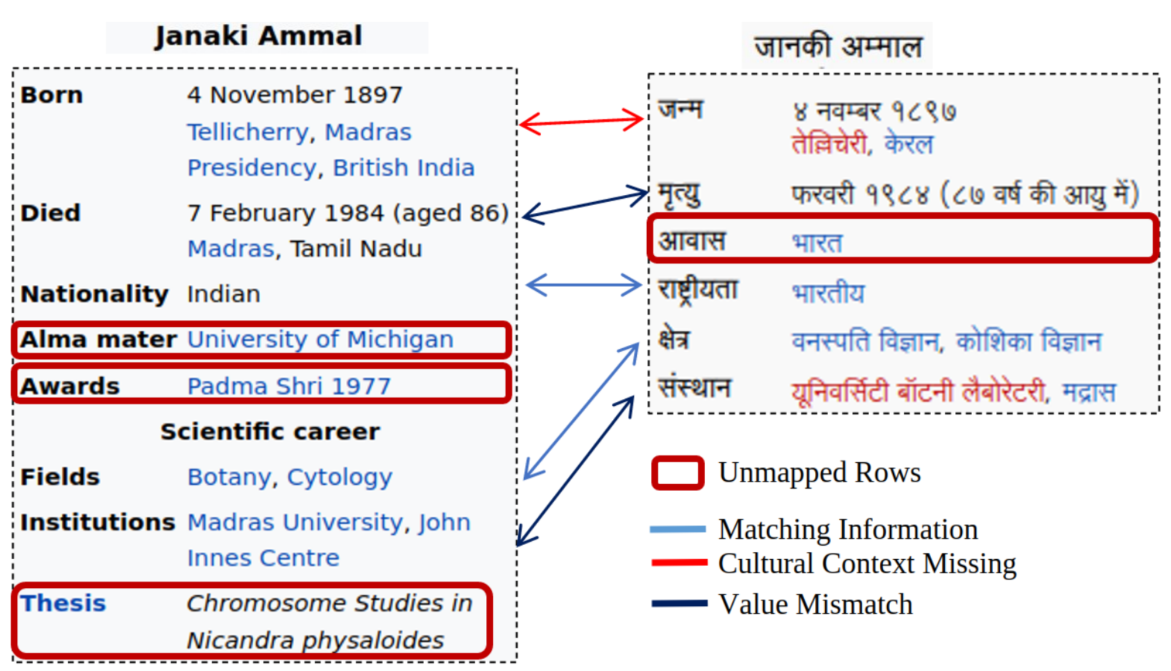
Information Synchronization across Multilingual Semi-structured Tables

Siddharth Khincha[1], Chelsi Jain[2], Vivek Gupta[3†], Tushar Kataria[3†], Shuo Zhang[4]
[1]IIT Guwahati, [2]CTAE, Udaipur, [3]University of Utah[2], [4]Bloomberg,

1. Information Mismatch in Tables Across Languages

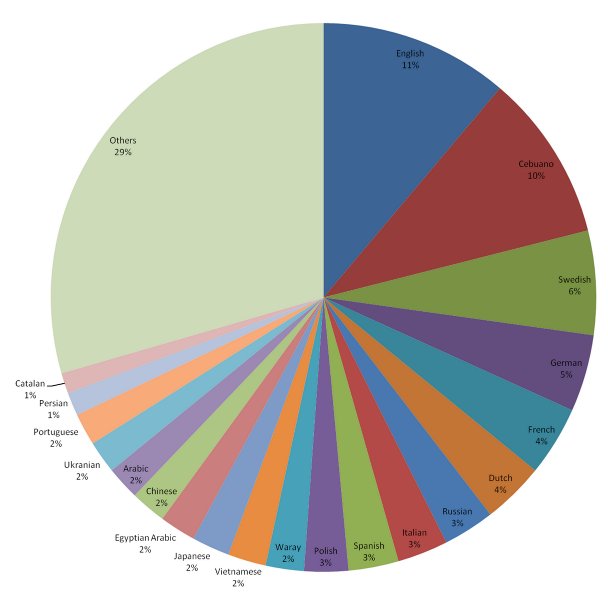


English Table Hindi Table

- Janaki Ammal Infoboxes: English (right) vs. Hindi (left). Hindi lacks "British Rule of India" context.
- Value mismatches: (a) Hindi table doesn't state Died key's state. (b) Institution values differ - Hindi mentions "residence," English doesn't.
- Missing keys in Hindi table: "Thesis," "Awards," and "Alma Mater." Neither mentions parents, early education, or honors.

