Annotating and Searching Web Tables Using Entities, Types and Relationships

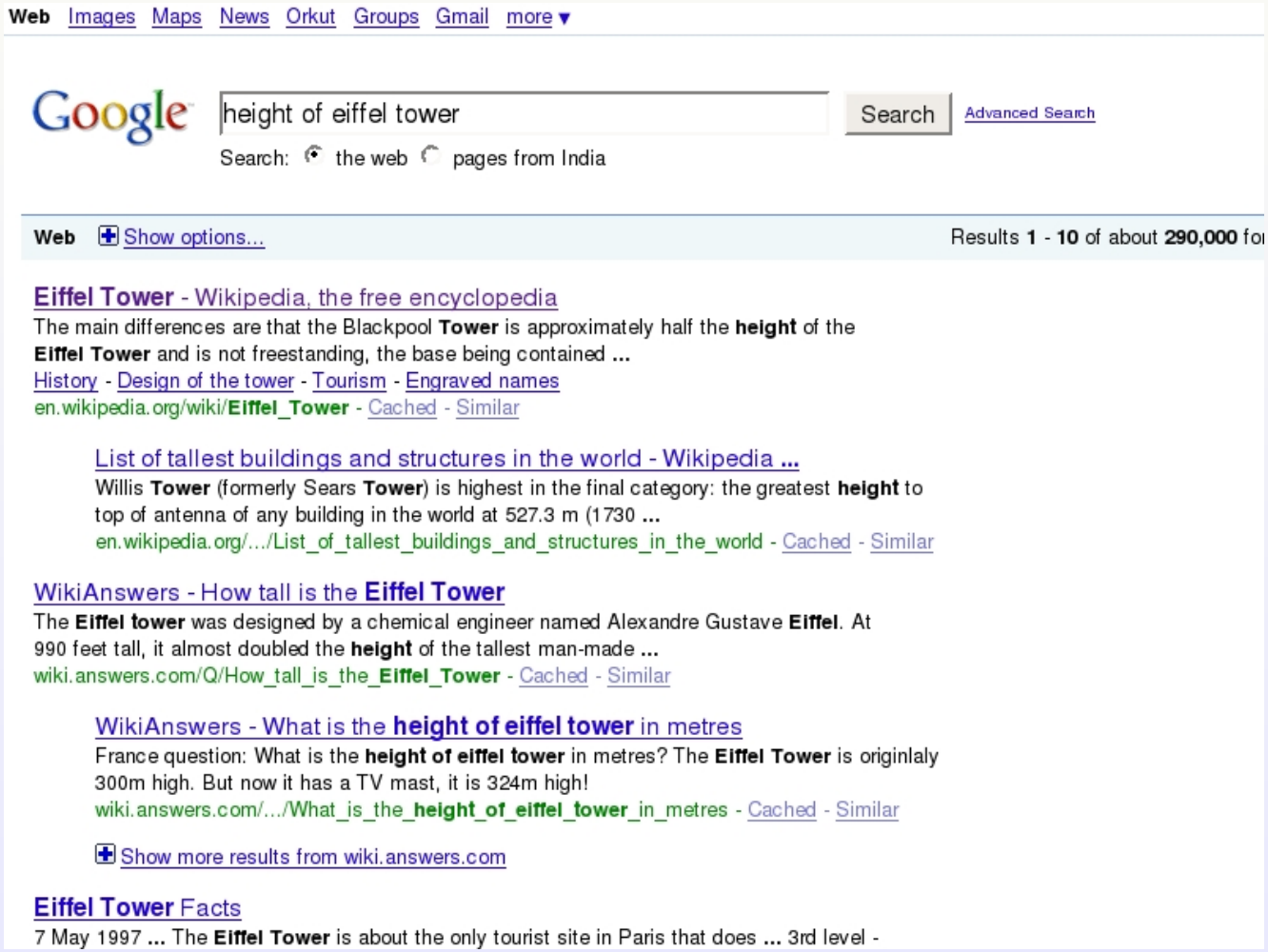
M.Tech. Project

Girija Limaye

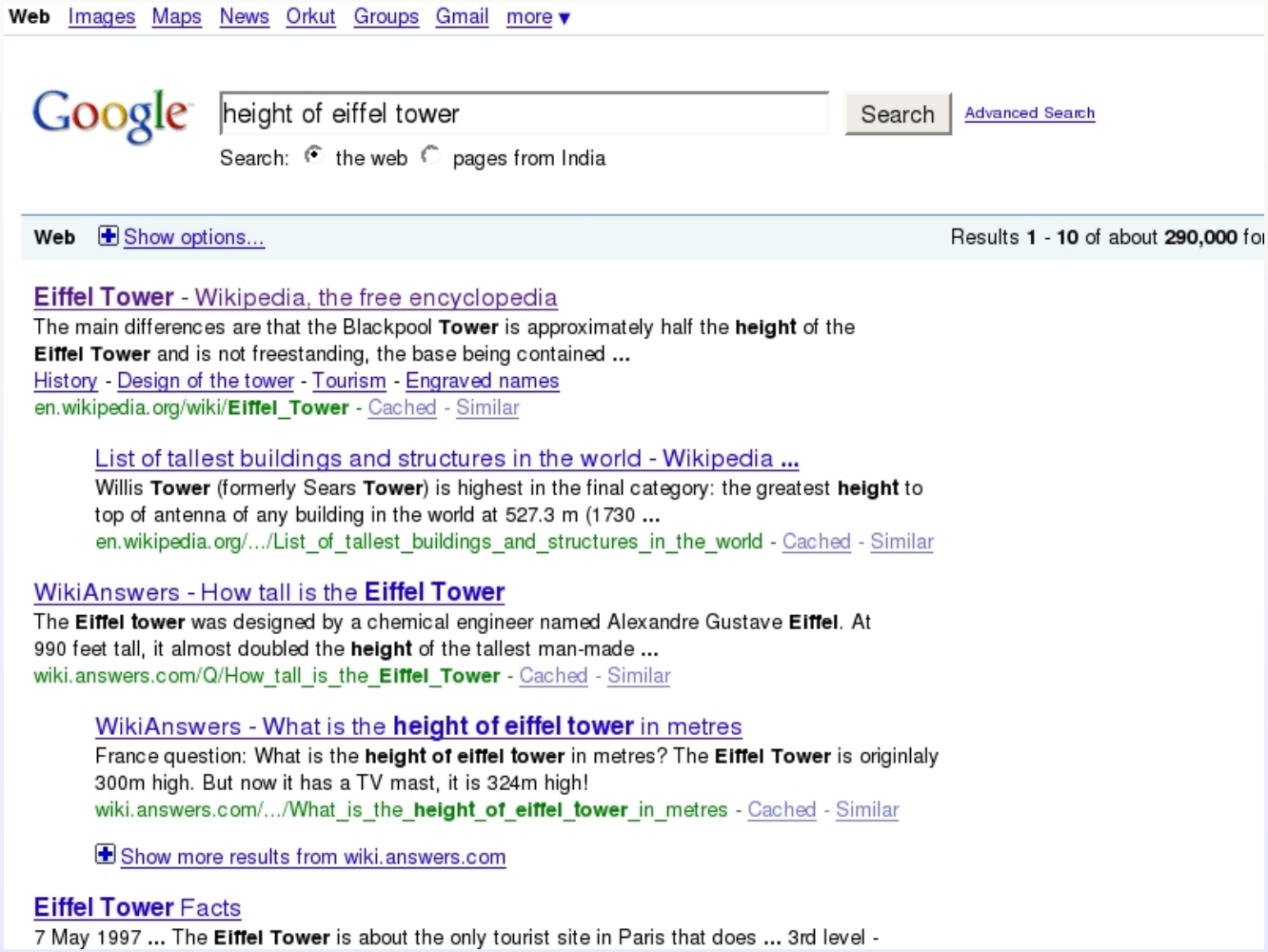
Under the guidance of

Prof. Sunita Sarawagi

Keyword based search

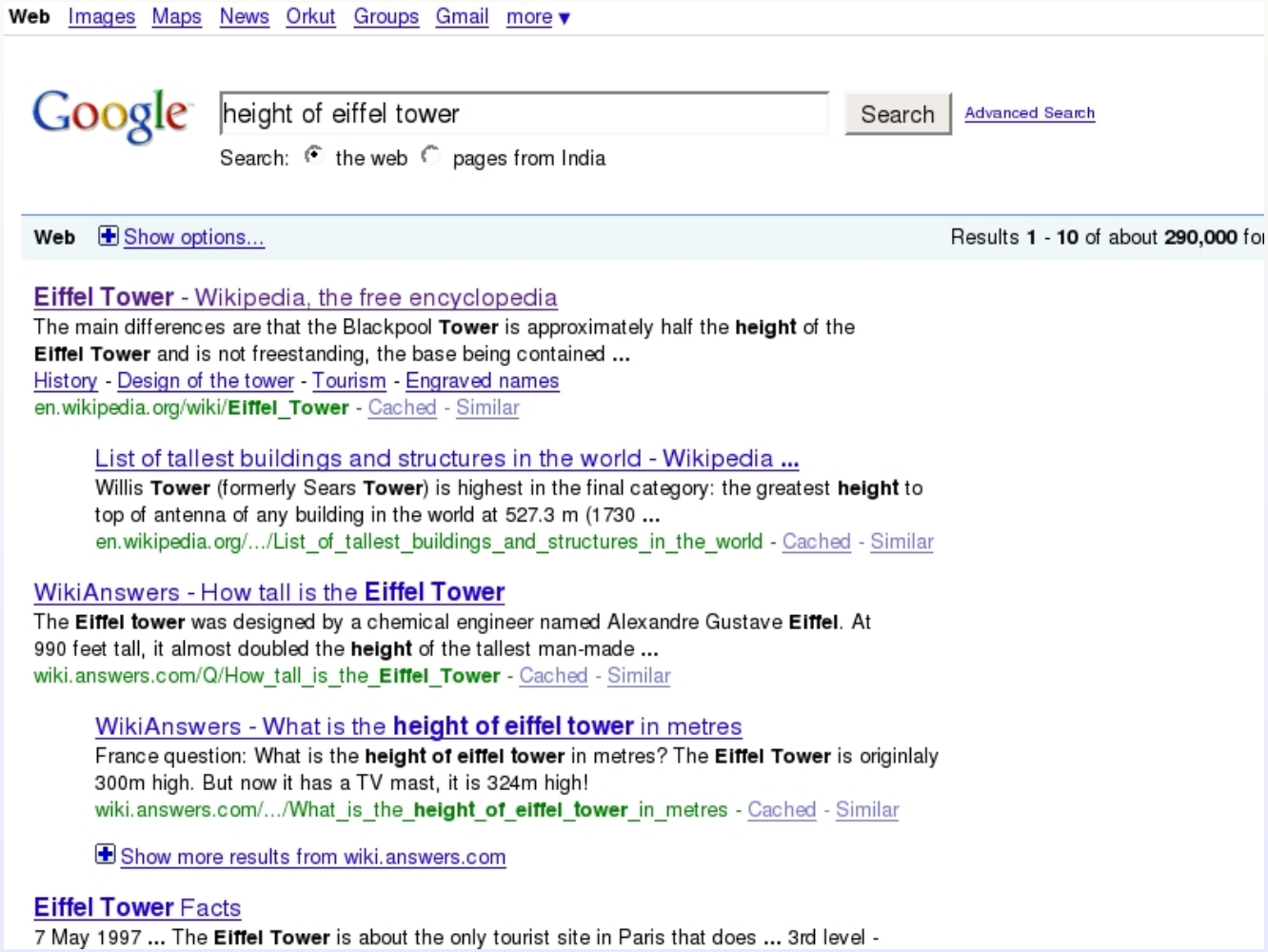


Keyword based search



Go through results

Keyword based search



Go through results

No precise answer

A web table with required information

record from	record to	Name and Location	Constructed	Height (m)	Height (ft)	Notes
c. 2700 BC	c. 2600 BC	Pyramid of Djoser, Egypt	c. 2700 BC	62	203	
c. 2600 BC	c. 2570 BC	Red Pyramid of Sneferu, Egypt	c. 2600 BC	105	345	
c. 2570 BC	c. AD 1311	Great Pyramid of Giza in Egypt	c. 2570 BC	146	481	By AD 1439, the Great Pyramid had eroded to a height of approximately 139 m (455 ft).
1311	1549	Lincoln Cathedral in England	1092–1311	160	525	The central spire was destroyed in a storm in 1549. While the reputed height of 525 ft (160 m) is doubted by A.F. Kendrick,[12] other sources[which?] agree on this height.
1549	1625	St. Olaf's Church in Tallinn, Estonia	1438–1519	159	522	The spire burnt down after a lightning strike in 1625 and was rebuilt several times. The current height is 123 m.
1625	1647	St. Mary's Church in Stralsund, Germany	1384–1478	151	495	The spire burnt down after a lightning strike in 1647. The current height is 104 m.
1647	1874	Strasbourg Cathedral in France	1439	142	469	
1874	1876	St. Nikolai in Hamburg, Germany	1846–1874	147	483	
1876	1880	Cathédrale Notre Dame in Rouen, France	1202–1876	151	495	
1880	1884	Cologne Cathedral in Germany	1248–1880	157	515	
1884	1889	Washington Monument in Washington D.C., United States	1884	169	555	
1889	1930	Eiffel Tower in Paris, France	1889	300	986	First structure to exceed 300 metres in height. The addition of a telecommunications tower in the 1950s brought the overall height to 324 m.
1930	1931	Chrysler Building in New York, United States	1928–1930	319	1,046	
1931	1967	Empire State Building in New York, United States	1930–1931	381	1,250	First building with 100+ stories. The addition of a pinnacle and antennas later increased its overall height to 1,472 ft/448.7 m.
