MantisTable an automatic approach for the Semantic Table Interpretation



Department of Computer Science, Systems and Communication (DISCo) University of Milano - Bicocca



Marco Cremaschi, Roberto Avogadro, and David Chieregato







Semantic Table Interpretation: an example

TABLE

Name	Coordinates	Height	Range
Mont Blanc	45°49'57"N 06°51'52"E	4808	Mont Blanc
Lyskamm	45°55′20″N 07°50′08″E	4527	Pennine Alps
Monte Cervino	45°58'35"N 07°39'31"E	4478	Pennine Alps

Subject column (S-column)

Named-Entity column (NE-column)

Literal column (L-column)





- 1. Data Preparation, which aims to prepare the data inside the table
- 2. Column Analysis, whose tasks are the semantic classification that assigns types to columns (NE-column or L-column), and the detection of the subject column (S-column)
- 3. Concept and Datatype Annotation, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG
- 4. Predicate Annotation, whose task is to find relations, in the form of predicates, between the main column and the other columns to set the overall meaning of the table
- 5. Entity Linking, which deals with mappings between cells and entities in a KG

Data Preparation, which aims to prepare the data inside the table

Name	Coordinates	Height	
mont blanc	45°49′57″N 06° 51′52″E	4808	mont
lyskamm	45°55′20″N 07°50′08″E	4527	ре
monte cervino	45°58'35"N 07°39'31"E	4478	ре

	 removal of HTML tags and stop
Range	words
blanc massif	 transformation of the text into
	 resolution of acronyms and
nnine alps	abbreviation
	 normalization of units of
nnine alps	measurement by applying regular
	expressions

(NE-column or L-column), and the detection of the subject column (S-column)

S-column	NE-column		
Name	Coordinates	Height	
mont blanc	45°49′57″N 06° 51′52″E	4808	mont
lyskamm	45°55′20″N 07°50′08″E	4527	ре
monte cervino	45°58'35"N 07°39'31"E	4478	ре

Column Analysis, whose tasks are the semantic classification that assigns types to columns

L-column	Detection
Range	(e.g., geo
mont blanc massif	 Color co Detection
pennine alps	differen
pennine alps	$subcol(c_j$

- ion of L-columns by 16 regular sions to identify regextype o coordinate, address, hex ode, URL)
- on of S-column considers nt statistic features

 $= \frac{2uc_{norm}(c_j) + aw_{norm}(c_j) - emc_{norm}(c_j)}{2uc_{norm}(c_j) + aw_{norm}(c_j) - emc_{norm}(c_j)}$ $\sqrt{df(c_j){+}1}$

Concept and Datatype Annotation, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG

Name	Coordinates	Height	
mont blanc	45°49′57″N 06° 51′52″E	4808	mont
lyskamm	45°55′20″N 07°50′08″E	4527	ре
monte cervino	45°58'35"N 07°39'31"E	4478	ре
MOUNTAIN	PLACE	t HEIGHT	

Range	 Retrieval of a set of candidate entities
t blanc massif	searching the Knowledge Graph w the content of a cell
ennine alps	 Retrieval of abstract and concepts abstract of retrieved
ennine alps	entities
MASSIF	 Application of heuristics for the identification of the most frequent concept of the column

Concept and Datatype Annotation, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG

 $econtext(e_{i,j}) = |bow(abstract(e_{i,j})) \cap bow(rcontent(i,j))| + |bow(abstract(e_{i,j})) \cap bow(hcontent(i,j))| + |bow(abs$ Header of the Row of the Abstract of the entity table column

inside the KG

 $ename(e_{i,j}) = editDistance(tx(i, j), e_{i,j})$

Text in the cell

column and the other columns to set the overall meaning of the table

Name	Coordinates	Height	
mont blanc	45°49′57″N 06° 51′52″E	4808	mont
lyskamm	45°55′20″N 07°50′08″E	4527	ре
monte cervino	45°58'35"N 07°39'31"E	4478	ре
		ţ	
MOUNTAIN	PLACE	HEIGHT	
geo	dbo:elevatio	n dbo:mou	ntainRange

Predicate Annotation, whose task is to find relations, in the form of predicates, between the main

Range

- blanc massif
- nnine alps
- nnine alps

- The winning concept of the S-column are considered as the subject of the relationship and annotations of the other columns as objects
- The Knowledge Graph is searched for the subject and the object to collect possible predicates

