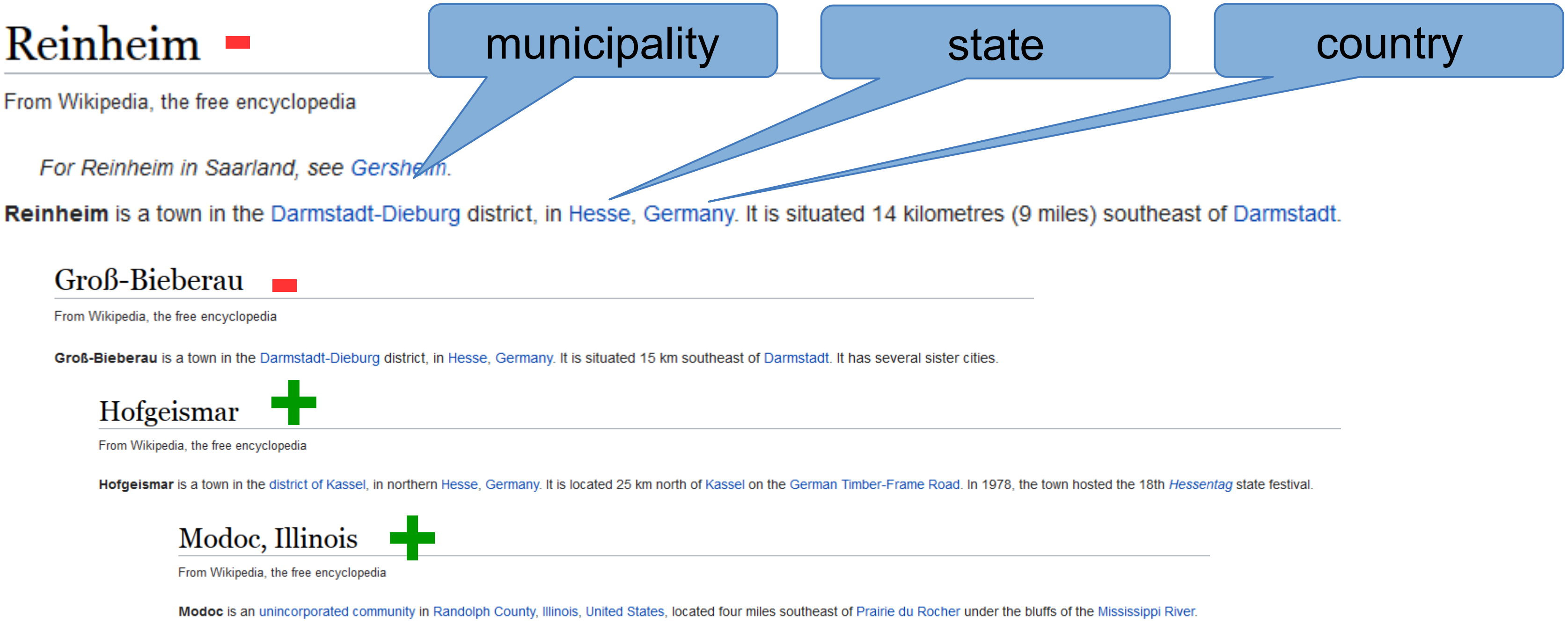
UNIVERSITÄT
MANNHEIM

Extracting Literal Assertions for DBpedia from Wikipedia Abstracts

Florian Schrage, Nicolas Heist, **Heiko Paulheim**

Flashback: ISWC 2017

- Heist, Paulheim (2017): “Language-agnostic relation extraction from Wikipedia abstracts”
- Main idea:
  - Find recurring patterns in abstracts



municipality

state

country

Reinheim −

From Wikipedia, the free encyclopedia

For Reinheim in Saarland, see Gersheim.

Reinheim is a town in the Darmstadt-Dieburg district, in Hesse, Germany. It is situated 14 kilometres (9 miles) southeast of Darmstadt.

Groß-Bieberau −

From Wikipedia, the free encyclopedia

Groß-Bieberau is a town in the Darmstadt-Dieburg district, in Hesse, Germany. It is situated 15 km southeast of Darmstadt. It has several sister cities.

Hofgeismar +

From Wikipedia, the free encyclopedia

Hofgeismar is a town in the district of Kassel, in northern Hesse, Germany. It is located 25 km north of Kassel on the German Timber-Frame Road. In 1978, the town hosted the 18th *Hessentag* state festival.

