RECA: Related Tables Enhanced Column Semantic Type Annotation Framework

Yushi Sun[1], Hao Xin[1], Lei Chen[1,2]

[1]The Hong Kong University of Science and Technology,
Hong Kong SAR, China
[2]The Hong Kong University of Science and Technology (Guangzhou),
Guangzhou, China

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

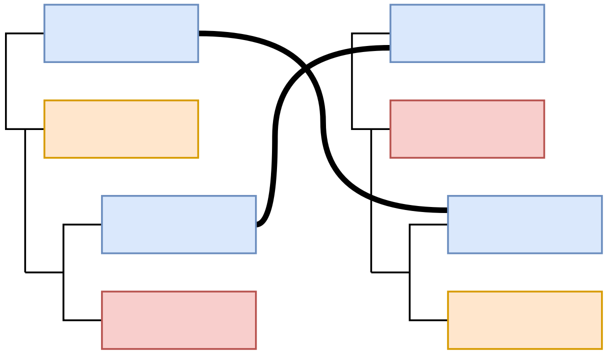
香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Outline

- Background and Motivation
- Definitions
- Methodology
- Experiments
- Summary

1. Background and Motivation

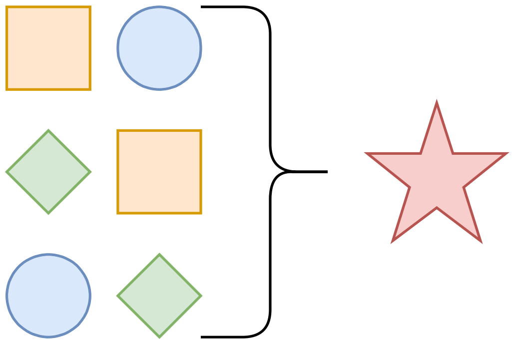
- Accurate column semantic type annotation is important for various applications:
  - schema matching, data cleaning, data integration, etc.



schema matching

Title 1	Title 2	Title 3
Value 1	Value 2	Value 3
Value 4	? ? ?	Value 6
Value 7	Value 8	Value 9
Value 10	Value 11	Value 12

data cleaning



data integration

1. Background and Motivation

- Two challenges exist:
  - The proper handle of wide tables
  - The utilization of inter-table context

?	?	?	?	?	...	?
Albania	27,398	$11,800	$2,949.57	2,994,667	…	Parliamentary Democracy
Algeria	2,381,740	$159,000	$3,948.01	34,994,937	…	Republic
Angola	1,246,700	$85,810	$5,003.43	13,338,541	…	Republic; Multiparty Presidential Regime
...	…	…	…	…	…	…

?	?	?	?	?	...	?
Bahamas	10,070	$7,538	$21,547.17	313,312	…	Constitutional Monarchy with a parliamentary system of government
Bangladesh	133,910	$100,100	$481.36	158,570,535	…	Parliame
Belgium	30,278	$461,300	$43,648.01	10,431,477	…	Federal F
...	…	…	…	…	…	…

?	?	?	?	?	...	?
Canada	9,984,670	$1,334,140	$40,457	34,733,000	…	Federal Parliamentary Democracy
Chile	748,800	$199,200	$10,058.50	16,888,760	…	republic
China	9,326,410	$5,745,000	$2,459.43	1,336,718,015	…	Communist State
...	…	…	…	…	…	…

1. Background and Motivation

- Two challenges exist:
  - The proper handle of wide tables
  - The utilization of inter-table context

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

WPPD

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

1. Background and Motivation

- Tables with the same/similar named entity schemata tend to be from the same/similar data source and thus tend to have the same/similar column semantic types.

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

WPPD

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

?	?	?	?
Donkey Kong Country	Nintendo	2006-12-08	2006
F-Zero	Nintendo	2006-12-08	2006
SimCity	Nintendo	2006-12-29	2006
Super Castlevania IV	Konami	2006-12-29	2006
Street Fighter II: The World Warrior	Capcom	2007-01-19	2007
...	...	...	...

WODD

- W: Work of art; P: Person; D: Date; O: Organization

Outline

- Background and Motivation
- Definitions
- Methodology
- Experiments
- Summary

2. Definitions - Concepts

- Named Entity Schema: Named Entity Schema is the table schema generated based on the most frequent named entity type extracted from each column.
- Related Tables: The tables that share the same named entity schema and are similar in content (Jaccard Similarity $> \delta$) with the original table.
- Sub-related Tables: The tables that share a similar named entity schema (the edit distance between their named entity schemata is less than a threshold) and are similar in content (Jaccard Similarity $> \delta$) with the original table.