1967	1975	Ostankino Tower in Moscow, Russia	1963–1967	537	1,762	Remains the tallest in Europe. Fire in 2000 led to extensive renovation.
1975	2007	CN Tower in Toronto, Canada	1973–1976	553	1,815	Remains the tallest in the Americas
2007	present	Burj Dubai in Dubai, United Arab Emirates	2004–2009	818	2,684	Current holder of world's tallest freestanding structure. Topped out at 818 m (2,684 ft).

Precise and instantaneous answer

Exploiting tables on the Web

- Is this a genuine relation?
- What is the table about?
- What do columns contain?
  - Do they have headers? *tower, film, . . .*
  - Non-informative? *name, title, . . .*
- What do the cells represent?
- Are columns similar?
  - {Namesake, God of small Things, White Tiger, . . .}
  - {Namesake, Hututu, Maachis, . . .}

An example - Ambiguity in cell text

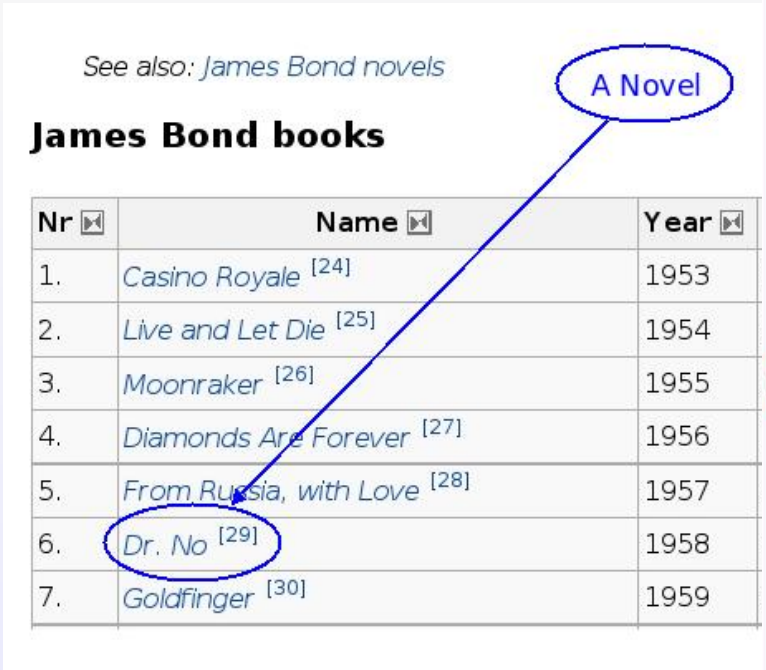
See also: James Bond novels

James Bond books

A Novel

Nr	Name	Year	
1.	*Casino Royale* [24]	1953	21 (o
2.	*Live and Let Die* [25]	1954	8
3.	*Moonraker* [26]	1955	11
4.	*Diamonds Are Forever* [27]	1956	7
5.	*From Russia, with Love* [28]	1957	2
6.	*Dr. No* [29]	1958	1
7.	*Goldfinger* [30]	1959	3

An example - Ambiguity in cell text



Another example - multiple specific types

2008

Title	Director
1920	Vikram Bhatt
Anamika	Anant Mahadevan
Aamir	Rajkumar Gupta
A Wednesday	Neeraj Pandey
Bachna Ae Haseeno	Siddharth Anand
Bhootnath	Vivek Sharma
Black and White	Subhash Ghai
Bombay to Bangkok	Nagesh Kukunoor
Chamku	Kabeer Kaushik
Coffee House	Gurbeer Garewal

Thus, we annotate Web tables

Thus, we annotate Web tables

with reference to a catalog

Dictionary / Ontology with set of types $\mathcal{T}$, set of entities $\mathcal{E}$, set of relationships $\mathcal{B}$

Thus, we annotate Web tables

with reference to a catalog

Dictionary / Ontology with set of types $\mathcal{T}$, set of entities $\mathcal{E}$, set of relationships $\mathcal{B}$

With the aim of

- Annotating cell text to an entity
- Annotating column to type(s)
- Finding relationship between columns

Annotated table

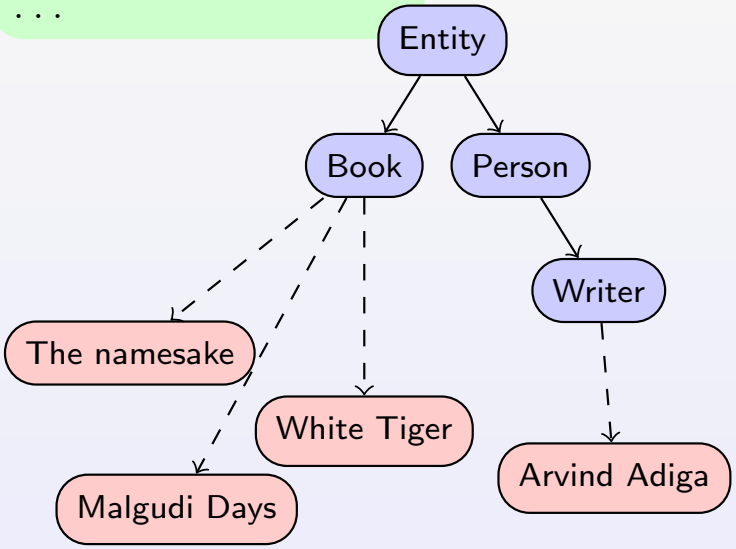
Title	Written by
Namesake	J. Lahiri
White Tiger	A. Adiga
...	...