Modoc, Illinois +

From Wikipedia, the free encyclopedia

Modoc is an unincorporated community in Randolph County, Illinois, United States, located four miles southeast of Prairie du Rocher under the bluffs of the Mississippi River.

Heist, Paulheim: *Language-agnostic relation exrtaction from Wikipedia Abstracts. In: ISWC 2017*

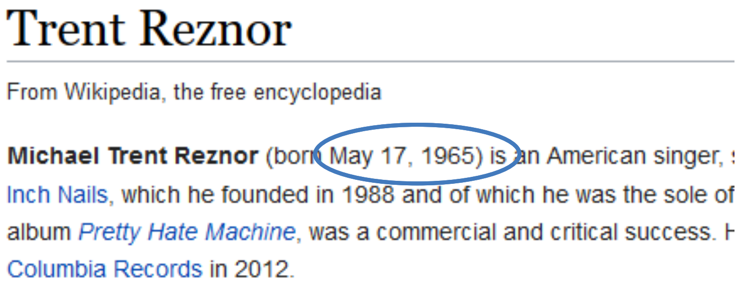
Observation: Typical Patterns

- The first three populated places linked in an abstract about a town are that town's *municipality*, *state*, and *country*
- All genres linked in an abstract about a writer are that writer's *genres*
- The first place linked in an abstract about a person is that person's *birthplace*

- Automatically finding those patterns:
  We can use existing relations as training data
  - Using a *local closed world assumption* for creating negative examples
- Training data:
  - Linked instances in an abstract, explicit relations extracted from infobox

From Entities to Numbers and Dates

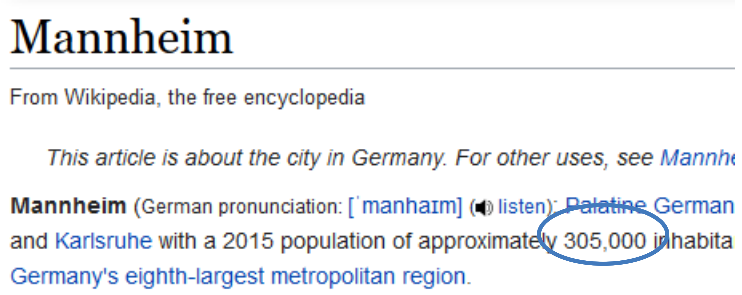
- Key assumption: such patterns also exist for numbers and dates
- Examples:
  - First date in an abstract about a *person* is the person's birthdate
  - First number in an abstract about a *city* is the city's population



Trent Reznor

From Wikipedia, the free encyclopedia

Michael Trent Reznor (born May 17, 1965) is an American singer,
Inch Nails, which he founded in 1988 and of which he was the sole of
album *Pretty Hate Machine*, was a commercial and critical success.
Columbia Records in 2012.