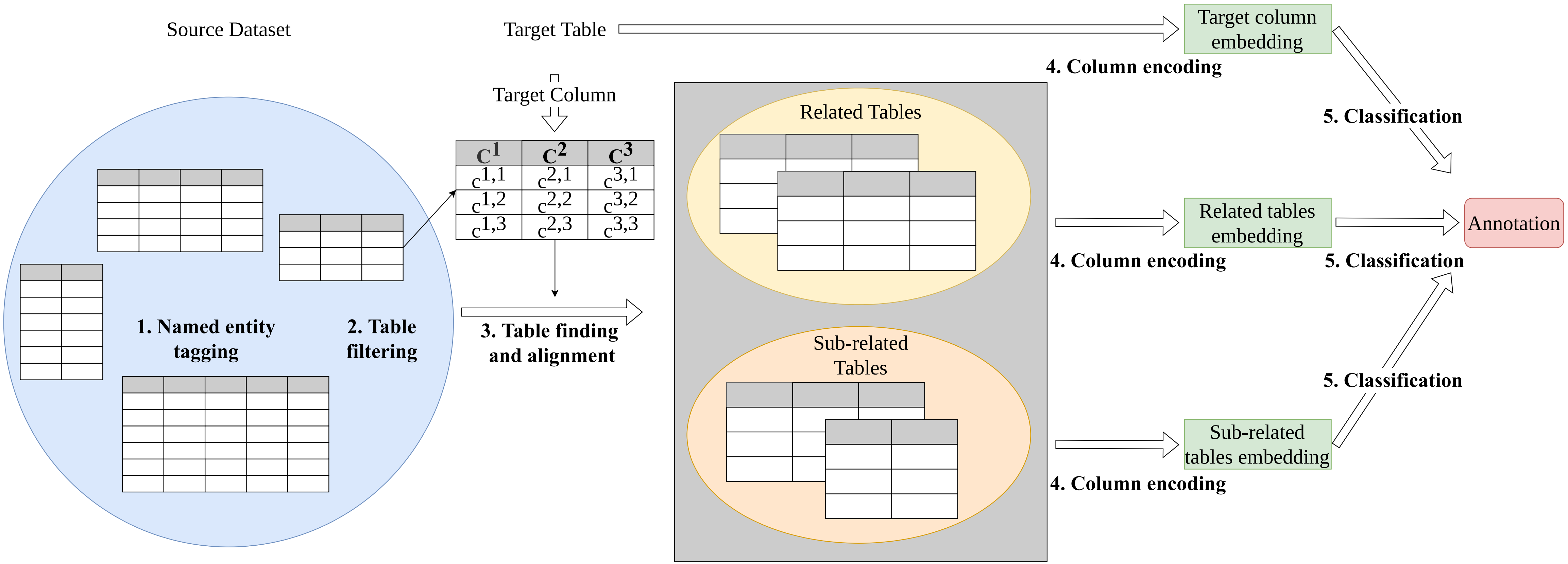
2. Definitions - Problem

- (Column semantic type annotation): Given a web table T (without table headers) from the dataset D, denote the target column as C_t in T. The column semantic type annotation model W annotates C_t with a semantic type $\bar{y}_t$ = W (C_t, T, D), such that $\bar{y}_t$ best fits the semantics of C_t.