Annotated table

author(Book, Person),

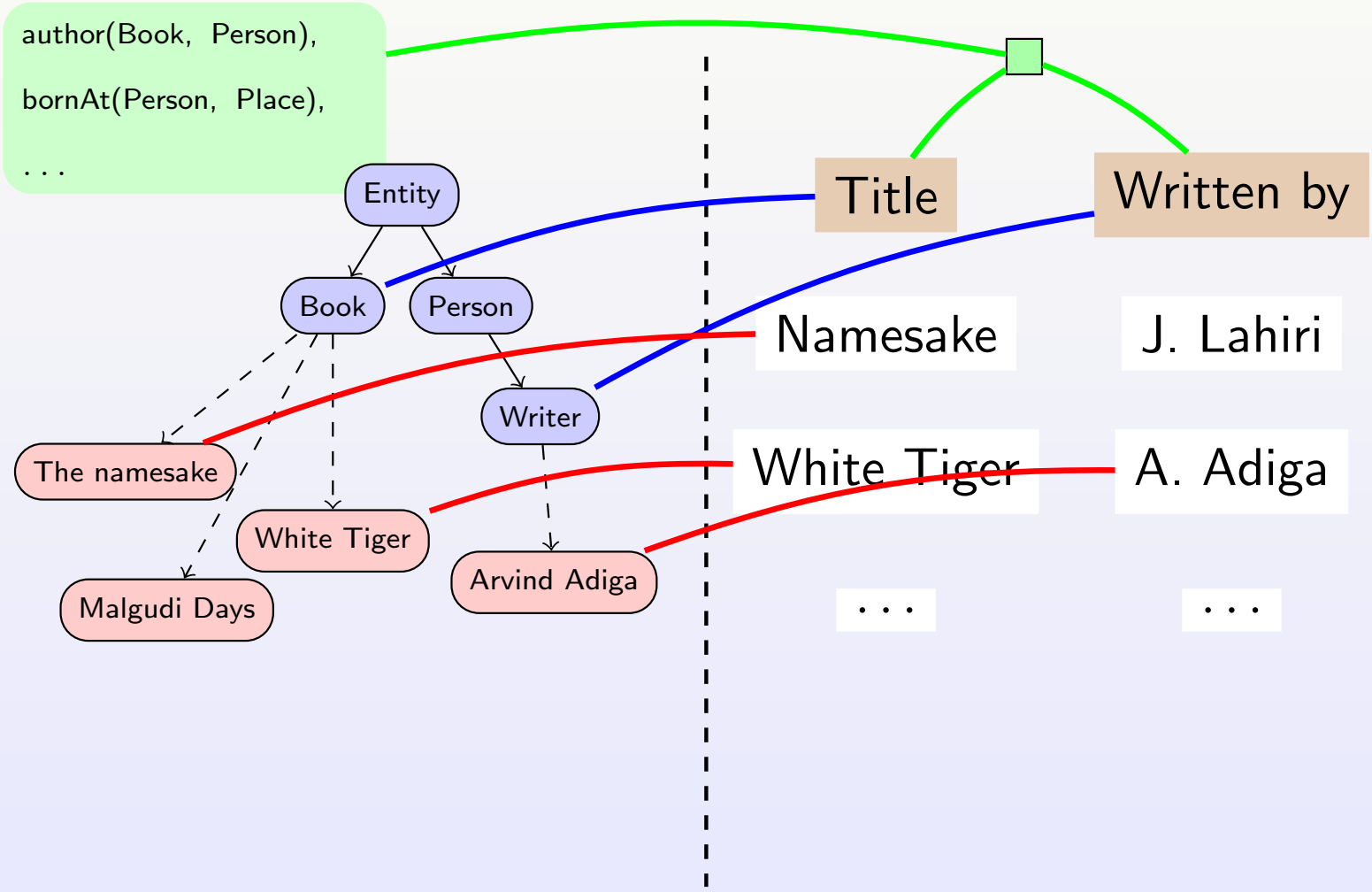
bornAt(Person, Place),

. . .



Title	Written by
Namesake	J. Lahiri
White Tiger	A. Adiga
. . .	. . .

Annotated table



Related Work

Web document annotation

- Linking to structured source
  - Learning to link with Wikipedia[MW08]
  - Collective annotation of Wikipedia entities in web text[KSRC09]

Querying on Web table corpus

- WebTables[CHW$^+$08]
  - Attribute based statistics
  - No interpretation of cells, columns
  - Plain text, header based answering

YAGO[SKW07][1]- Overview

Sources

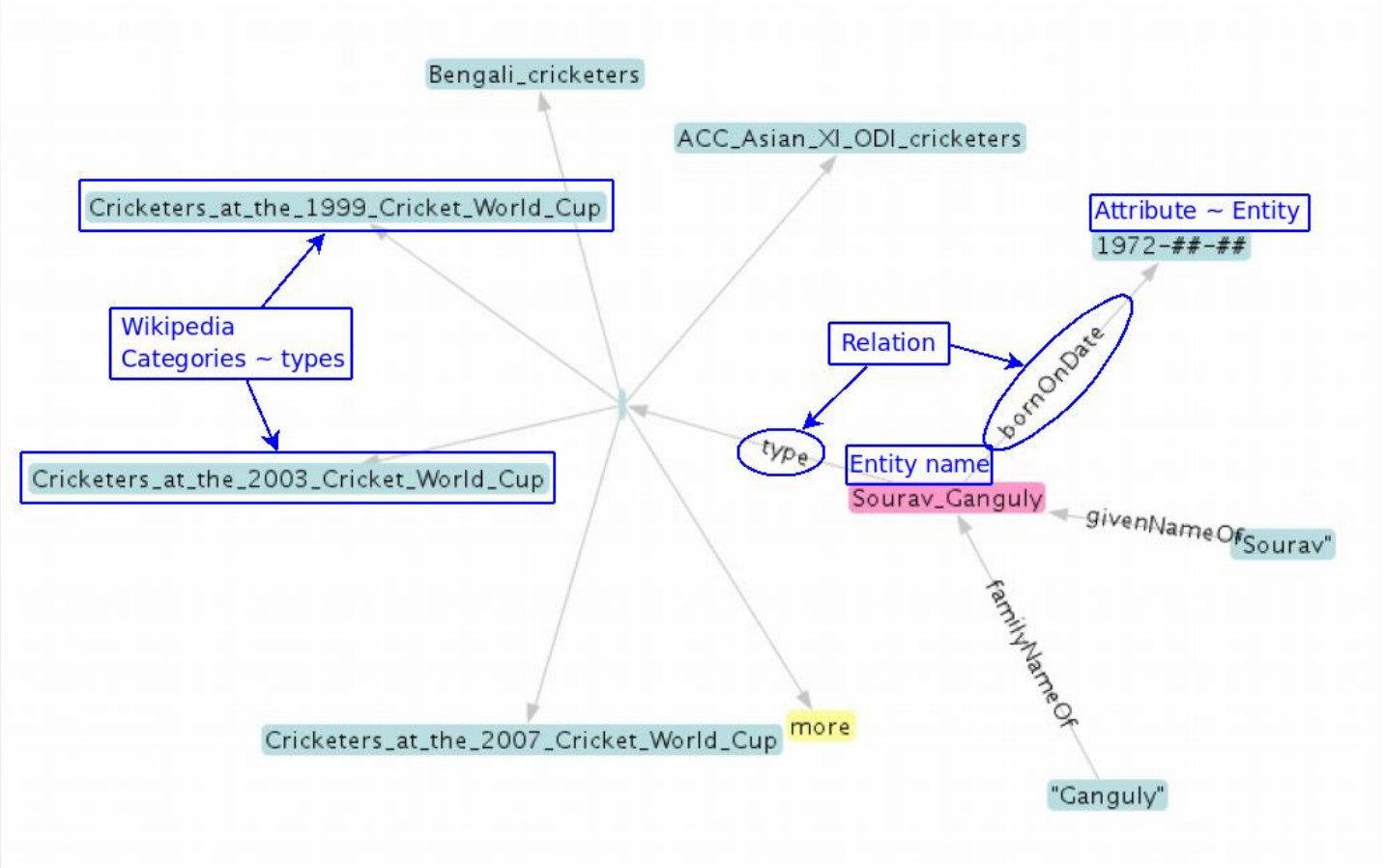
- Wikipedia article names = entities (`http://en.wikipedia.org/wiki/Sourav_Ganguly` → Sourav_Ganguly)
- Wikipedia categories = classes i.e. types (Marathi_Language_Singer)
- Wikipedia Category lists = TYPE relations (Lata_Mangeshkar TYPE Marathi_Language_Singer)
- WordNet - borrows concepts like singer, politician
- WordNet - borrows taxonomy

[1]Yet Another Great Ontology

YAGO - Example facts

fact identifier	Arg 1	relation	Arg 2
#1	"Einstein"	MEANS	Albert_Einstein
#2	"A. Einstein"	MEANS	Albert_Einstein
#3	"Albert Einstein"	MEANS	Albert_Einstein
#4	"Deutschland"	MEANS	Germany
#5	Albert_Einstein	TYPE	Physicist
#6	Physicist	SubClassOf	Scientist
#8	Albert_Einstein	HasWonPrize	NobelPrize
#9	#8	YEAR	1921

YAGO Terminologies[2]



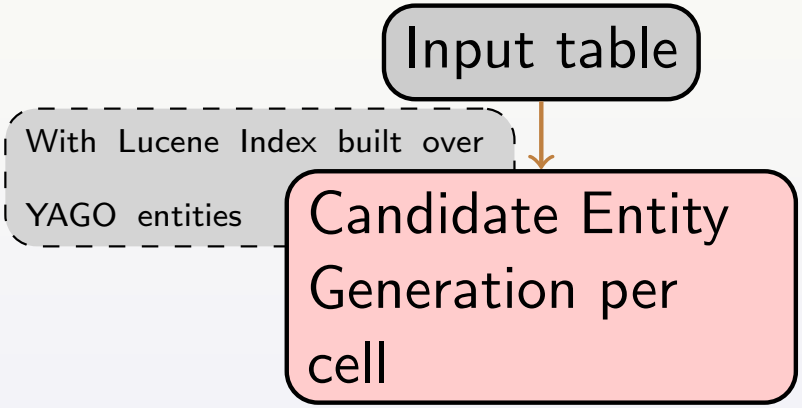
[2] Figure taken from http://uniat5401.ag5.mpi-sb.mpg.de:8180/webyago/html/graphViewer.html?entity=Sourav_Ganguly&n=5