Predicate Annotation, whose task is to find relations, in the form of predicates, between the main column and the other columns to set the overall meaning of the table [Zhang 2017]

$$pcontext(p_j) = dice(p_j, x_j) = \underbrace{\frac{2 \cdot \sum_{w \in bowset(p_j) \cap bowset(x_j)} (freq(w, bow(p_j)) + freq(w, bow(x_j)))}{|bow(p_j)| + |bow(x_j)|}}_{\text{Predicate}}$$

pfreq

$$(p_j) = rac{|p_j|}{\sum_j |p_j|}$$

Entity Linking, which deals with mappings between cells and entities in a KG

Name	Coordinates	Height	Range
mont blanc <u>dbr:Mont Blanc</u>	45°49′57″N 06°51′52″E	4808	mont blanc mas dbr:Mont_Blanc_n
lyskamm <u>dbr:Lyskamm</u>	45°55′20″N 07°50′08″E	4527	pennine alps <u>dbr:Pennine_Al</u>
monte cervino <u>dbr:Monte Cervin</u> <u>O</u>	45°58′35″N 07°39′31″E	C Non sicuro dbpedia.org/page/M Browse usin About: Mont Bla An Entity of Type : natural place, from M	Iont_Blanc
		(This article is about the m (disambiguation).) Mont Bl Bianco (Italian pronunciati Mountain", is the highest m the Caucasus peaks. It ris ranked 11th in the world in	nountain. For other uses, see Mont Blanc lanc (French pronunciation: [mɔ̃ blɑ̃]) or Monte on: [ˈmonte ˈbjaŋko]), both meaning "White nountain in the Alps and the highest in Europe at es 4,808.73 m (15,777 ft) above sea level and is topographic prominence.

Range

blanc massif t_Blanc_massif

nnine alps ennine Alps

國 ☆	9	2	4	;
ceted Browser 🛛 🗹 Sparql Endpoint				
Data Space : dbpedia.org				
ee Mont Blanc				
: [mɔ̃ blɑ̃]) or Monte				
meaning "White				
e highest in Europe after				

- Already discovered annotations are used to create a query for the disambiguation of the cell content
- If more than one entity is returned for a cell, the one with a smaller edit distance is taken

	CTA	
	Primary score	Secondary score
Round 1	.929	.933
Round 2	1.049	.247
Round 3	1.648	.269
Round 4	1.682	.322

CEA				
	Primary score	Secondary score		
Round 1	1	1		
Round 2	.614	.673		
Round 3	.633	.679		
Round 4	.973	.983		

CPA								
	Primary score	Secondary score						
Round 1	.965	.991						
Round 2	.460	.544						
Round 3	.518	.595						
Round 4	.787	.841						

Search for the path in the graph that links all the entities in the row

11

	10	- D	WHEN WHEN	1 B 1000	B. Browness	n dimension		1.00	· II Berer
	10.000	Concernment of the local division of the loc	140	Contraction of the second	Torong big parts			the second data and the second second	
	1	And Approve Automatical	440		144/12 1940	Concerning in a concerning			All local 2 pt. and annual
	199		4.441		watering in the little				And Description of The Local Division of The
	pad	and a	4.bit		and the second s	\$1.00 million and the approximate			
	400		Lat.		Phys 21, 2007				
	444		Lat		The second se				
	471		4.10		mark. store	In our other sectors and the second sectors and			
	144		1.00		Service come	is in many production of the second			
-	-		a hite		- 10 h 10 m				
	0.00	and the second s	149		Telefolder	14 million Al Approximate			
	4444	and a second second	1.04	inclusion in the local diversion of the local	and the		104	on administra of print participants.	
•	41.01	Through Street	"bar		are \$ 100				
	1111	and the second s	ade .		and 1 (1986)				
	20.00	which an all the presentation	Tant.		10012-000				
	40.00	gaterine in	LND .		agt toth	10 June 19, 50 August Street Street		on an and the log of the partners of the	provide a state of the second
	1.00	and a second	Like .		and the same	in the second second second			
	100	And a second sec	Torn.		1000.000				provide a second
	1.00	daming rest.	Take .		ar 8.788	PROPERTY A. TAXABLE PROPERTY.		on president and in for particular party	And in case of a local week.
	1000	Designation of the	100		The Rolling				International Party of the local division of
		anapter of	1.00		allest the				complete a to decayly of the

the second se

The other side of the share in the same

MANTISTABLE

Acres 1

- Load tables in JSON format
- Download annotations (RDF/XML, N3, NTriples, Turtle and JSON-LD)
- Possibility to explore the output of each phase
- Manual annotation editing function
- Integration of the API provided by ABSTAT for auto-completion and suggestions

rmat es, Turtle ne ting

Marco Cremaschi

PhD Student@UNIMIB

marco.cremaschi@unimib.it