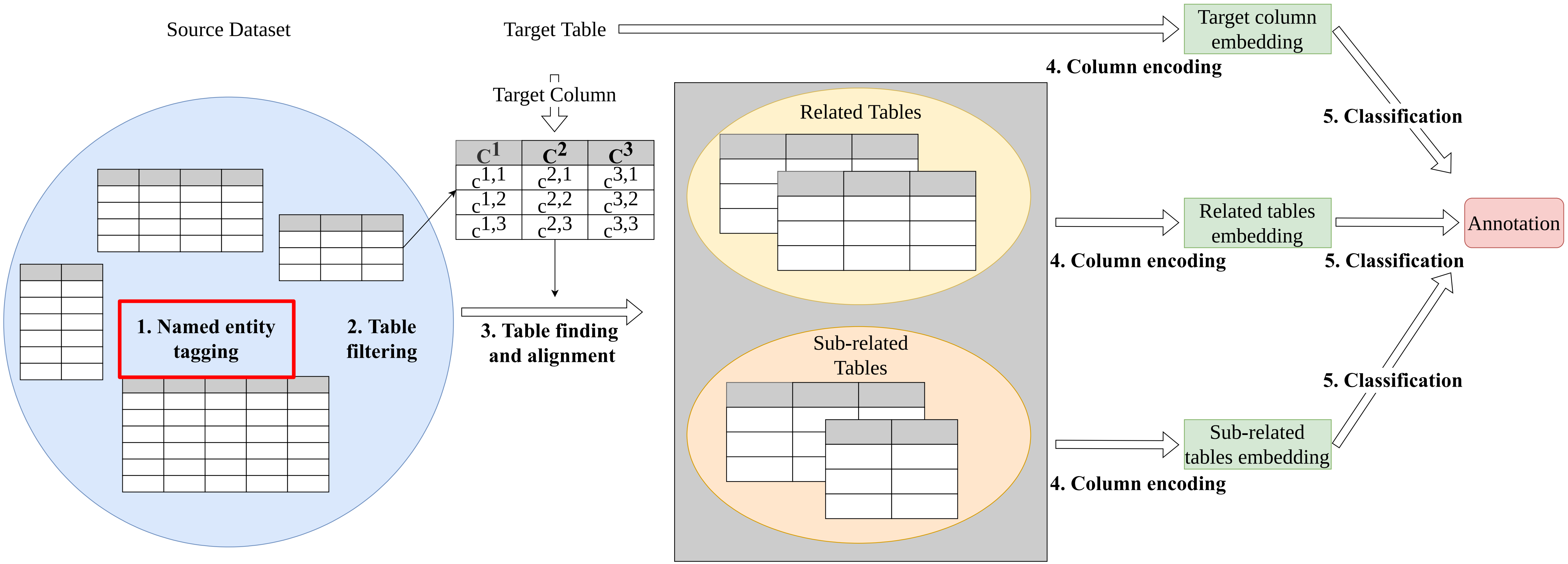
Outline

- Background and Motivation
- Definitions
- Methodology
- Experiments
- Summary

3. Methodology



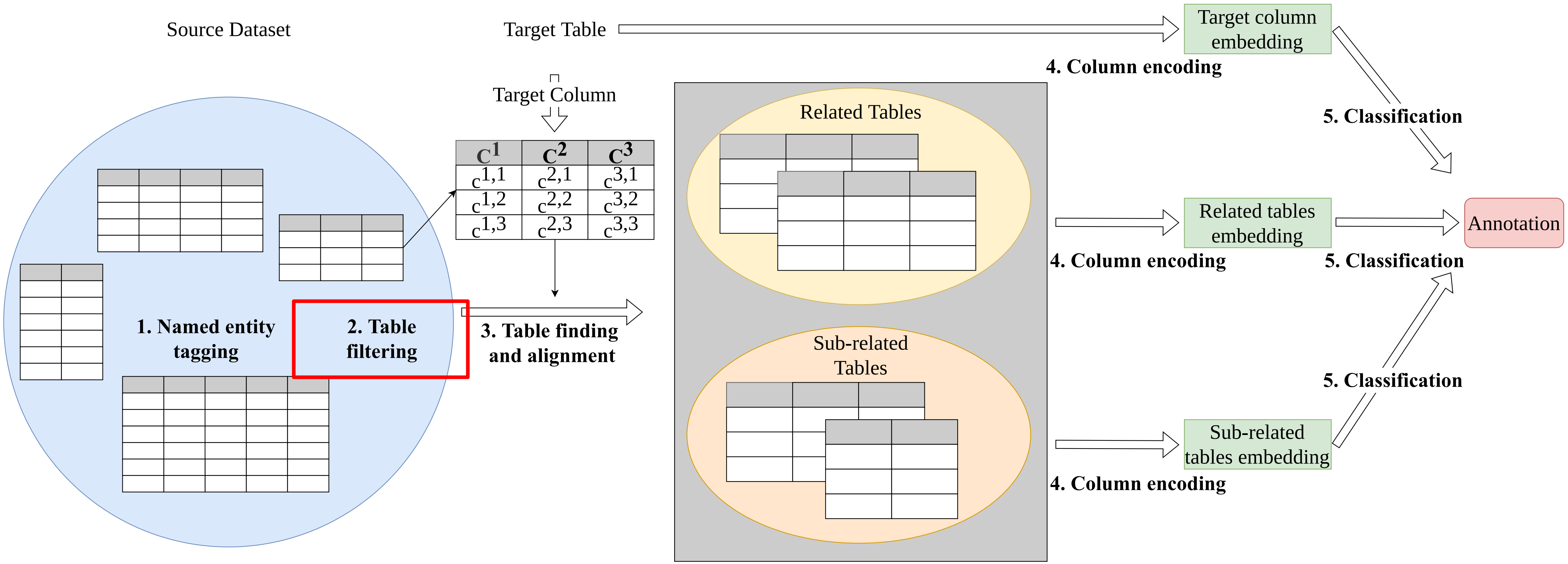
3. Methodology



3. Methodology - Named Entity Tagging

- Given a table T with M columns and N rows, we use the spaCy tagging tool to identify the named entities in each column and tag them.
- We further classify the DATE and PERSON types based on the data format.
  - E.g. DD-MM-YYYY; YYYY; January 16^{th} 2022; 2023