2. Problem Magnitude

- Articles in More than 300 languages.
- English has the most significant Wikipedia covering 23% (11%) of total pages (articles).



- Most users' edits (76%) are also done in English Wikipedia.

3. Our Contributions

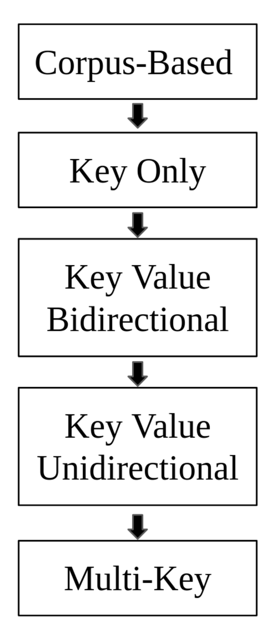
1. InfoSyncDataset
   - 100K entity-centric wikipedia Infoboxes table across 14 languages
   - Approximately 3.5K human annotated table alignment pairs
2. Proposed a two-step approach as a solution, include Information
   - **Alignment** to mapped similar rows
   - **Update** update missing/outdated rows for aligned tables

   across multilingual entity centric tables

4. Dataset Details :- Language and category selection

1. Languages
   - Languages are selected to cover all the continents.
   - 4 low resource Hindi(hi), Cebuano(ceb), , Turkish(tr), and Afrikaans(ak)
   - 7 medium resource German(de), Korean(ko), Russian(ru), Arabic(ar), Chinese(zh), Swedish(sv),Dutch(nl)
   - 3 high resource - English(en), French(fr), Spanish(es)
2. Entities
   - Each Entity selected contains an Infobox in at least 5 languages
3. Categories Selection
   - 21 simple, diverse, and popular topics: Airport, Album,Animal, Athlete, Book, City, College, Company,Country, Food, Monument, Movie Musician, Nobel, Painting, Person, Planet, Shows, and Stadiums.

5. Method: Alignment



Corpus-based : Align rows based on keys using their cosine similarity across a category using majority voting.

Key-only : This module aligns rows with ***key*** similarity score greater than a threshold value, only if they are mutually most similar keys

Key value bidirectional : This module aligns rows with ***key+value*** similarity score greater than a threshold value, only if they are mutually most similar rows.

Key value unidirectional: This module aligns rows with ***key+value*** similarity greater than a threshold. They **do not** have to be mutually most similar.

Multi-key : This module considers the case where one row from table needs to be mapped to multiple rows in the second table. It is valid multi-key alignment when the merge value-combination similarity score exceeds that of the most similar key.

