Towards Disambiguating Web Tables

{Stefan.Zwicklbauer, Christoph.Einsiedler, Christin.Seifert, Michael.Granitzer }@uni-passau.de

Problem

Tables on web sites or in scientific papers represent a valuable source of information for the human reader. The information itself is meaningless to machines – unless it is enriched with semantic information. The Semantic Web, and specifically the Linked Open Data initiative provide means for representing any kind of knowledge semantically.
If tables were enriched semantically a variety of new applications could evolve, as is the idea of Google Fusion Tables [1] where the annotation is done by humans.

Contributions

We formulated column type inference as the majority search of cell types.
The method is based on the majority vote algorithm proposed in [2]. Our main contributions are:

1. Simple algorithm to annotate table headers with semantic types based on the types cell entities
2. Investigating the influence of the number of cells on the accuracy of the header-type inference

Data Set

We used DbPedia as knowledge base with type relations `rdf:type` and `dcterms:subject` from the Dublin Core Metadata Ontology. We evaluated our system on 50 tables extracted from Wikipedia including tables mentioned in [2].

# Columns	132
# Rows	2707
# RDF:type annotations	169
# DCT:subject annotations	160
ø RDF:type annotations	1.29
ø DCT:subject annotations	1.21

Result

As a result our work contributes:

⇒ Algorithm to annotate table headers

⇒ Similar accuracy as previous work with more complex methods

⇒ Reasonable to use only a small number of cells for annotating the header

References

[1] Gonzalez, H., Halevy, A. Y., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., Shen, W. and Goldberg-Kidon, J., Google Fusion Tables: Web-Centered Data Management and Collaboration In *Proc. ACM SIGMOD 2010*

[2] Limaye, G., Sarawagi, S. and Chakrabarti, S. Annotating and Searching Web Tables Using Entities, Types and Relationships In *VLDB 2010*

[3] Zwicklbauer, S., Seifert, C. and Granitzer, M. Do We Need Entity-Centric Knowledge Bases for Entity Disambiguatin? In *i-Know 2013*

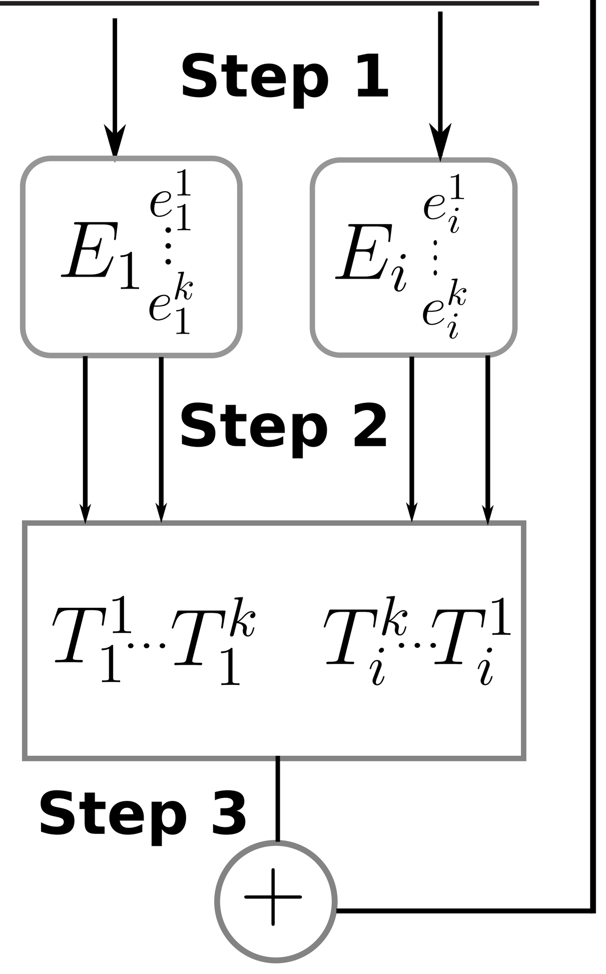
Method

Web Table of Nuclear Power Plants

Operator	Location	Reactor	Type
Aerotest	San Ramon	TRIGA Mark I	
Armed Forces	Bethesda	TRIGA Mark F	l_1
Cornell University	Ithaca	TRIGA Mark I	⋮
Idaho University	Pocatello	AGN-201	l_i
MIT	Cambridge	Tank Type HWR	⋮

Let l_i $1 < i \leq n$ be the labels of non-header cells i and $E_i = \{e_i^k\}$ is the set of all possible semantic meanings of label l_i. The set T_i^k is the set of all type labels assigned to entity e_i^k. **We make no assumptions of interrelationships between columns**, i.e. all columns are treated separately. Additionally, we assume that the tables do not have merged cells. The annotation of table headers is performed in three steps:

1. For each cell label l_i we derive a list of k possible entity candidates E_i using a search-based disambiguation method [3].
2. For each entity candidate e_i^k in E_i a set of types is retrieved by following the `rdf:type` and `dcterms:subject` relations yielding the set of types T_i^k.
3. The types assigned to the table header are the t types that occur most frequently in the set of all types of all cells $\bigcup_i \bigcup_k T_i^k$. We set $t = 1$ in our experiments, e.g. only use the most frequent type as result.



Column Type Inference

We assessed the overall performance on the complete data set. The table below shows the results for three different type vocabularies (using Rdf-Type relations only, using DublinCore subjects only, or using both), whereby macro-averaged precision is denoted as π and recall is denoted as ρ. **In terms of precision the combined vocabulary performs best (0.64)**, however only slightly better than using DublinCore subjects only (0.59), whereas Rdf-type annotations are worst (0.24). **For the combined approach, F_1 is low due to the low recall**, which is because we have more correct results in the ground-truth but consider only the best result in the evaluation.

Vocabulary	π^M	ρ^M	F_1^M
Rdf-Type	0.24	0.22	0.23
DublinCore	0.59	**0.51**	**0.55**
Rdf-Type + DublinCore	**0.64**	0.27	0.38

Influence of Number of Cells

We assessed the influence of the number of cells on the accuracy of table header disambiguation. From all 192 columns we randomly selected k cells for the cell-entity annotation step and assessed the header disambiguation accuracy using the DublinCore vocabulary. We repeated the experiment 10 times with different randomly selected cells for each $k \in \{1, 2, ..7, 8, 10, 12, 15, 20\}$. As expected **for small numbers of cells the performance increases significantly when adding one more cell** (e.g. from 3 to 4 cells the F_1 measure increases from 0.27 to 0.35 a growth of 26%). **For larger numbers of cells there is less information gain by adding one more cell** resulting in smaller increases in performance (all below 10%). Using 20 cells results in F_1 of 0.514, which is 94% of the F_1 achieved with all cells (0.547).