Mannheim

From Wikipedia, the free encyclopedia

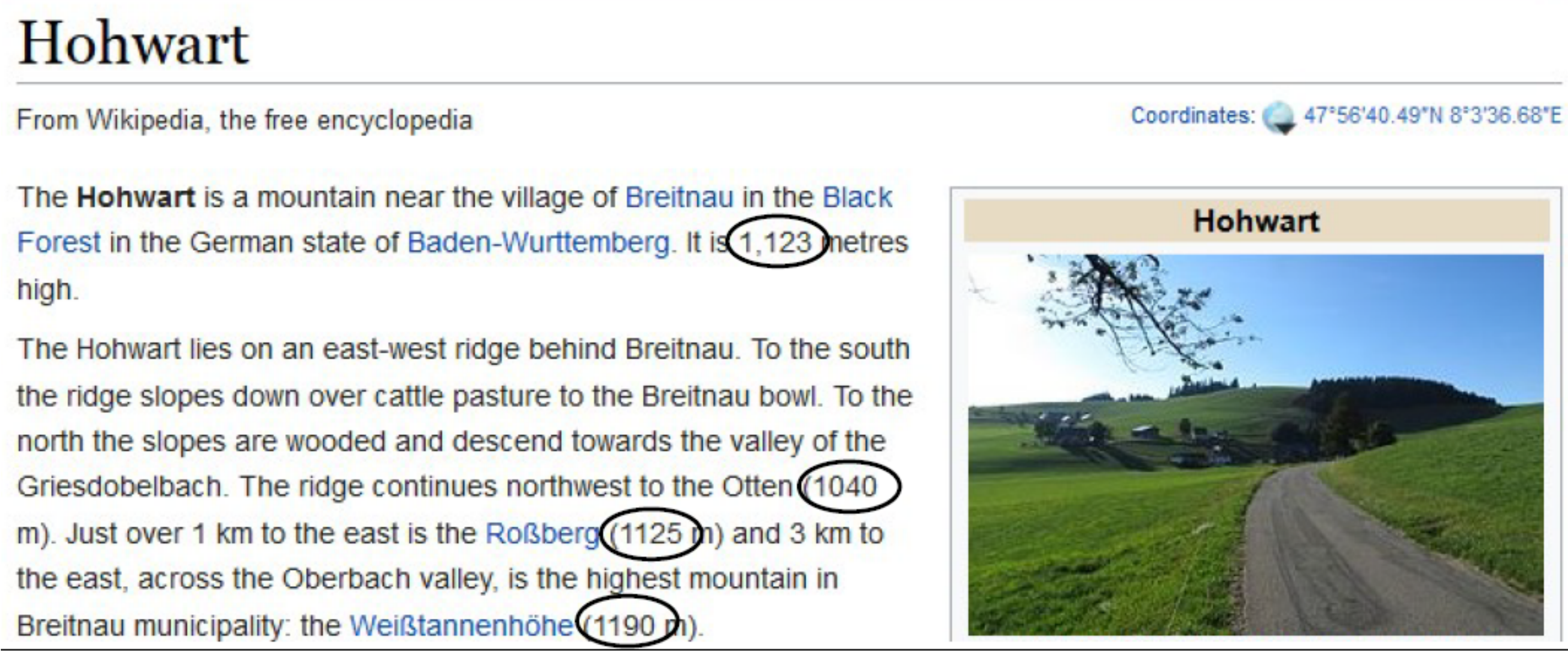
This article is about the city in Germany. For other uses, see Mannh

Mannheim (German pronunciation: [ˈmanhaɪm] (🔊 listen); Palatine German
and Karlsruhe with a 2015 population of approximately 305,000 inhabita
Germany's eighth-largest metropolitan region.

- Differences to entity-based extraction (aka: challenges)
  1. numbers/dates are neither tagged nor typed
  2. numbers/dates come in different formats
  3. infobox value and value in abstract may use a different format and/or unit of measure and/or rounding

Challenge: Number/Date Formats

- Sometimes even inconsistent *within* a single Wikipedia page



Hohwart

From Wikipedia, the free encyclopedia

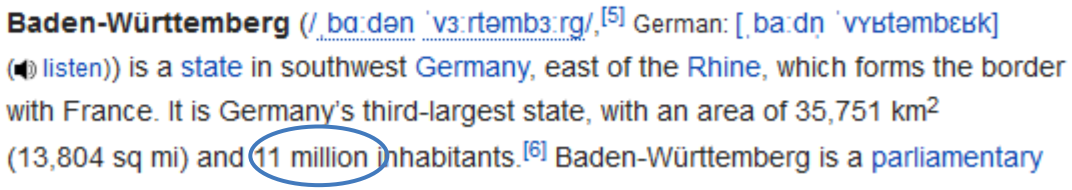
Coordinates: 47°56'40.49"N 8°3'36.68"E

The **Hohwart** is a mountain near the village of Breitnau in the Black Forest in the German state of Baden-Wurttemberg. It is 1,123 metres high.

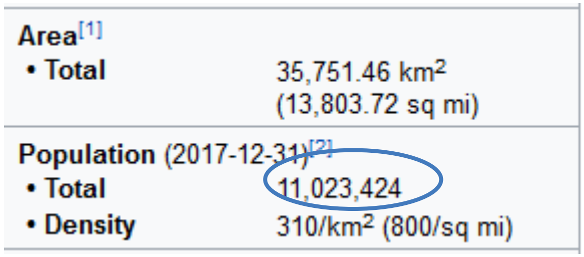
The Hohwart lies on an east-west ridge behind Breitnau. To the south the ridge slopes down over cattle pasture to the Breitnau bowl. To the north the slopes are wooded and descend towards the valley of the Griesdobelbach. The ridge continues northwest to the Otten (1040 m). Just over 1 km to the east is the Roßberg (1125 m) and 3 km to the east, across the Oberbach valley, is the highest mountain in Breitnau municipality: the Weißtannenhöhe (1190 m).