6. Method: Rule-Based Update

P.R.	Rule Name	Logical Rule $\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}$ L $\mapsto$ R	Update Type
1	Row Transfer	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=0$ $\mapsto T_y \cup tr_x^y(\mathrm{R}_{T_x}) \bigwedge \mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x}; tr_x^y(\mathrm{R}_{T_x}))=1$	Row Addition
2	Multi-Match	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}(\sum_{\mathrm{R}_{T_y}}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y}))>1$ $\mapsto \{T_y \setminus \cup_{(\forall \mathrm{R}_{T_y}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1)}\mathrm{R}_{T_y}\} \bigcup tr_x^y(\mathrm{R}_{T_x}) \bigwedge \mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x}; tr_x^y(\mathrm{R}_{T_x}))=1$	Row Delete
3	Time-based	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1 \bigwedge(\mathrm{isTime}(\mathrm{R}_{T_x},\mathrm{R}_{T_y})=1)$ $\bigwedge(\mathrm{exTime}(\mathrm{R}_{T_x})>\mathrm{exTime}(\mathrm{R}_{T_y}))\mapsto \mathrm{R}_{T_y}\leftarrow tr_x^y(\mathrm{R}_{T_x})$	Value Substitute
4	Positive Trend or	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y},\mathrm{PosTrend})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1 \bigwedge \mathrm{exKey}(\mathrm{R}_{T_x})\in \mathrm{PosTrend}$ $\bigwedge \mathrm{R}_{T_x}>\mathrm{R}_{T_y}\mapsto \mathrm{R}_{T_y}\leftarrow \mathrm{R}_{T_x}$	Value Substitute
	Negative Trend	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y},\mathrm{NegTrend})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1 \bigwedge \mathrm{exKey}(\mathrm{R}_{T_x})\in \mathrm{NegTrend}$ $\bigwedge \mathrm{R}_{T_x}<\mathrm{R}_{T_y}\mapsto \mathrm{R}_{T_y}\leftarrow \mathrm{R}_{T_x}$	Value Substitute
5	Append Value	$\mathrm{R}_{T_x}=\mathrm{V}\bigwedge\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1\bigwedge\lvert\mathrm{R}_{T_x}[k]\rvert>\lvert\mathrm{R}_{T_y}[k]\rvert$ $\mapsto \forall_{(v\in \mathrm{R}_{T_x}[k]\,\wedge\,\notin tr_x^y(\mathrm{R}_{T_x}[k]))}\mathrm{R}_{T_y}\leftarrow \mathrm{R}_{T_y}\cup tr_x^y(v)$	Value Addition
6	HR to LR	$(T_x,T_y)\in(HR,LR)\bigwedge\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1$ $\bigwedge tr_x^{en}(\mathrm{R}_{T_x})\neq tr_y^{en}(\mathrm{R}_{T_y})\mapsto \mathrm{R}_{T_y}\leftarrow tr_x^y(\mathrm{R}_{T_x})$	Value Substitute
7	# Rows	$\lvert T_x\rvert >> \lvert T_y\rvert \bigwedge\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1\bigwedge tr_x^{en}(\mathrm{R}_{T_x})\neq tr_y^{en}(\mathrm{R}_{T_y})$ $\mapsto \mathrm{R}_{T_y}\leftarrow tr_x^y(\mathrm{R}_{T_x})$	Value Substitute
8	Rare Keys	$\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y},\mathrm{RarKeys})}\mathrm{Al}_{T_x}^{T_y}(\mathrm{R}_{T_x};\mathrm{R}_{T_y})=1\bigwedge \mathrm{tr}_x^{en}(\mathrm{R}_{t_x})\neq \mathrm{tr}_y^{en}(\mathrm{R}_{t_y})$ $\bigwedge\forall_{(\mathrm{R}_{T_x},\mathrm{R}_{T_y}}\lvert\mathrm{exKey}(\mathrm{R}_{T_x})\in\mathrm{RarKey}\rvert>\lvert\mathrm{exKey}(\mathrm{R}_{T_y})\in\mathrm{RarKey}\rvert\mapsto\mathrm{R}_{T_y}\leftarrow\mathrm{R}_{T_x}$	Value Substitute

7. Result: Alignment

Proposed similarity-based alignment method outperforms different multi-lingual baseline.

Method	Match				UnMatch			
	$T_{en}\leftrightarrow T_x$	$T_x\leftrightarrow T_y$	$T_{en}\overset{*}{\leftrightarrow}T_{hi}$	$T_{en}\overset{*}{\leftrightarrow}T_{zh}$	$T_{en}\leftrightarrow T_x$	$T_x\leftrightarrow T_y$	$T_{en}\overset{*}{\leftrightarrow}T_{hi}$	$T_{en}\overset{*}{\leftrightarrow}T_{zh}$
SimCSE	75.78	68.46	77.93	80.47	79.11	76.3	73.31	74.91
LaBSE	85.25	78.44	88.98	89.1	87.03	81.7	88.98	85.06
mBERT-mp	80.98	73.74	82.9	86.73	82.68	80.22	76.73	81.85
XLM-R	83.38	75.02	86.85	88.08	85.42	80.65	83.14	83.1
MPNet	82.85	78.63	86.08	87.58	84.2	83.45	83.14	83.76
distill mBERT	84.55	77.45	87.64	88.7	86.3	82.28	83.14	84.3
Our Approach								
Corpus-based	61.86	56.74	57.34	69.33	70.51	71.73	54.01	63.11
+ Key Only	70.41	62.14	73.4	74.67	73.85	73.52	62.49	66.23
+ Key-Val-Bi	87.71	84.2	90.07	93.04	89.51	85.52	85.06	89.2
+ Key-Val-Uni	87.89	84.33	90.34	93.12	89.52	85.42	85.16	88.62
+ Multi-Key	**87.91**	**84.36**	**90.14**	**92.8**	**89.3**	**85.46**	**84.98**	**88.15**