  - E.g. J. K. Rowling; Anna
- We include an additional EMPTY type.
- The most frequent named entity type in each column forms the named entity schema.

3. Methodology



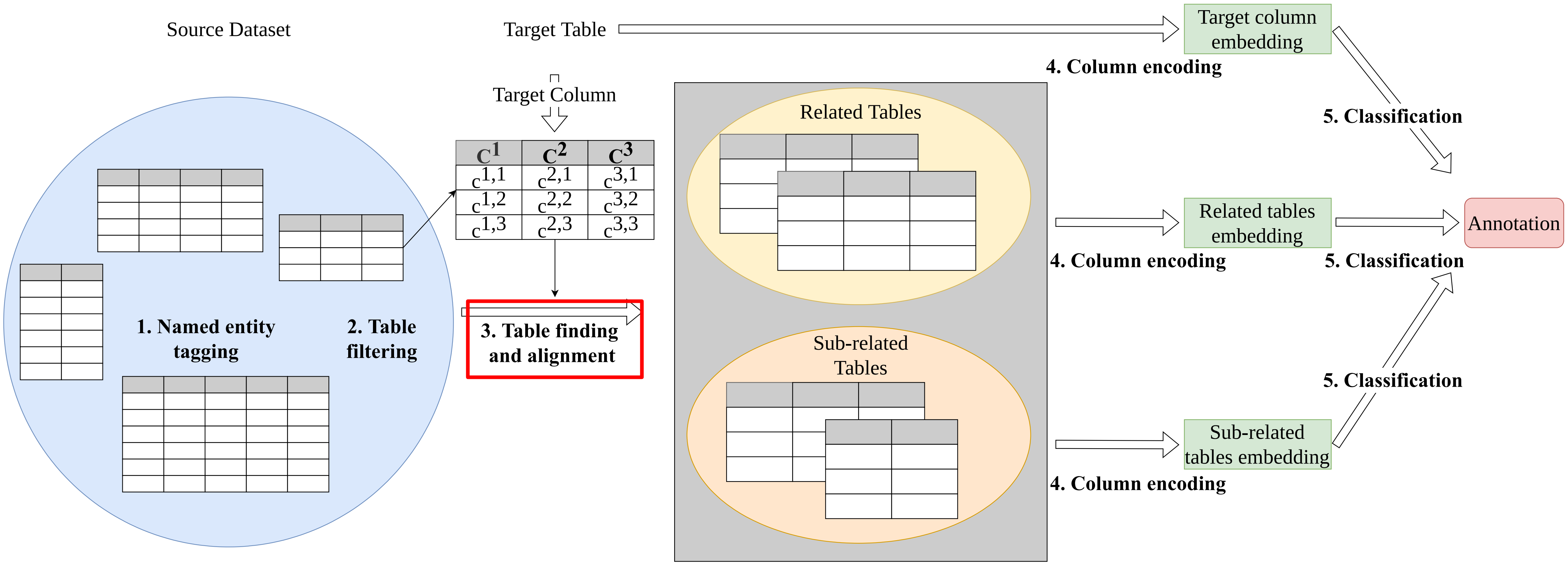
3. Methodology - Table Filtering

- To filter out tables that are irrelevant in content, we compute the Jaccard similarity between the set of words for each table pair.

$$\text{Jaccard}(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

- If $\text{Jaccard}(A_i, A_j) > \delta$, include T_j as a candidate table of T_i.

