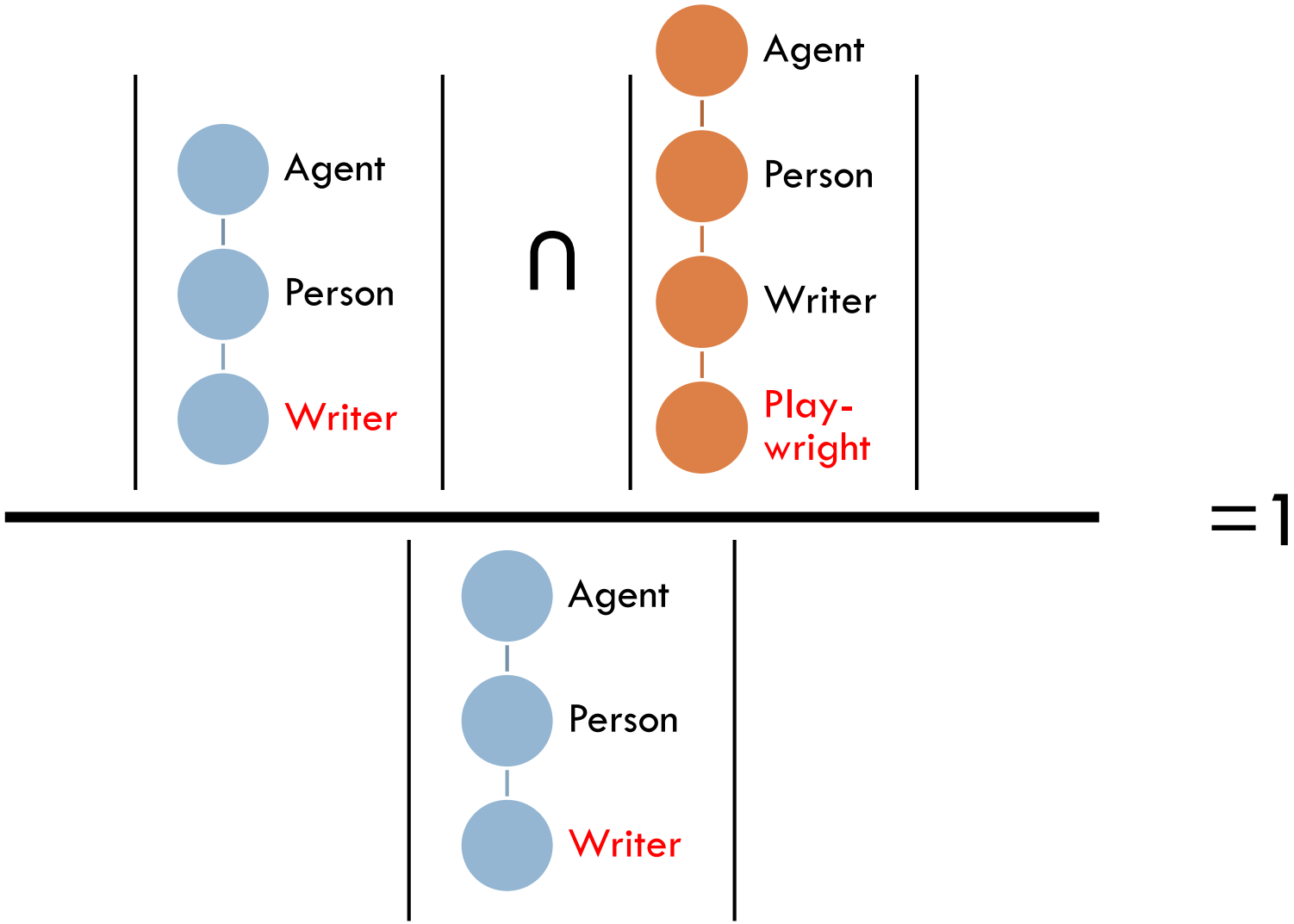
Gold standard



Type assignment by our algorithms

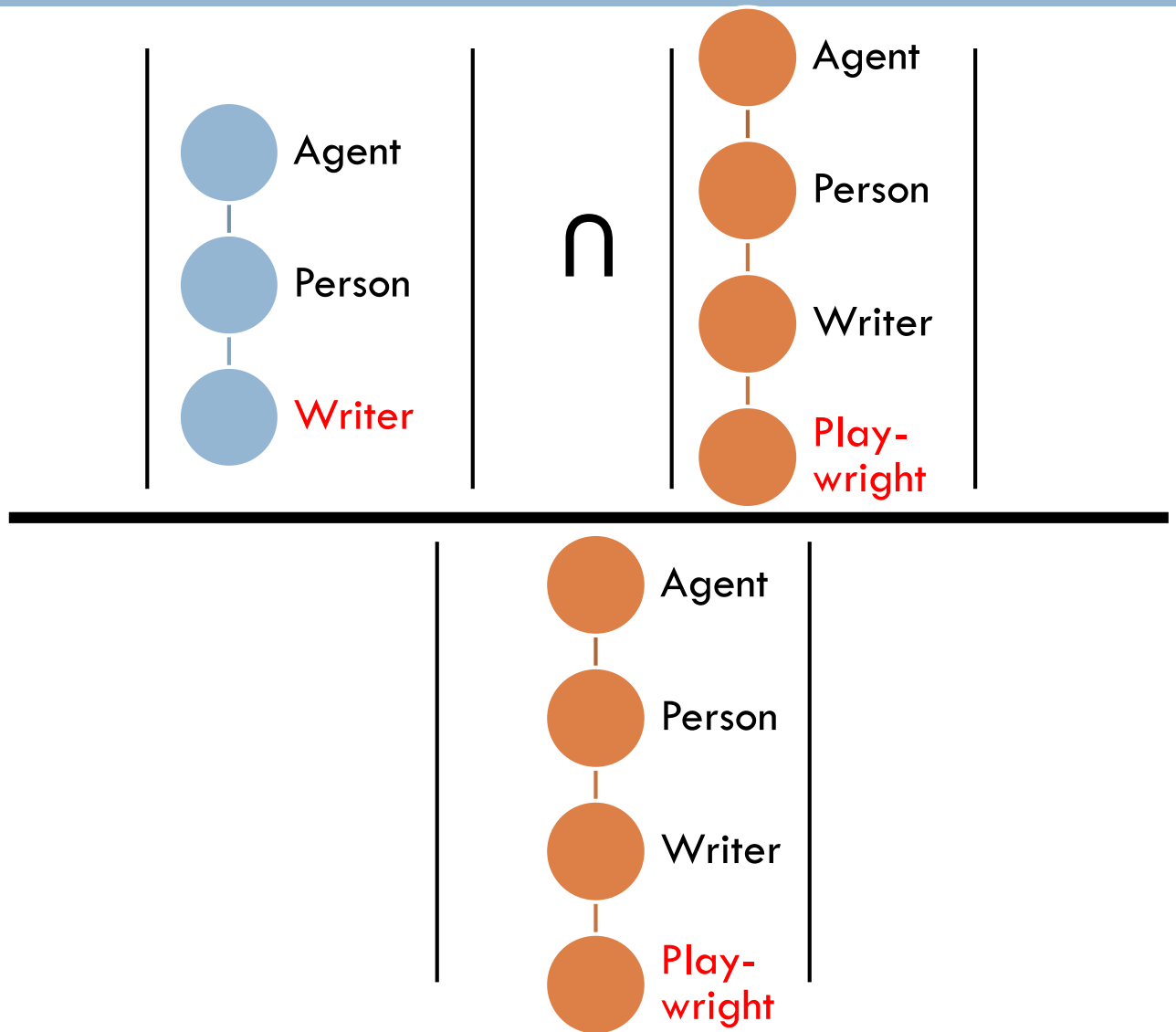
Hierarchical precision

$$hP = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{P}_i|}$$



Hierarchical recall

$$hR = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{T}_i|}$$



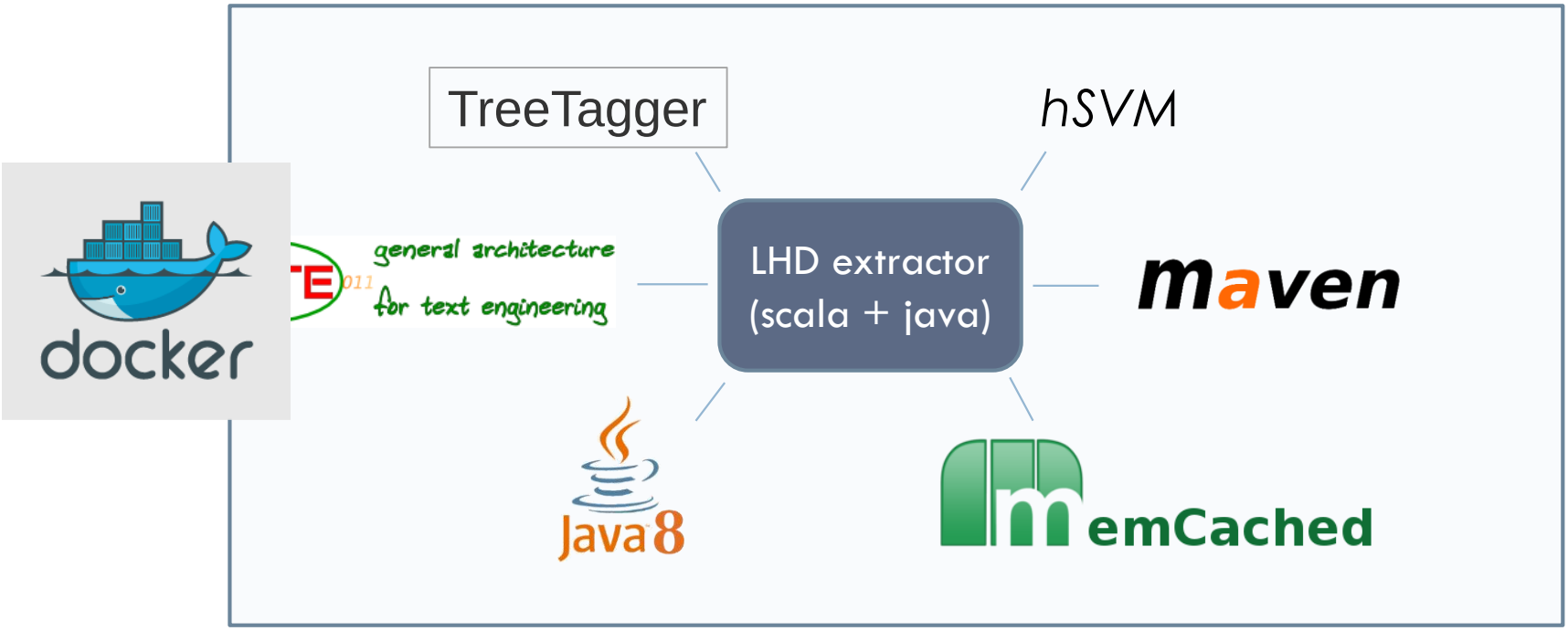
$=3/4$

Evaluation results

Classifier	GS3 (randomly drawn articles)				
	entities	P_{exact}	hP	hR	hF
DBpedia	715	.537	.902	.611	.729
SDType					
Core	402	.654	.864	.371	.519
STI_{prune}	379	.449	.754	.274	.403
$hSVM_{text}\alpha$	750	.307	.747	.597	.663
$hSVM_{text}STI\alpha$	765	.327	.757	.621	.682
$Core + STI_{prune}$	781	.554	.814	.645	.720
$Core + hSVM_{text}\ STI\alpha$	864	.439	.786	.720	.752
$Core + STI_{prune} + hSVM_{text}\alpha$	896	.465	.800	.724	.760

- LHD lexico-syntactic patterns match/exceed exact precision of DBpedia (infoboxes)
- LHD hSVM have lower precision, but higher recall than DBpedia

Dockerized LHD framework



Comparison with state-of-the-art

Paulheim, Heiko, and Christian Bizer. "Type inference on noisy rdf data." International Semantic Web Conference. Springer Berlin Heidelberg, 2013.

Evaluation on gold standard GS1 (1021 entities) and GS2 (160 entities).