Paulheim: *A Robust Number Parser based on Conditional Random Fields.* In: KI 2017

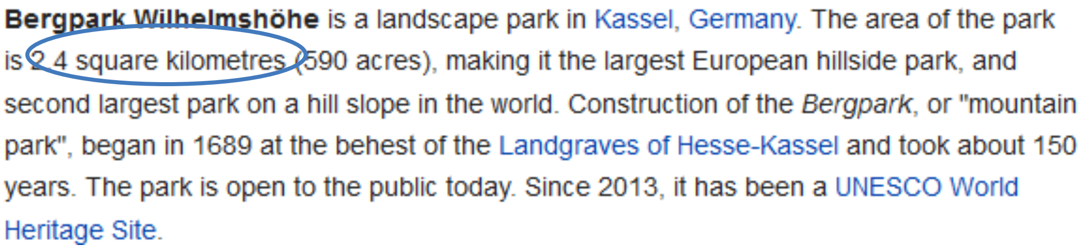
Challenge: Infobox vs. Text Mismatch



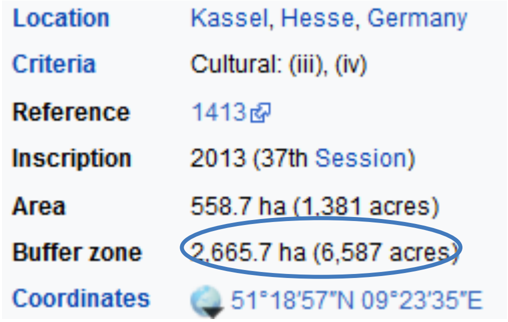
Baden-Württemberg (/ˌbɑːdən ˈvɜːrtəmbɜːrɡ/,[5] German: [ˌbaːdn̩ ˈvʏʁtəmbɛʁk] (listen)) is a state in southwest Germany, east of the Rhine, which forms the border with France. It is Germany's third-largest state, with an area of 35,751 km^2 (13,804 sq mi) and 11 million inhabitants.[6] Baden-Württemberg is a parliamentary



Area[1]	
• **Total**	35,751.46 km^2 (13,803.72 sq mi)
Population (2017-12-31)[2]	
• **Total**	11,023,424
• **Density**	310/km^2 (800/sq mi)



Bergpark Wilhelmshöhe is a landscape park in Kassel, Germany. The area of the park is 2.4 square kilometres (590 acres), making it the largest European hillside park, and second largest park on a hill slope in the world. Construction of the *Bergpark*, or "mountain park", began in 1689 at the behest of the Landgraves of Hesse-Kassel and took about 150 years. The park is open to the public today. Since 2013, it has been a UNESCO World Heritage Site.



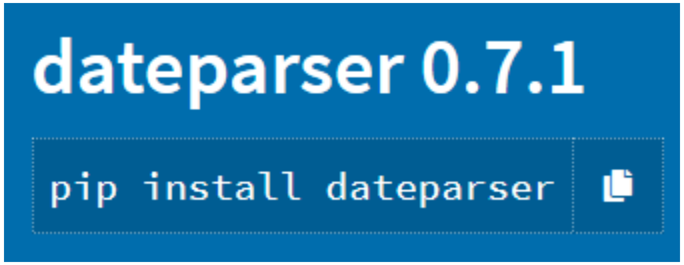
Location	Kassel, Hesse, Germany
Criteria	Cultural: (iii), (iv)
Reference	1413
Inscription	2013 (37th Session)
Area	558.7 ha (1,381 acres)
Buffer zone	2,665.7 ha (6,587 acres)
Coordinates	51°18′57″N 09°23′35″E

Creating Training Data

- First step: spot any character sequences containing numbers
  - Those could be numbers, dates, and others