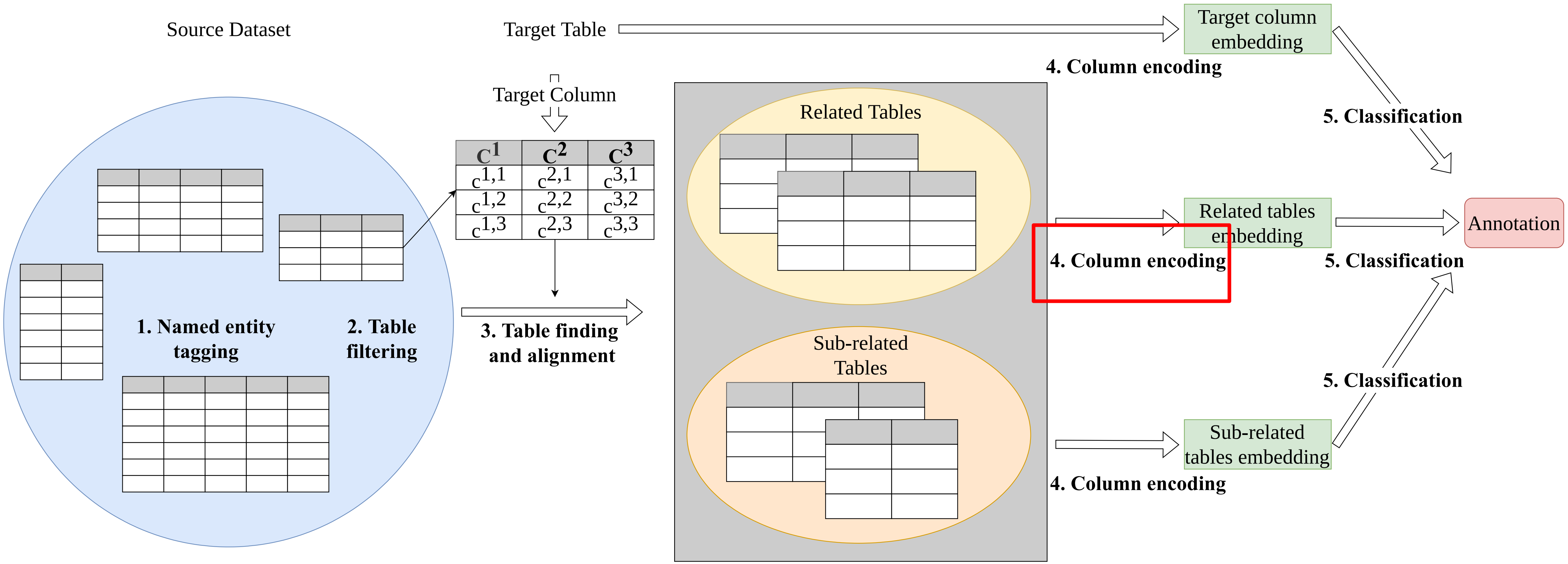
3. Methodology



3. Methodology - Table Finding and Alignment

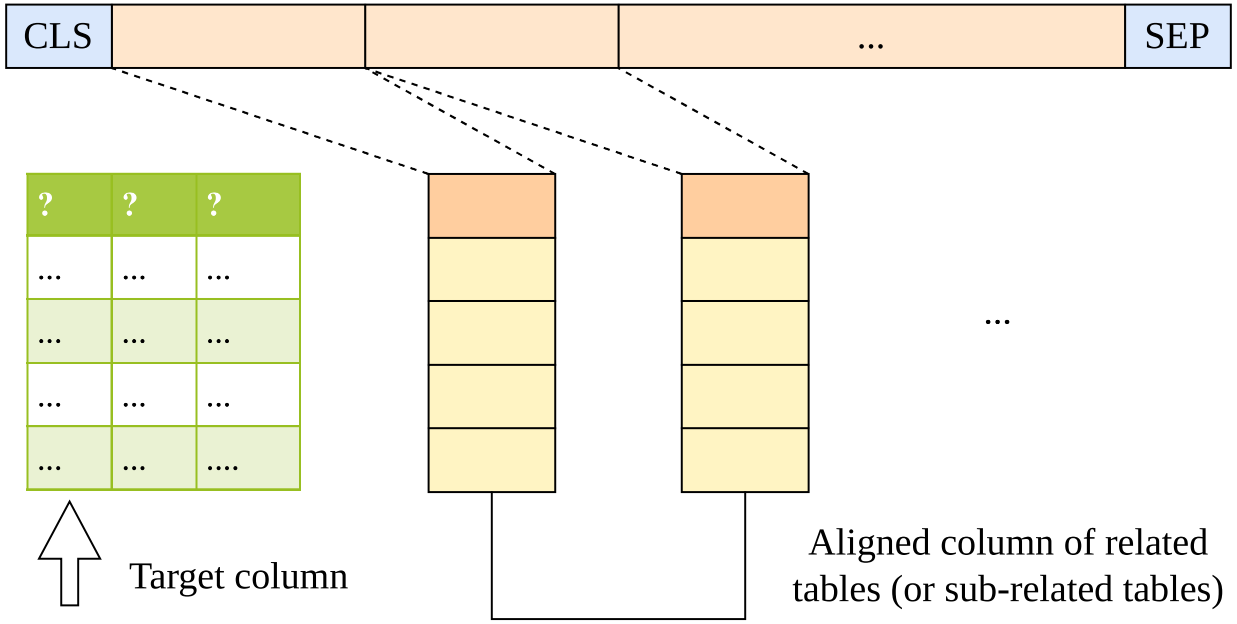
- Related tables: candidate tables T_j that share the same named entity schema as T_i.
- Sub-related tables: we consider the following two requirements:
  - Schema similarity: the named entity schemata should not be very different (edit distance less than a threshold).
  - Column location alignment: The named entity type of the target column matches with that of the column at the identical location in the sub-related table.