System architecture

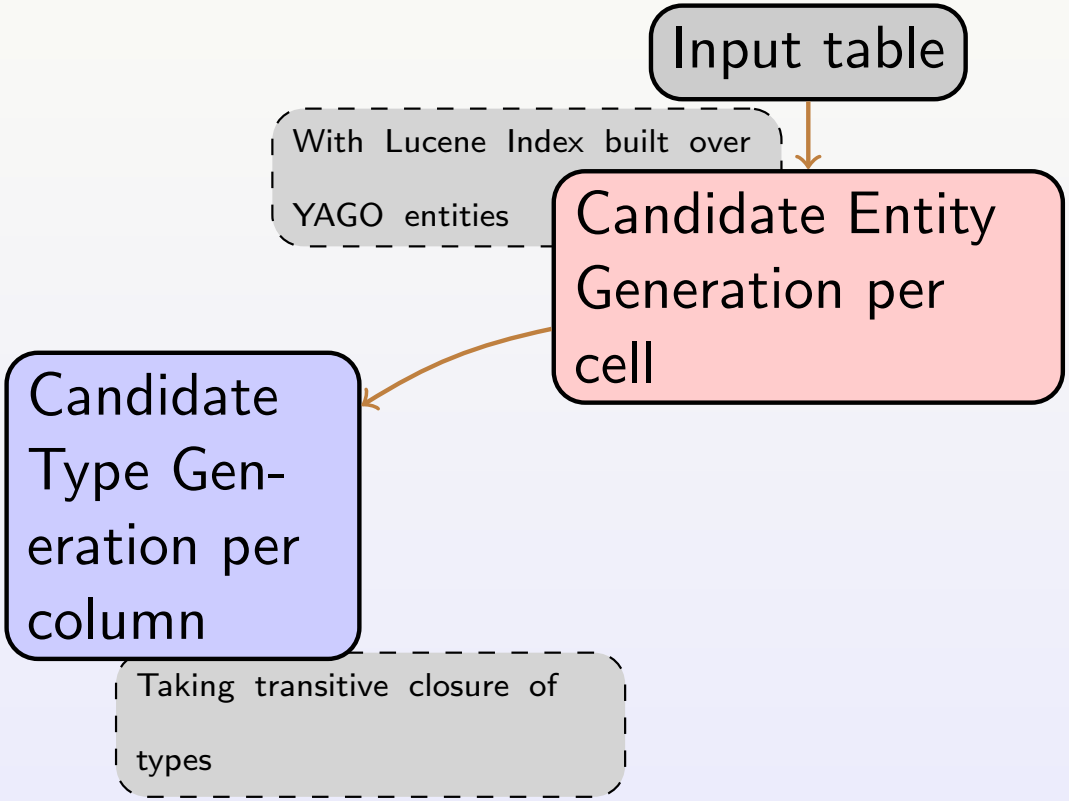
System architecture

Input table

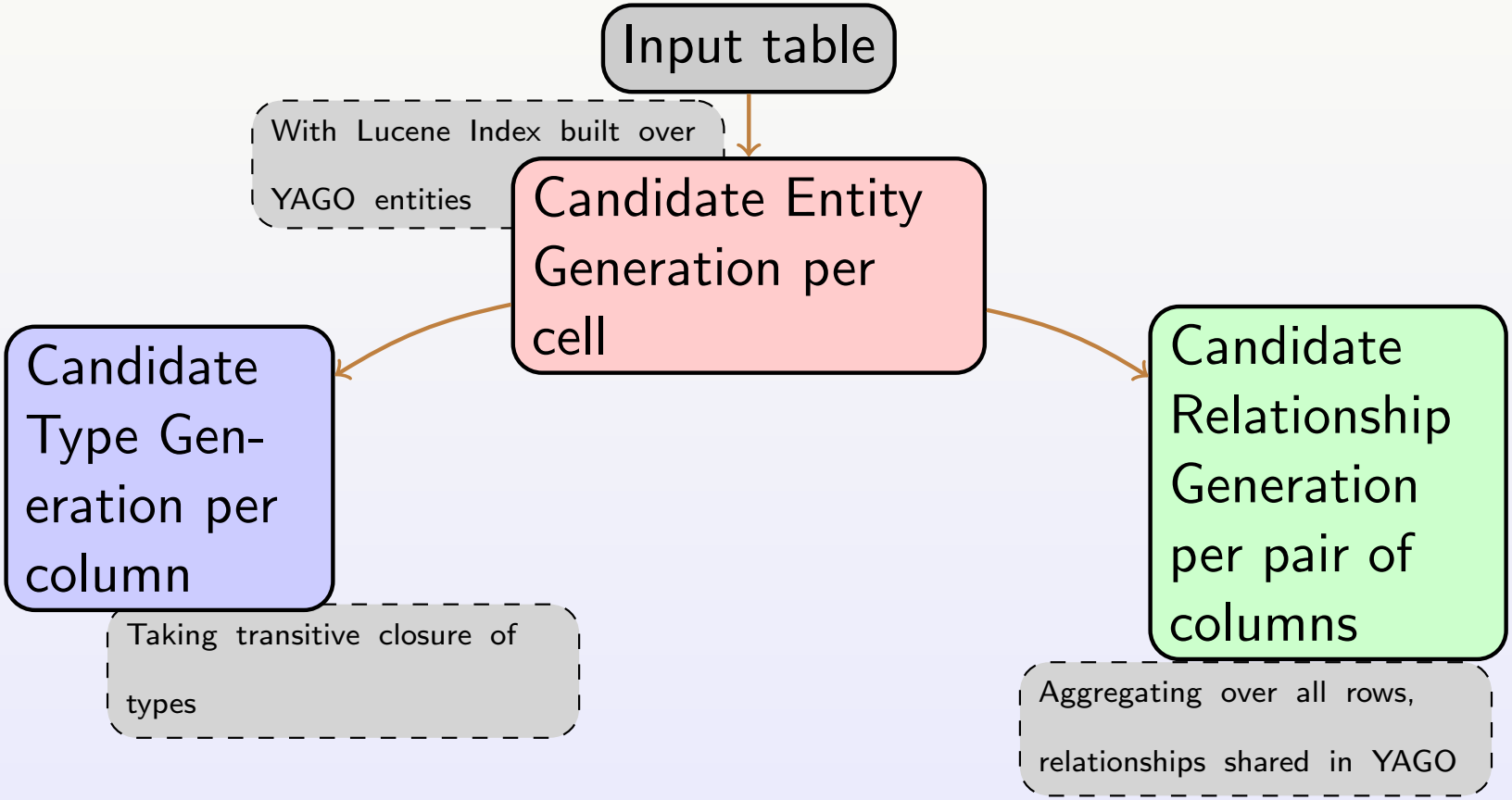
System architecture



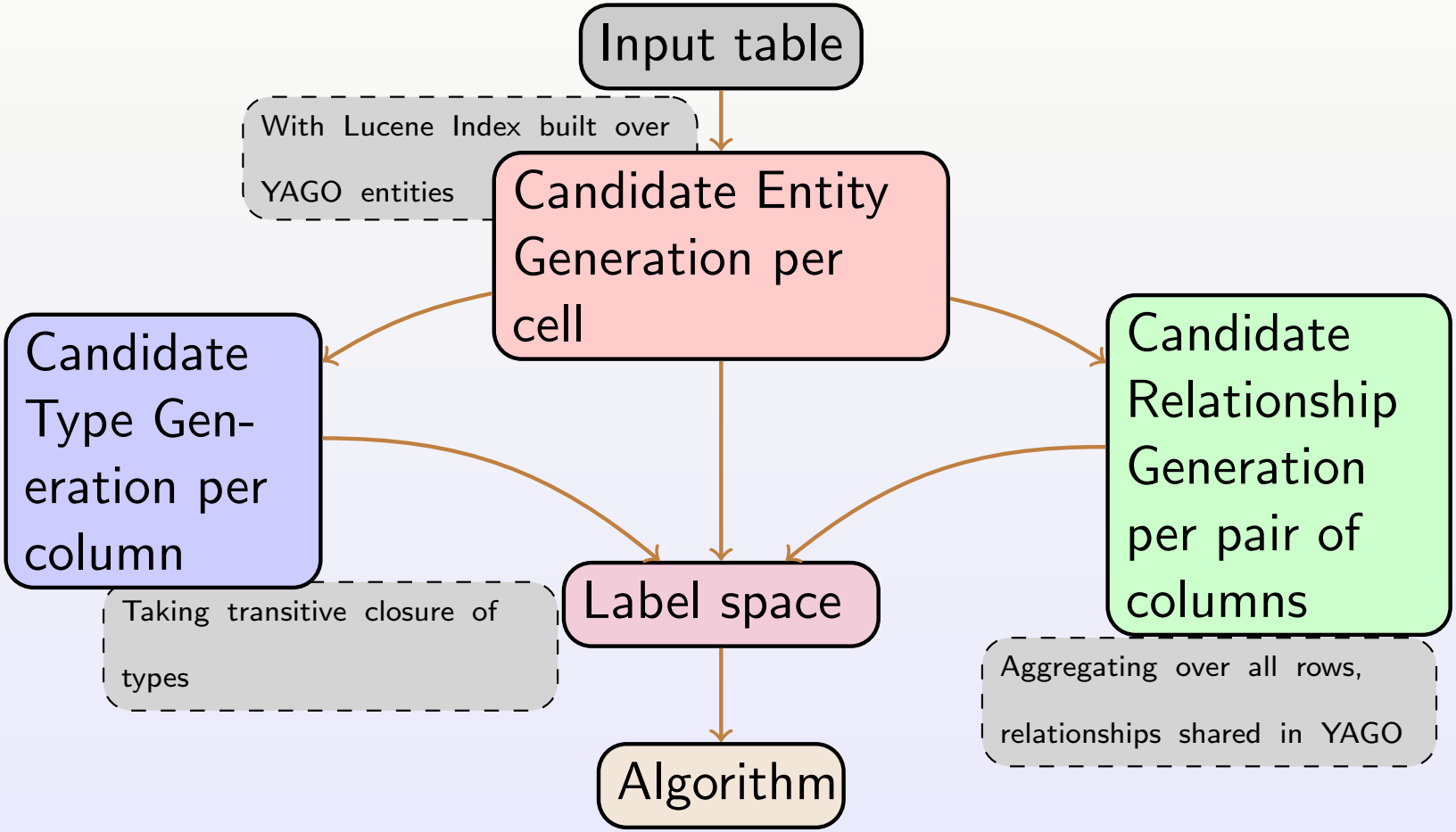
System architecture



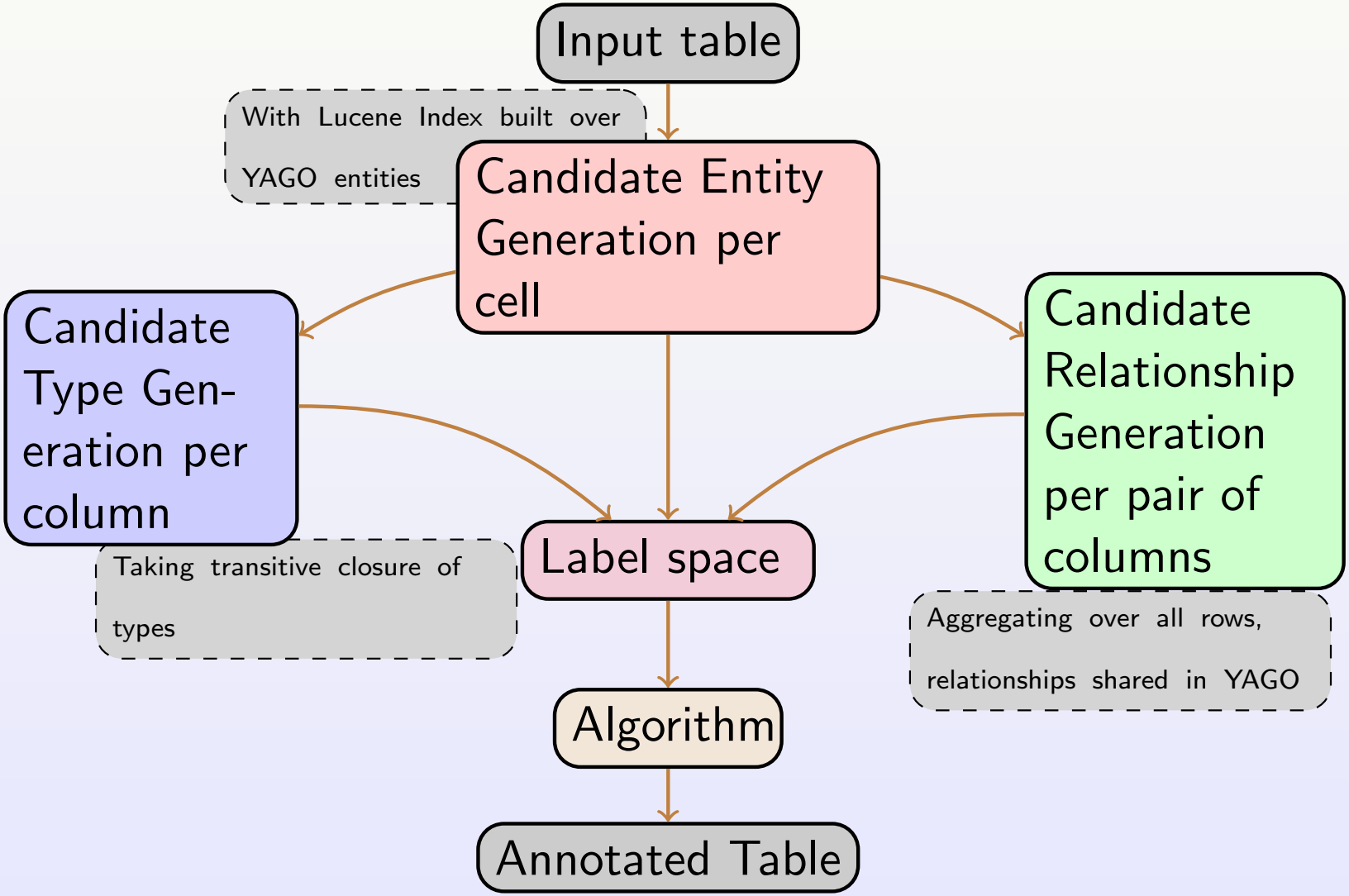
System architecture



System architecture



System architecture



An example

Cell Text	**Candidate Entities**	**Candidate Types for this cell**	**Candidate Types for entire column**
Naushad Ali	Naushad, NA	Filmfare award winners, . . .	Filmfare award winners, . . ., Entity
		Indian Muslims, . . .	Indian Muslims, . . .

An example

Cell Text	**Candidate Entities**	**Candidate Types for this cell**	**Candidate Types for entire column**
Naushad Ali	Naushad, NA	Filmfare award winners, . . .	Filmfare award winners, . . ., Entity
		Indian Muslims, . . .	Indian Muslims, . . .
	Naushad_Ali _(cricketer), NA	Pakistani cricketor, . . .	Pakistani cricketor, . . . Cricketor, Athlet, . . .

An example

Cell Text	**Candidate Entities**	**Candidate Types for this cell**	**Candidate Types for entire column**
Naushad Ali	Naushad, NA	Filmfare award winners, . . .	Filmfare award winners, . . ., Entity
		Indian Muslims, . . .	Indian Muslims, . . .
	Naushad_Ali _(cricketer), NA	Pakistani cricketor, . . .	Pakistani cricketor, . . . Cricketor, Athlet, . . .
Rajesh Roshan	Rajesh_Roshan, NA	Filmfare award winners, . . .	Filmfare award winners, . . .,
		Indian film score composers, . . .	Indian film score composers, . . . Composer, Musician, . . ., NA