Classifier	P_{exact}	hP	hR	hF
STI_{prune}	.446	.780	.589	.671
$STI + hSVM_{text}\ \alpha$	.400	.763	.734	.748
$hSVM^{add}_{text}\ \beta$	.365	.719	.706	.712
$hSVM^{add}_{text}\ STI\ \beta$	.294	.817	.652	.726
DBpedia (2014)	.548	.890	.665	.761
GS2				
SDType (3.9)	.338	.809	.641	.715

Excerpt of results from our LHD 2.0 paper

- Results for our approach are comparable to SDType in terms of hP and hR
- We found that SDType and our approach are largely complementary w.r.t. entities covered
- SDType types entities based on ingoing/outgoing links (properties) why our approach uses text

ner.vse.cz/thd

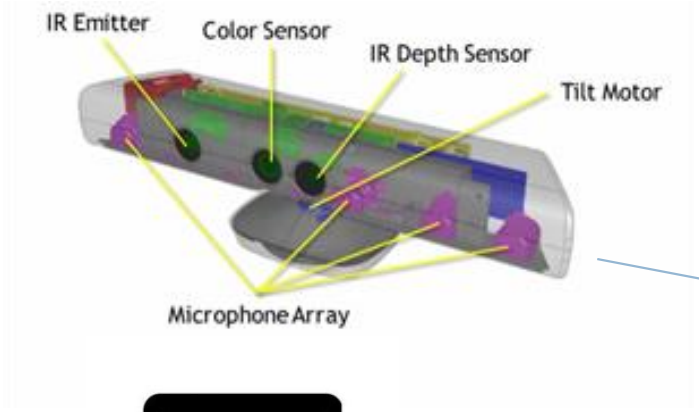
github.com/entityclassifier-eu/

Entityclassifier.eu NER

Recognition, Classification, Linking of Salient Named Entities

- Entity spotting
  - TreeTagger + GATE JAPE
  - Stanford NER
- Entity linking
  - String similarity
  - Lucene
  - Wikipedia Search
  - Surface form index
- Entity salience
  - SVM
- Languages
  - English, German, Dutch
- Knowledge bases
  - DBpedia, YAGO, LHD
- Stability
  - The system runs since 2012
  - Was used to annotate hundreds of thousands web pages
- Benchmarks
  - NIST TAC 2013, 2014
  - The Wikipedia search method had median performance in TAC 2013
  - GERBIL

Inbeat.eu: Our “Orwellian Eye”



general architecture
GATE
for text engineering

LEARNING PREFERENCE RULES

DBpedia

USER PREFERENCE
- **REMOTE CONTROL**
- **GAZE**

RECOMMENDATION OF CONTENT

SEMANTIC REPRESENTATION OF VIDEO CONTENT

Tomáš Kliegr, Jaroslav Kuchař: Orwellian Eye: Video Recommendation with Microsoft Kinect. In: Prestigious Applications Of Intelligent Systems. ECAI 2014. IOS Press

Credits and resources

Dataset

`ner.vse.cz/datasets/linkedhypernyms`

- Supplementary datasets (fine grained types, ontology alignment)
- Evaluation resources: gold standard datasets, guidelines, etc.

Václav Zeman

`github.com/KIZI/LinkedHypernymsDataset`

- LHD generation framework wrapped in Docker container

Ondřej Zamazal

`github.com/OndrejZamazal/hSVM3`

- hSVM implementation

`github.com/kliegr/hierarchical_evaluation_measures`

- Evaluation of DBpedia entity type algorithms

Use cases

`ner.vse.cz/thd` & github repositories

- Free to use API and open source entity classification software
- GATE plugin

Milan Dojchinovski

`Inbeat.eu` & github repository

- Inbeat semantic recommenders with sensor support

Jaroslav Kuchař

Publications

LHD algorithms

- T. Kliegr: Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. Journal of Web Semantics, Elsevier, 2015
- T. Kliegr and O. Zamazal: LHD 2.0: A text mining approach to typing entities in knowledge graphs. Journal of Web Semantics. Elsevier, 2016

LHD framework

- T. Kliegr, V. Zeman and M. Dojchinovski. Linked Hypernyms Dataset - Generation Framework and Use Cases. 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, At Reykjavik, Iceland. 2014

Applications/Use cases

- M. Dojchinovski and T. Kliegr: Entityclassifier.eu: Real-time Classification of Entities in Text with Wikipedia, European Conference on Machine Learning (ECML PKDD'13). Prague, Czech Republic, Springer, 2013
- T. Kliegr, J. Kuchař: Orwellian Eye: Video Recommendation with Microsoft Kinect. Prestigous Applications of Intelligent Systems, European Conference on Artificial Intelligence (PAIS/ECAI 2014), *Prague, Czech Republic,* IOS PRESS, 2014