- Second step:
  - Try to parse sequences with spaCy and dateparser
  - Tolerant, language-independent Python based number and date parsers

spaCy

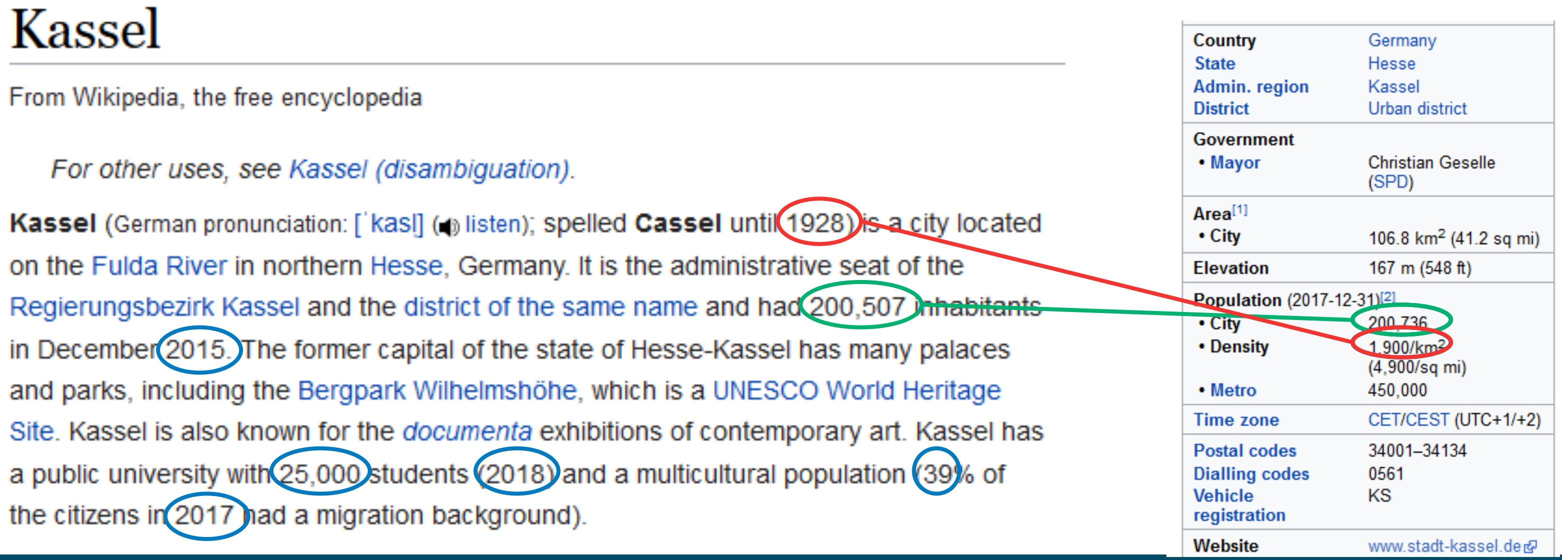


dateparser 0.7.1

```
pip install dateparser
```

Creating Training Data

- Challenge: abstracts often use rounded values
  - Or there are slight deviations
  - Experimented with 1%, 1.5%, 2% tolerance
  - Precision drops at 2% → we use 1.5%
- Gain: more training data
- Loss: false positives



Kassel

From Wikipedia, the free encyclopedia

For other uses, see Kassel (disambiguation).

Kassel (German pronunciation: [ˈkasl] (listen); spelled **Cassel** until 1928) is a city located on the Fulda River in northern Hesse, Germany. It is the administrative seat of the Regierungsbezirk Kassel and the district of the same name and had 200,507 inhabitants in December 2015. The former capital of the state of Hesse-Kassel has many palaces and parks, including the Bergpark Wilhelmshöhe, which is a UNESCO World Heritage Site. Kassel is also known for the *documenta* exhibitions of contemporary art. Kassel has a public university with 25,000 students (2018) and a multicultural population (39% of the citizens in 2017 had a migration background).

Country	Germany
State	Hesse
Admin. region	Kassel
District	Urban district
Government	
• **Mayor**	Christian Geselle (SPD)
Area[1]	
• **City**	106.8 km^2 (41.2 sq mi)
Elevation	167 m (548 ft)
Population (2017-12-31)[2]	
• **City**	200,736
• **Density**	1,900/km^2 (4,900/sq mi)
• **Metro**	450,000
Time zone	CET/CEST (UTC+1/+2)
Postal codes	34001–34134
Dialling codes	0561
Vehicle registration	KS
Website	www.stadt-kassel.de

Creating Training Data

- Challenge: different units of measure,
  mixed number-text notation (e.g., “3.4 million”)
- Approach: train a linear (b=0) model for context words
  - i.e., context words can be linked to linear factors
  - Accept models with at least 100 examples and R² value >0.85