Baseline Algorithms - 1

LCA based annotator

For a column c,

- Set of all types T such that per cell (r, c) has at least one candidate entity E and $E \in^{+} T$
- AND no T' such that $T' \subseteq T$ is in the set

Baseline Algorithms - 1

LCA based annotator

For a column c,

- Set of all types T such that per cell (r, c) has at least one candidate entity E and $E \in^{+} T$
- AND no T' such that $T' \subseteq T$ is in the set
- Choose entities which are instances of chosen types

Baseline Algorithms - 2

Majority based annotator

Define vote(T) =

$$\left|\left\{E : T \in \bigcup_{E \in \mathcal{E}_{r,c}} \mathcal{T}(E)\right\}\right|$$

For a column c,

- Set of all types T Ruth that vote(T) is at least 50%
- AND no T' such that $T' \subseteq T$ is in the set

Baseline Algorithms - 2

Majority based annotator

Define vote(T) =

$$\left|\left\{E : T \in \bigcup_{E \in \mathcal{E}_{r,c}} \mathcal{T}(E)\right\}\right|$$

For a column c,

- Set of all types T Ruth that vote(T) is at least 50%
- AND no T' such that $T' \subseteq T$ is in the set
- Choose entities independently per cell (r, c)

Failure cases : LCA

Missing links

Over-generalization

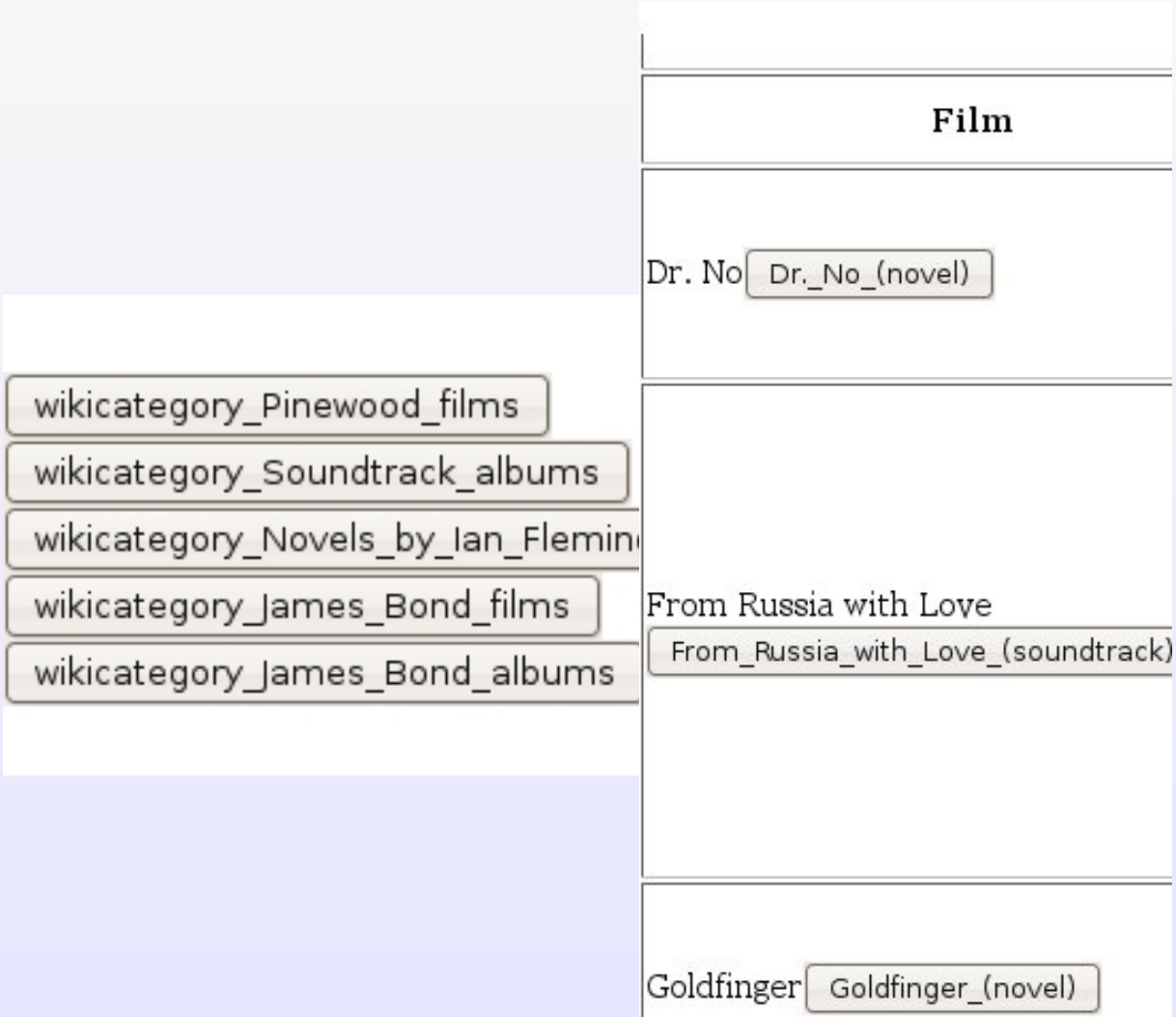
wordnet_entity_100001740

Title	
1. The Secret of the Old Clock	The_Secret_of_the_Old_Clock
2. The Hidden Staircase	The_Hidden_Staircase
3. The Bungalow Mystery	The_Bungalow_Mystery
4. The Mystery at Lilac Inn	The_Mystery_at_Lilac_Inn