3. Methodology

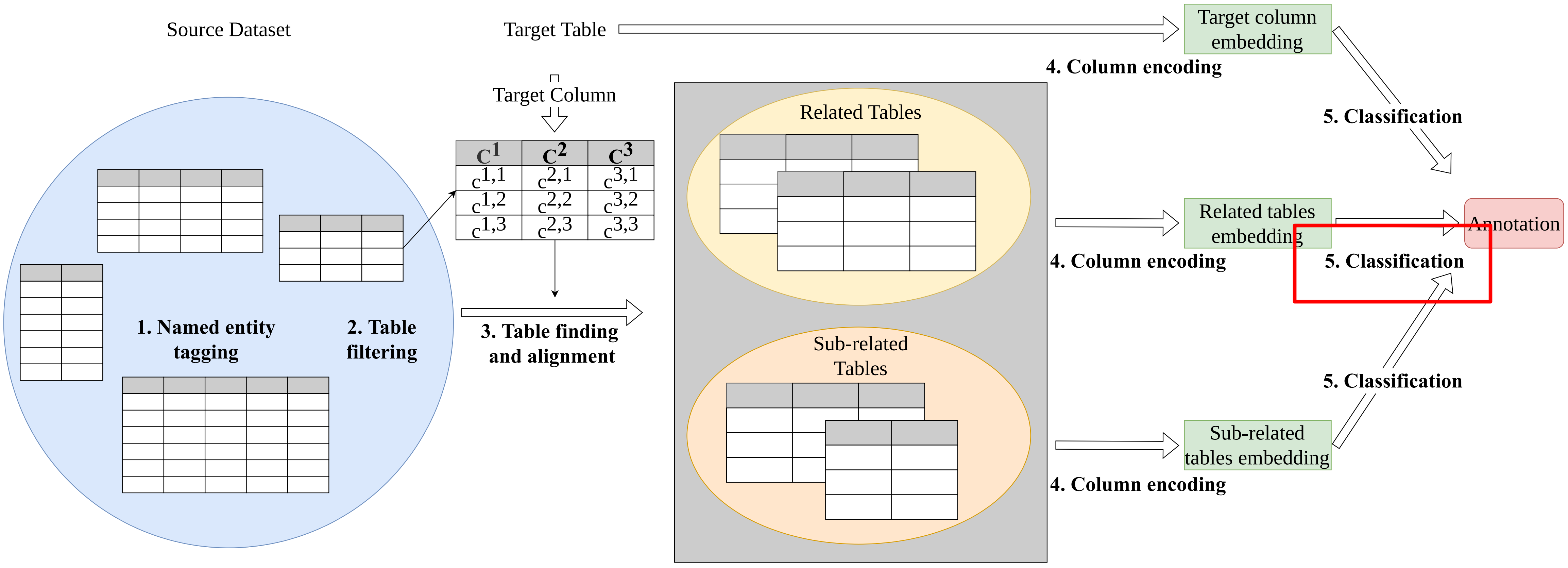


3. Methodology - Column Encoding

- The target column is encoded with BERT solely.
- The aligned columns in related tables and sub-related tables are encoded separately with BERT.
- The tokens are allocated fairly to each related table (or sub-related table).