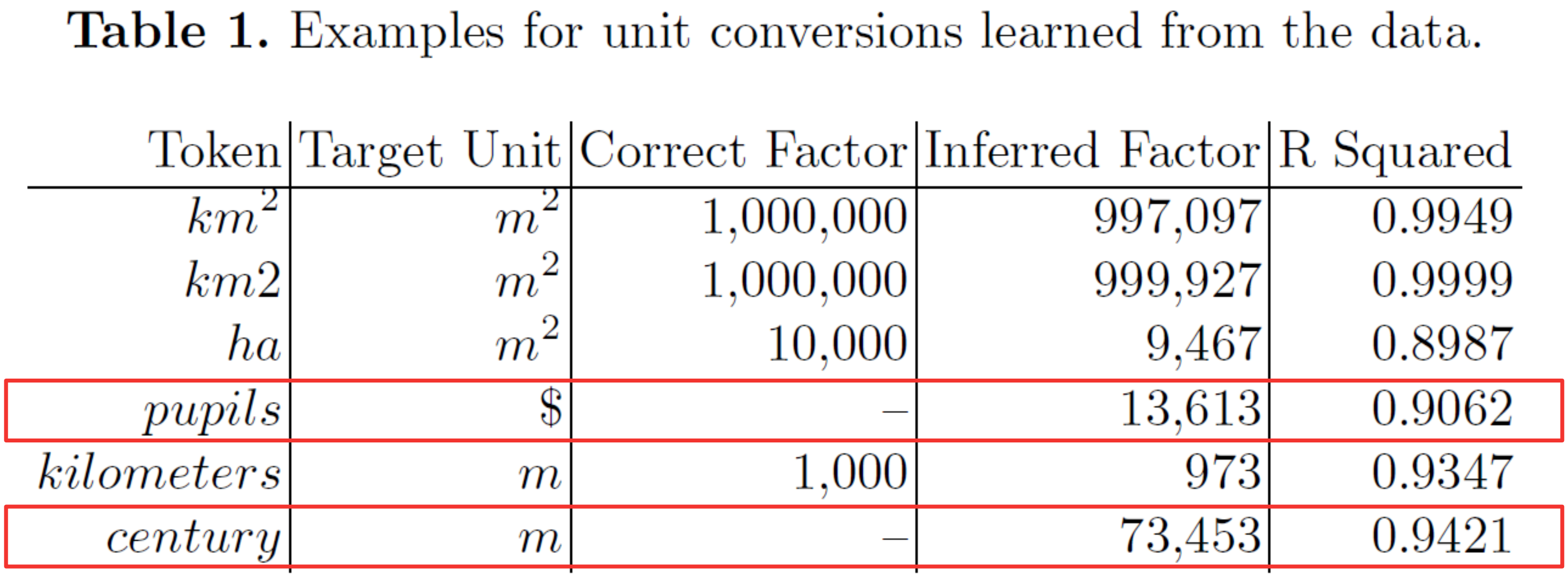
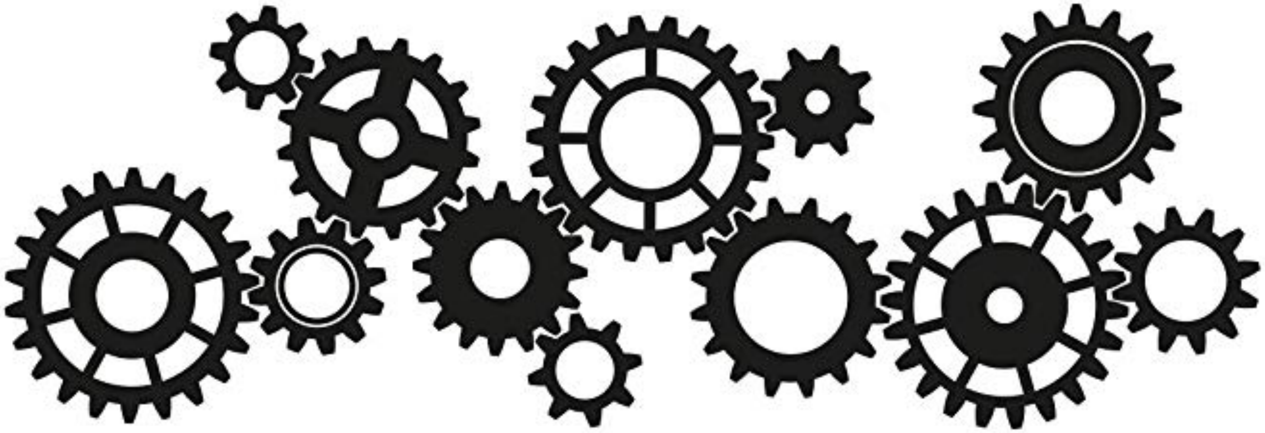


Table 1. Examples for unit conversions learned from the data.