Failure cases : Majority

Ambiguity in Cell text

Low Precision

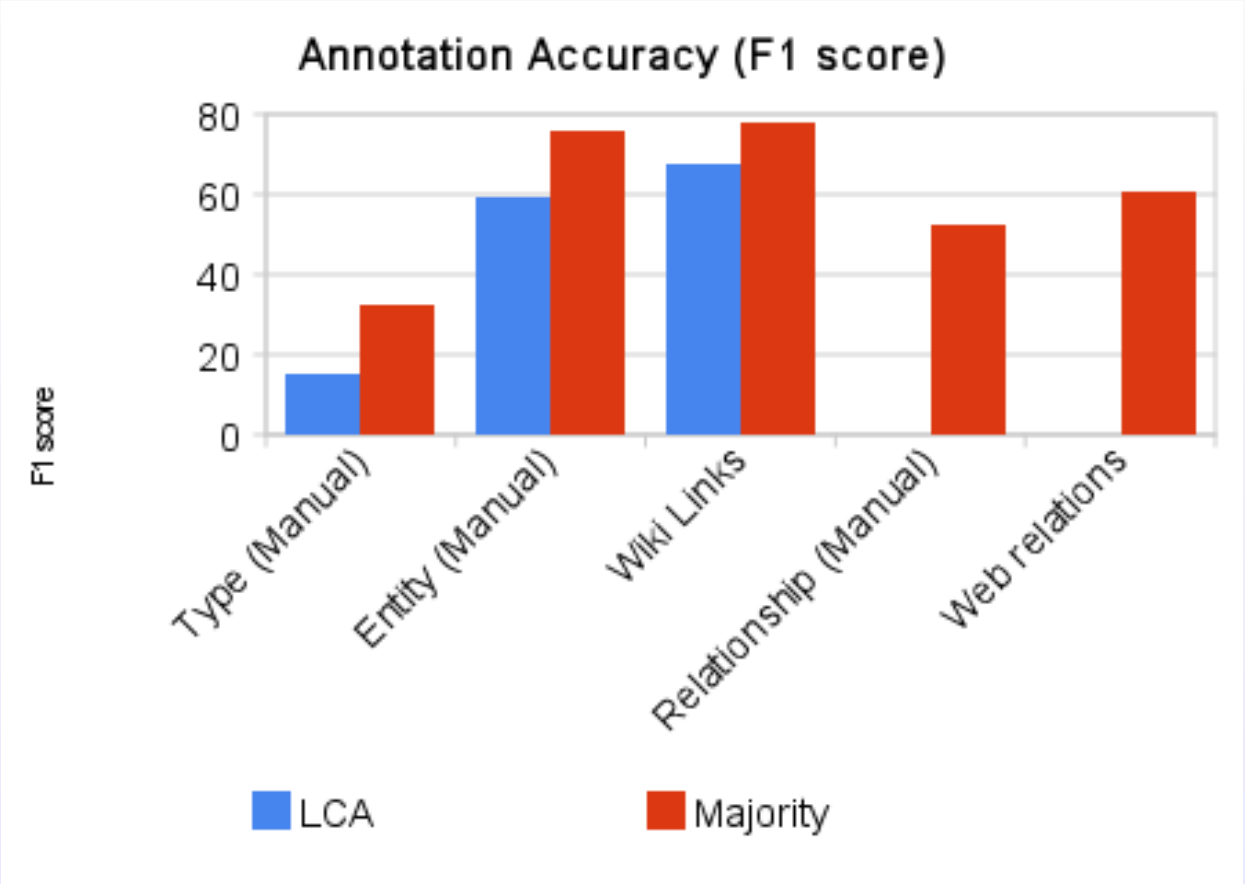


Dataset and Evaluation

Dataset	#Tables	Average #rows	Total annotations Entity	Type	Rel
Manually annotated					
Wiki_Manual	36	37	1647	72	10
Web_Manual	371	35	9239	674	44
Web_Relations	30	51	-	-	36
Wiki_Link	6085	20	131807	-	-

F1 score of for entity, type and relationship annotation

A glance at performance of baseline algorithm



Why baseline algorithms fail?

- Cell text is ambiguous
- Ontology is incomplete
  - Missing types and relationships, and their instances
- Links are missing in ontology
  - Not sure if link is absent or actually the entity is not an instance of that type
- Evidence from outside table text is ignored
  - say header text, when they are useful

Combined Annotator Model

Features considering table text

Cell Text - Entity	
$\mathbf{f}_1(r, c, E)$	$\max_{\ell \in L(E)} \mathit{similarity}(D_{r,c}, \ell)$

Features considering table text

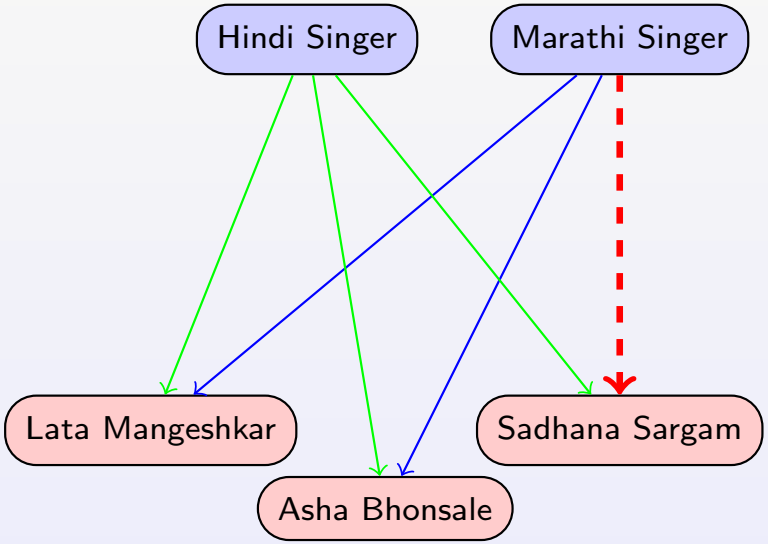
Cell Text - Entity	
$\mathbf{f}_1(r, c, E)$	$\max_{\ell \in L(E)} similarity(D_{r,c}, \ell)$

Column Header Text - Type	
$\mathbf{f}_2(c, t_c)$	$\max_{\ell \in L(t_c)} similarity(H_c, \ell)$

Features considering ontology structure

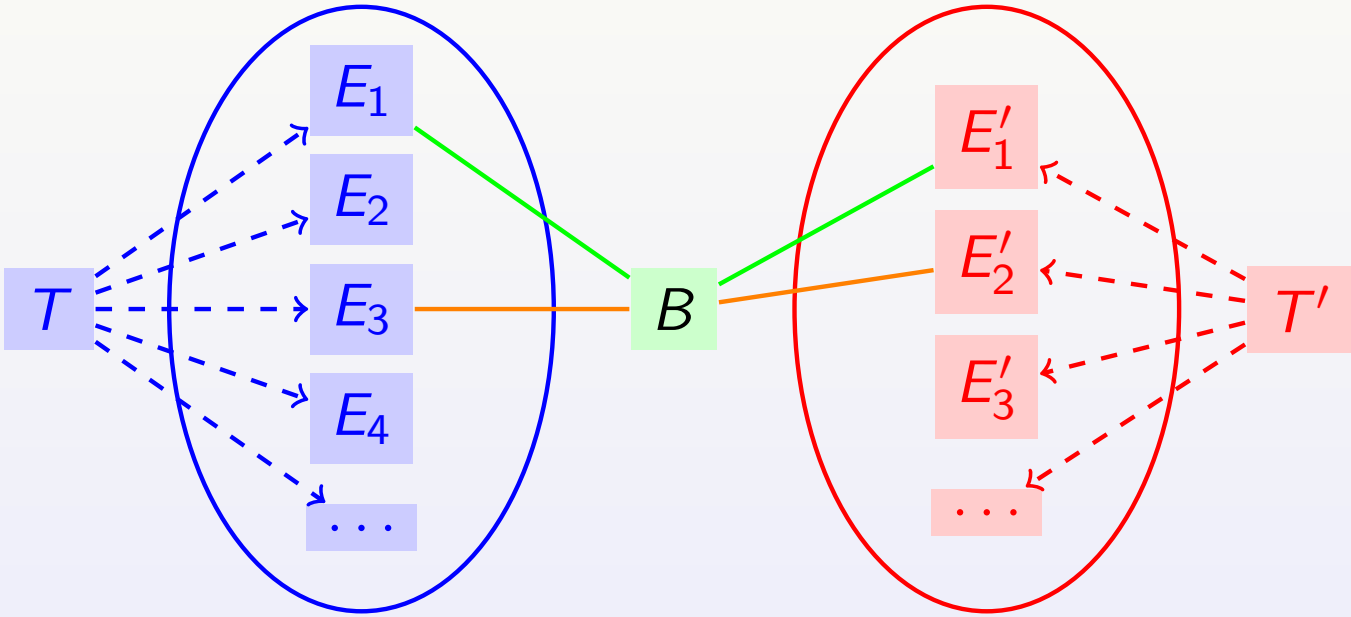
$\mathcal{E}(T)$ = the set of all entities which are direct or transitive instanced of type T

Specificity of T	$\frac{1}{dist(T)}$, $\frac{1}{\sqrt{dist(T)}}$
	$\frac{\lvert\mathcal{E}\rvert}{\lvert\mathcal{E}(T)\rvert}$
Missing links	$\frac{\lvert\mathcal{E}(T')\cap\mathcal{E}(T)\rvert}{\lvert\mathcal{E}(T')\rvert}$



Type - Entity Compatibility	
$\mathbf{f}_3(t_c, e_{r,c})$	$\min_{E\in T'} \frac{\lvert\mathcal{E}(T')\cap\mathcal{E}(T)\rvert}{\lvert\mathcal{E}(T')\rvert} \frac{1}{\sqrt{dist(T)}}$