3. Methodology



3. Methodology - Classification

- The embeddings of the target column, related tables, and sub-related tables are passed to three corresponding classification modules.
- Each classification module contains two layers: dropout and linear layers.
- The generated output embeddings are combined with learnable weights:

$$a_i^t = \alpha * \hat{v}_i^t + \beta * \hat{r}_i^t + \gamma * \hat{x}_i^t$$

- We use the cross-entropy loss as the loss function.

Outline

- Background and Motivation
- Definitions
- Methodology
- Experiments
- Summary

4. Experiments – Datasets and Metrics

- Datasets:

	WebTables	Semtab2019
# semantic types	78	275
# tables	32262	3045
# annotated columns	74141	7603
Avg. # rows	20.0	69.0
Avg. # columns	2.3	4.5
Avg. # annotated columns	2.3	2.5

- Metrics:
  - Support-weighted F1: weighted support of per type F1 scores
  - Macro average F1: average of per type F1 scores (emphasize on long-tail types)

4. Experiments – Main Results

- RECA outperforms all the state-of-the-arts in terms of the F1 scores.

	Semtab2019 dataset		WebTables dataset	
Model names	Support-weighted F1	Macro average F1	Support-weighted F1	Macro average F1
Sherlock [15]	0.646 ± 0.006	0.440 ± 0.009	0.844 ± 0.001	0.670 ± 0.010
TaBERT [35]	0.768 ± 0.011	0.413 ± 0.019	0.896 ± 0.005	0.650 ± 0.011
TABBIE [16]	0.799 ± 0.013	0.607 ± 0.011	0.929 ± 0.003	0.734 ± 0.019
DODUO [30]	0.820 ± 0.009	0.630 ± 0.015	0.928 ± 0.001	0.742 ± 0.012
RECA	**0.853** ± 0.005	**0.674** ± 0.007	**0.937** ± 0.002	**0.783** ± 0.014

4. Experiments – Ablation Study

- We conducted ablation study on RECA:
  - RECA target only: only encode the target column
  - RECA w/o re: encode both target column and aligned columns in sub-related tables
  - RECA w/o sub: encode both target column and aligned columns in related tables
- Performance drops on macro average F1 scores are greater than that on support-weighted F1 scores – incorporating inter-table context can improve the annotation quality on less-populated semantic types.

	Semtab2019 dataset		WebTables dataset	
Model names	Support-weighted F1	Macro average F1	Support-weighted F1	Macro average F1
RECA *target only*	0.808 ± 0.017	0.586 ± 0.039	0.911 ± 0.001	0.688 ± 0.014
RECA *w/o re*	0.836 ± 0.012	0.641 ± 0.037	0.927 ± 0.001	0.748 ± 0.024
RECA *w/o sub*	0.848 ± 0.009	0.650 ± 0.019	0.936 ± 0.002	0.774 ± 0.011
RECA	**0.853** ± 0.005	**0.674** ± 0.007	**0.937** ± 0.002	**0.783** ± 0.014

4. Experiments - Learning and Input Data Utilization

- RECA is efficient in utilizing the learning data and the input data.