Token	Target Unit	Correct Factor	Inferred Factor	R Squared
km^2	m^2	1,000,000	997,097	0.9949
$km2$	m^2	1,000,000	999,927	0.9999
ha	m^2	10,000	9,467	0.8987
$pupils$	\$	–	13,613	0.9062
$kilometers$	m	1,000	973	0.9347
$century$	m	–	73,453	0.9421

Overall Approach

- Extract numbers from abstract
- Match them to numbers in the infobox
  - Matching: positive example
  - Non-matching: negative example
- Train a classifier
  - Self-assessment: estimate precision
  - Only classifiers above 95% precision are used to produce statements



Experiments

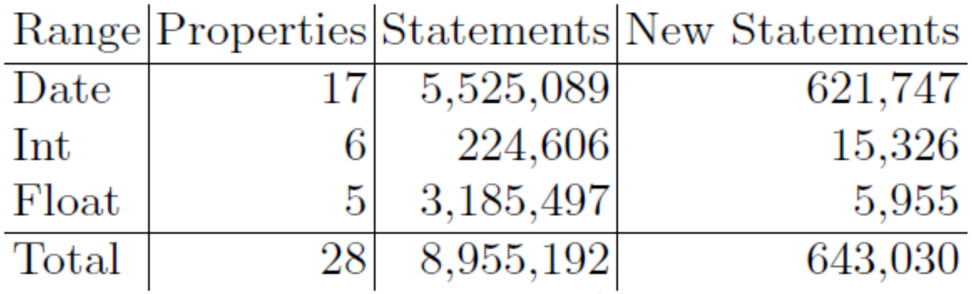
- Training example generation
  - Extracted by identifying matching pairs in abstract and infobox
  - Allowing deviation and linear factors (as above)
  - Negatives: non-matching numbers/dates in the same abstract
- Datasets used for classification (true/false extraction)
  - DBpedia 2016-10 and corresponding Wikipedia dump
  - 120 number and date valued properties
    w/ at least 100 positive training examples
    - 120 classifiers trained
    - 75%/25% split to allow self-assessment of trained models
    - 28 reach a precision >95%

Experiments

- Feature set
  - Motivation: patterns such as “The first number in an abstract...”
  - Features used: position in sentence, sentence in abstract, …
  - Plus: bag of words around literal (e.g., “birth”, “population”, …)
  - For numbers: deviation from mean
- Classifiers
  - SGD, Naive Bayes, SVM, Decision Trees, Random Forest, Extra Trees, Bagging Decision Trees, XGBoost
  - RandomForest used and fine-tuned after initial experiment

Results

- 28 properties for which a model with 95% precision is trained
  - Those generate 9M facts
  - 7% are not contained in DBpedia
    - Mostly dates, not numbers



Range	Properties	Statements	New Statements
Date	17	5,525,089	621,747
Int	6	224,606	15,326
Float	5	3,185,497	5,955
Total	28	8,955,192	643,030

Table 2. Number of statements extracted at 95% precision according to internal validation.

- Posterior validation on 500 newly generated facts
  - Precision is 94.2%
  - i.e., estimated precision is valid

Take Aways

- Literal-valued relations are challenging
- Tweaks to original entity-based approach
  - Number/date tagging and parsing
  - Tolerance intervals
  - Learned model for unit conversion
- 9M statements could be extracted (600k new)

- Code: https://github.com/FlorianSchrage/DBpediaLiteralRelations

Future Challenges

- Deeper analysis of deviations
  - Is the correct value more likely in the abstract or the infobox?

- Better training data and learning
  - Robustly discarding false matches
  - Learning models for smaller datasets

- Learning complex formulae
  - e.g., population density

- Transfer to other datasets
  - e.g., DBkWik

Extracting Literal Assertions for DBpedia from Wikipedia Abstracts

Florian Schrage, Nicolas Heist, **Heiko Paulheim**