Features for relationships

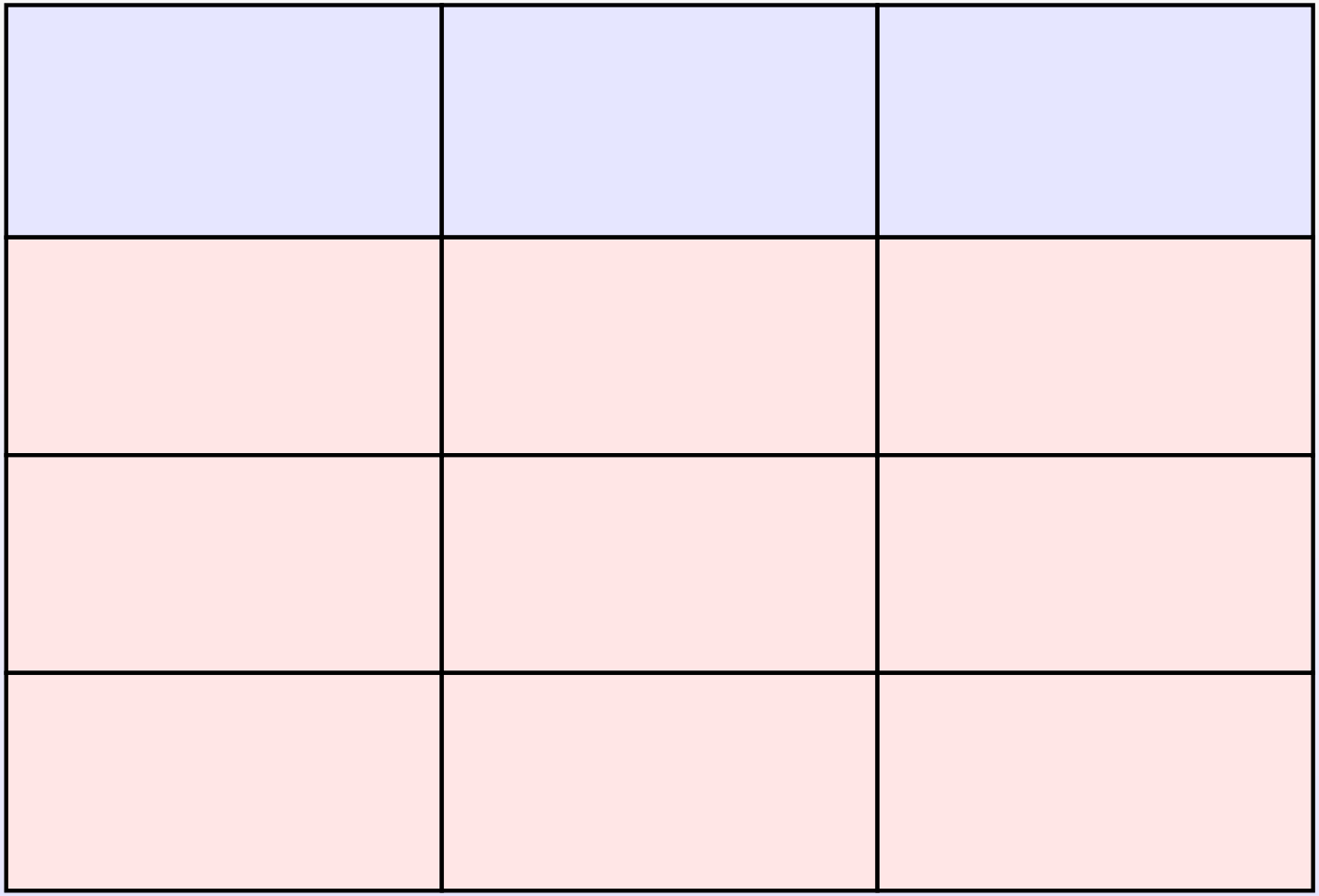


Type - Type - Relation	
$\mathbf{f}_4(b_{c,c'}, t_c, t_{c'})$	number of entities under t (or t') that appear in relationship b with an entity in t' (or t)

Features for relationships

Entity - Entity - Relation	
$\mathbf{f}_5(b_{c,c'}, e_{r,c}, e_{r,c'})$	1 : $b_{c,c'}(e_{r,c}, e_{r,c'})$ exists 0.1 : $b_{c,c'}$ is one-to-many and $b_{c,c'}(e_{r,c}, e_{r,c'})$ doesn't exist -1 : $b_{c,c'}$ is one-to-one and $b_{c,c'}(e_{r,c}, E')$ for $E' \neq e_{r,c'}$