Learning data utilization

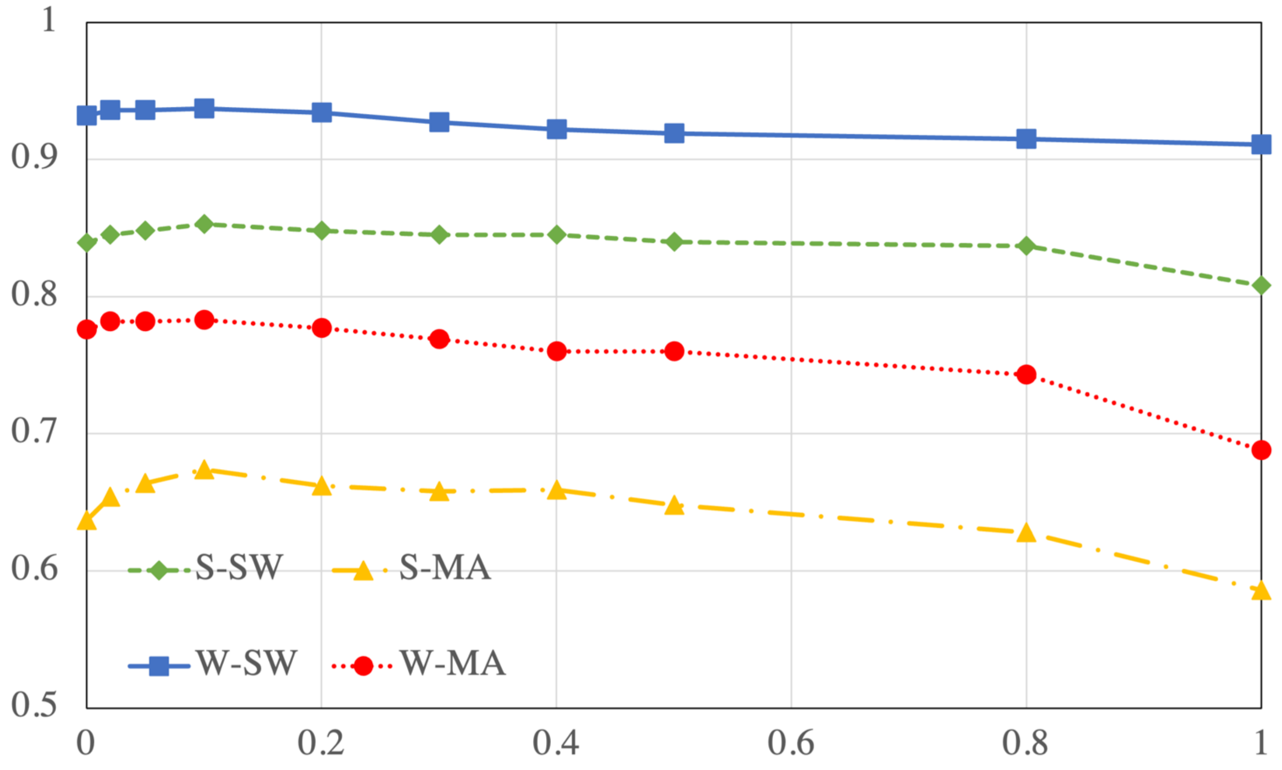
Datasets	[%]	Support-weighted F1	Macro average F1
Semtab2019	25	0.697 ± 0.041	0.442 ± 0.074
Semtab2019	50	0.792 ± 0.020	0.566 ± 0.045
Semtab2019	75	0.820 ± 0.021	0.631 ± 0.047
Semtab2019	100	**0.853** ± 0.005	**0.674** ± 0.007
WebTables	25	0.909 ± 0.002	0.680 ± 0.008
WebTables	50	0.924 ± 0.004	0.738 ± 0.019
WebTables	75	0.930 ± 0.002	0.772 ± 0.013
WebTables	100	**0.937** ± 0.002	**0.783** ± 0.014

Input data utilization

Datasets	Max	Support-weighted F1	Macro average F1
Semtab2019	8	0.540 ± 0.009	0.319 ± 0.010
Semtab2019	16	0.654 ± 0.013	0.436 ± 0.006
Semtab2019	32	0.728 ± 0.010	0.507 ± 0.020
Semtab2019	128	0.816 ± 0.017	0.620 ± 0.033
Semtab2019	256	0.851 ± 0.011	0.662 ± 0.024
Semtab2019	512	**0.853** ± 0.005	**0.674** ± 0.007
WebTables	8	0.907 ± 0.004	0.737 ± 0.011
WebTables	16	0.923 ± 0.002	0.762 ± 0.011
WebTables	32	0.931 ± 0.002	0.780 ± 0.010
WebTables	128	**0.937** ± 0.002	**0.783** ± 0.014
WebTables	256	0.936 ± 0.003	**0.783** ± 0.020
WebTables	512	0.936 ± 0.001	0.780 ± 0.011

4. Experiments – Parameter Sensitivity

- RECA achieves stable performance when the Jaccard threshold is in the range of [0, 0.3].



- S-SW and S-MA stand for the support-weighted and macro average F1 scores on the Semtab2019 dataset; W-SW and W-MA stand for the support-weighted and macro average F1 scores on the WebTables dataset.

Outline

- Background and Motivation
- Definitions
- Methodology
- Experiments
- Summary

5. Summary

- We propose RECA for column semantic type annotation. RECA extracts and leverages inter-table context to enhance the annotation quality of the target column, thus resolving the wide table issue.
- We define a novel named entity schema for RECA to efficiently align related and sub-related tables, which resolves the difficulty of incorporating inter-table context.
- We conduct extensive experiments on two real-world web table datasets to show that RECA outperforms all the state-of-the-art methods. The result demonstrates the effectiveness of utilizing the inter-table context to annotate column semantic types accurately.
- We show that RECA is data efficient and learning efficient, since it requires shorter input token sequences and fewer training data to achieve high performance.