8. Result: Update

Our rule-based method efficiently updates a large number of rows, with the highest number of updates being in row transfers.

Rules	Gold		Live Set	Predicted	
	$T_{en}\rightarrow T_x$	$T_x\rightarrow T_y$		$T_{en}\rightarrow T_x$	$T_x\rightarrow T_y$
R1	20320	18055	4213	21246	17675
R2	648	502	207	1395	1852
R3	546	399	75	443	347
R4	142	151	4	120	147
R5	3507	2116	784	3193	1960
R6	5237	3047	332	5062	2891
R7	2748	1899	990	2732	1855
R8	25	77	5	29	82
Al	14967	9715	2851	14864	10657

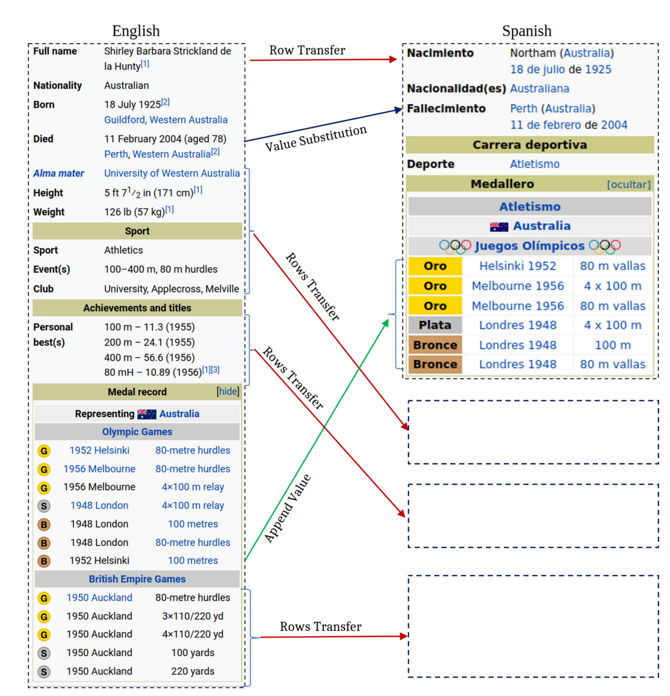
9. Human Assisted Wikipedia Updates

Human evaluator update the Wikipedia Infobox with our mehtod recommendation.

Type	Total	Accept	Reject
Row Transfer	461	368(79.82%)	93(20.17%)
Value Substitution	70	52(74.28%)	18(25.72%)
Append Value	72	46(63.88%)	26(36.12%)
Total	603	466 (77.28%)	136(22.72%)

Ln Pairs	Total	Accept	Reject
$T_{en}\rightarrow T_x$	204	161(78.92%)	43(21.07%)
$T_x\rightarrow T_y$	216	169(78.25%)	47(21.75%)
$T_x\rightarrow T_{en}$	183	136(74.31%)	47(25.68%)
Total	603	466(77.28%)	137(22.71%)

10. Example



11. Key Takeaways

1. Multilingual Tabular Information Synchronization is challenging problem.
2. Taking Wikipedia Infoboxes as our case study, we created InfoSync
3. A two-step sequential approach (a.) Alignment and (b.) Updation
   - Alignment method outperforms baseline with an F1-score > 0.85
   - The rule-based method received a 77.28 % approval rate on Wikipedia updates.