Model variables



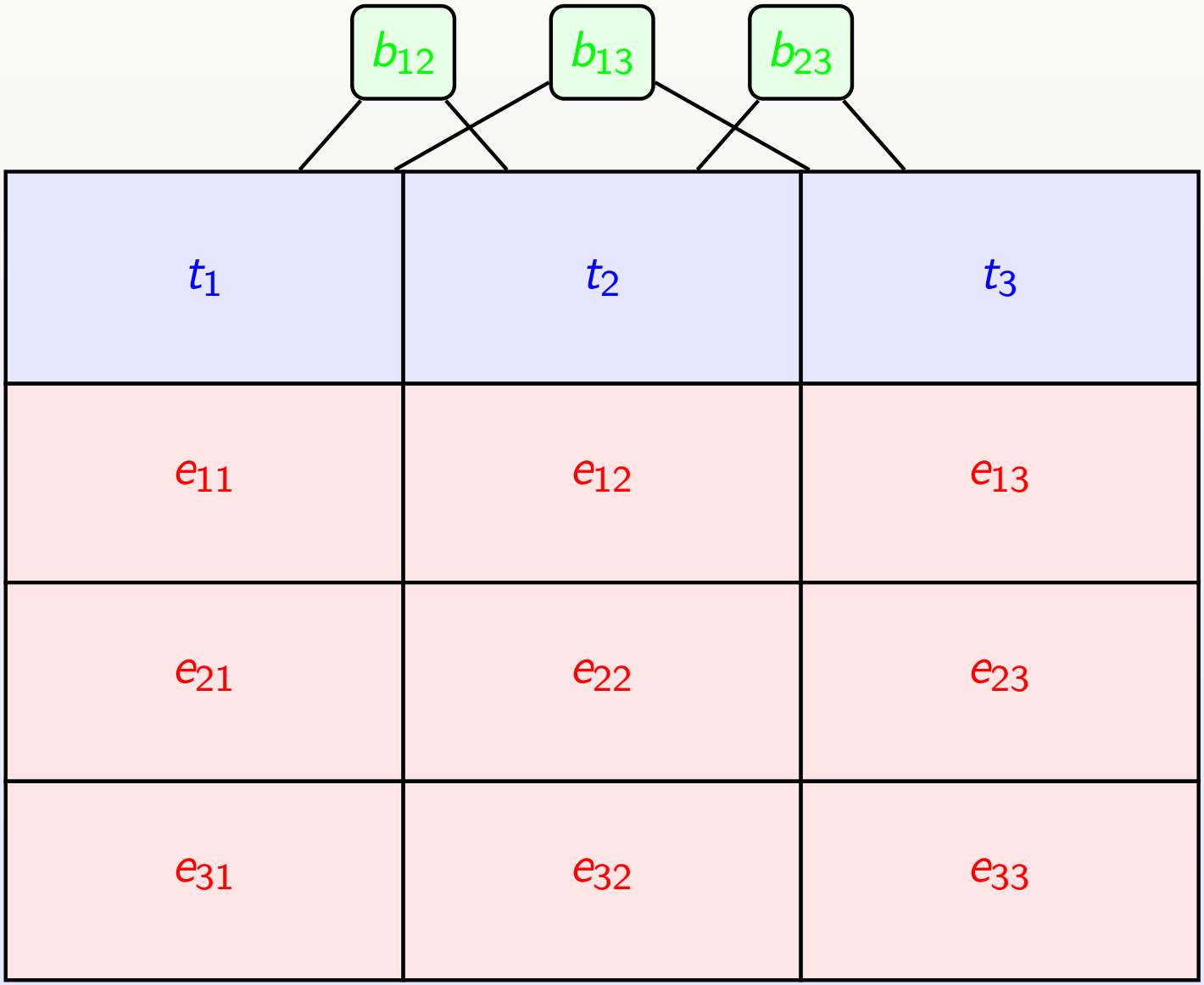
Model variables

t_1	t_2	t_3

Model variables

t_1	t_2	t_3
e_{11}	e_{12}	e_{13}
e_{21}	e_{22}	e_{23}
e_{31}	e_{32}	e_{33}

Model variables

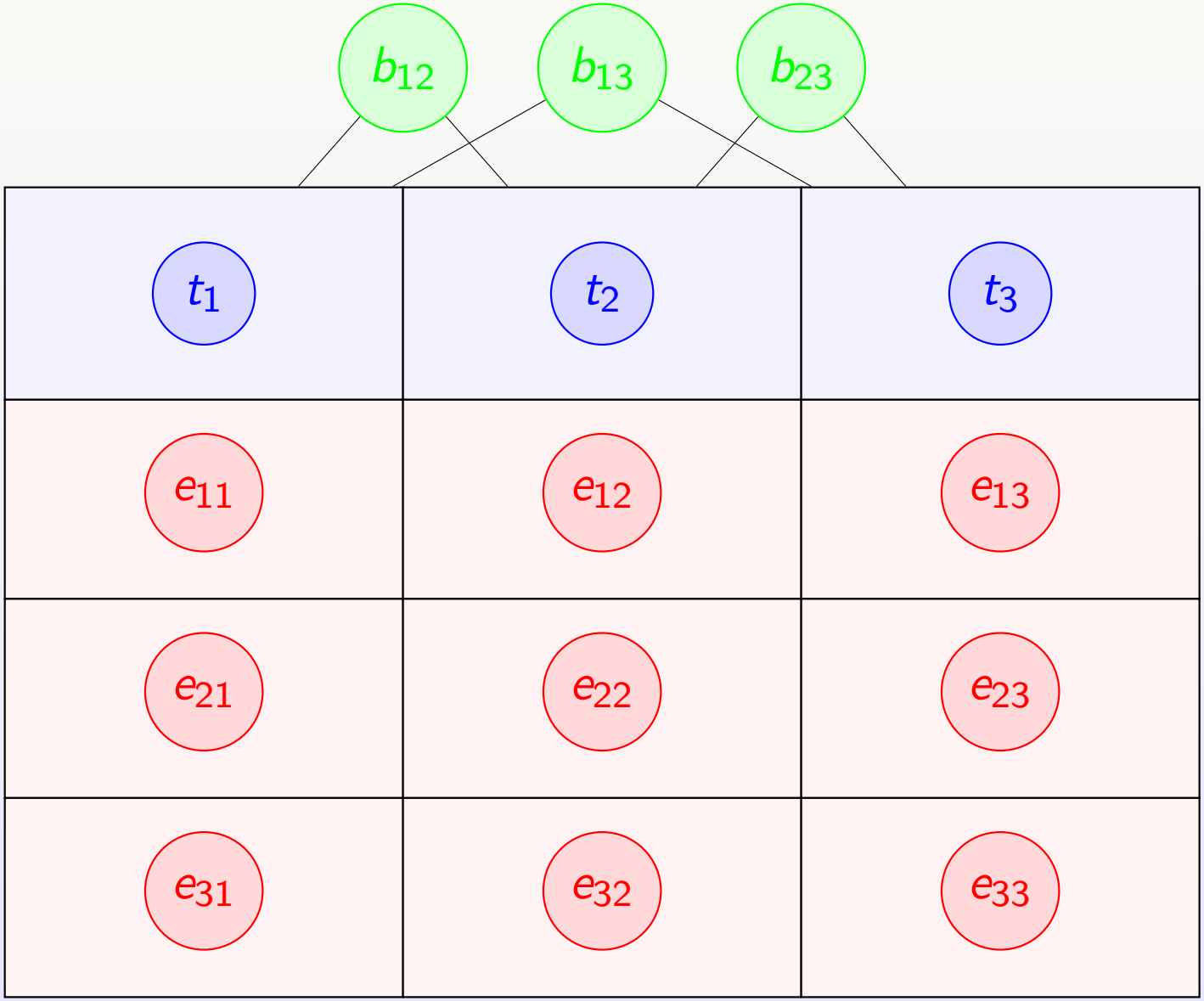


Potentials

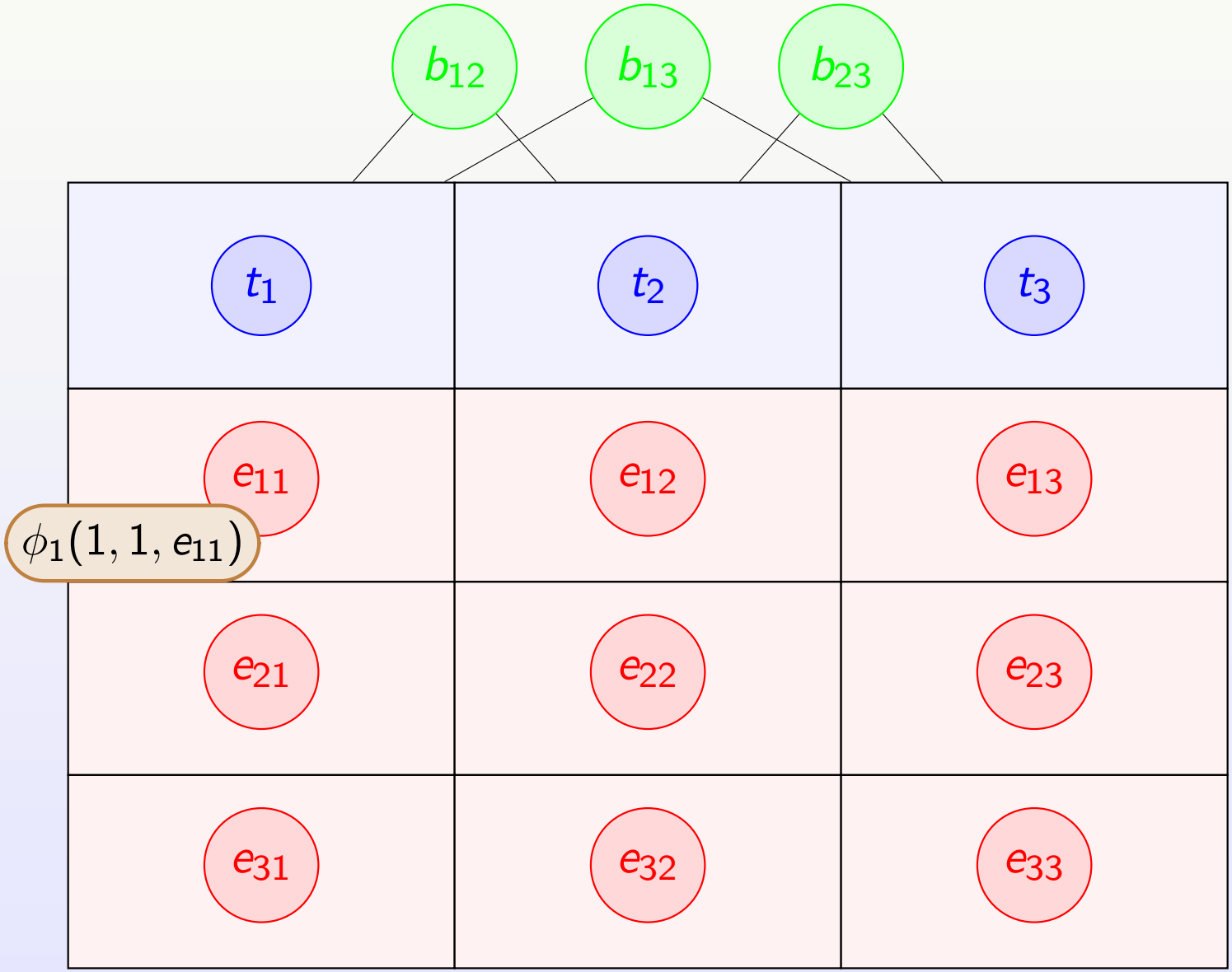
Potential	Notation	Definition
E Node	$\phi_1(r, c, e_{r,c})$	$\exp\left(\mathbf{w}_1^\top \mathbf{f}_1(r, c, e_{r,c})\right)$
T Node	$\phi_2(c, t_c)$	$\exp\left(\mathbf{w}_2^\top \mathbf{f}_2(c, t_c)\right)$
E-T Edge	$\phi_3(t_c, e_{r,c})$	$\exp\left(\mathbf{w}_3^\top \mathbf{f}_3(t_c, e_{r,c})\right)$
T-T'-R Clique	$\phi_4(b, t, t')$	$\exp\left(\mathbf{w}_4^\top \mathbf{f}_4(b, t, t')\right)$
E-E'-R Clique	$\phi_5(b, e, e')$	$\exp\left(\mathbf{w}_5^\top \mathbf{f}_5(b, e, e')\right)$

Table: Potentials assigned to the model

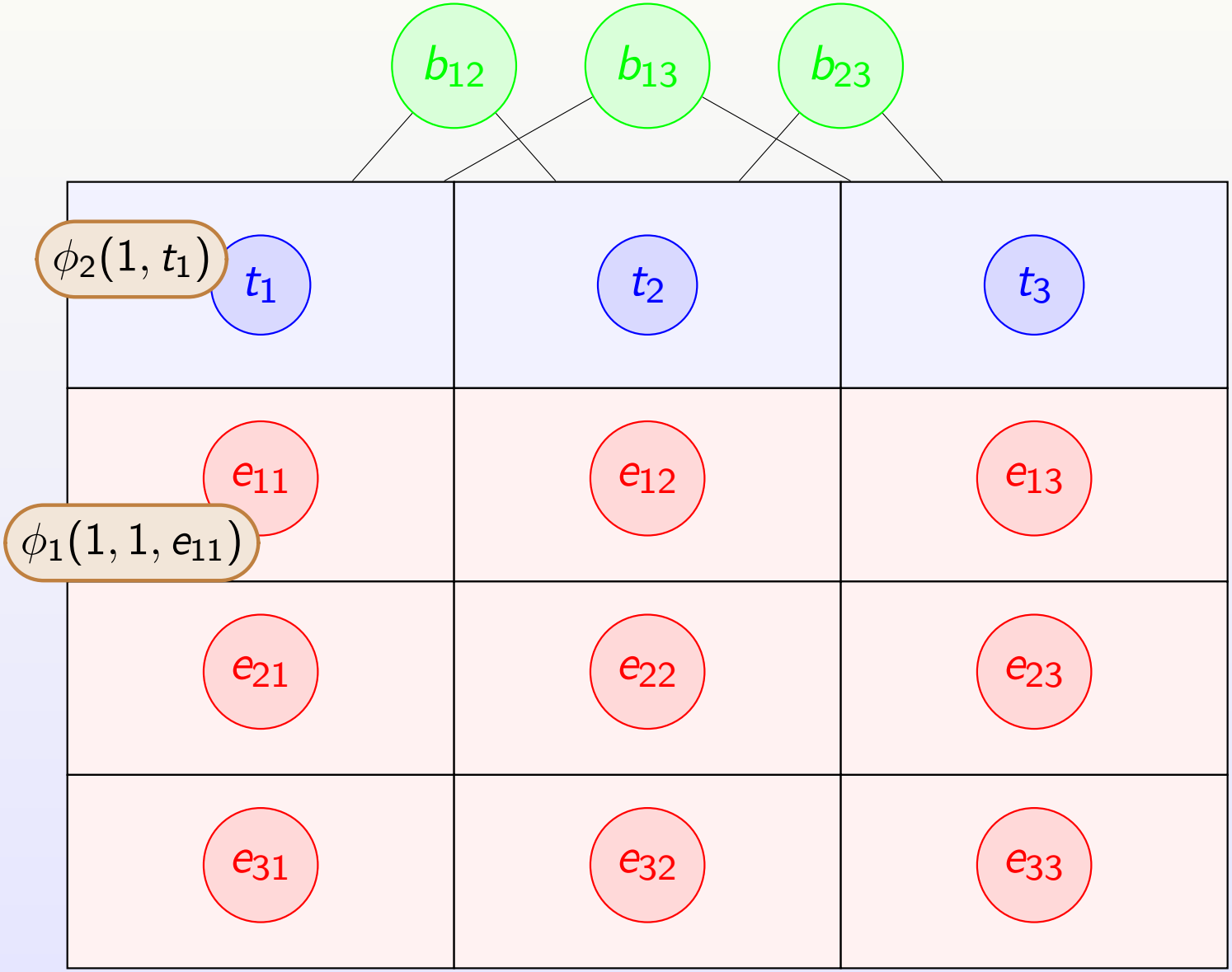
Potentials



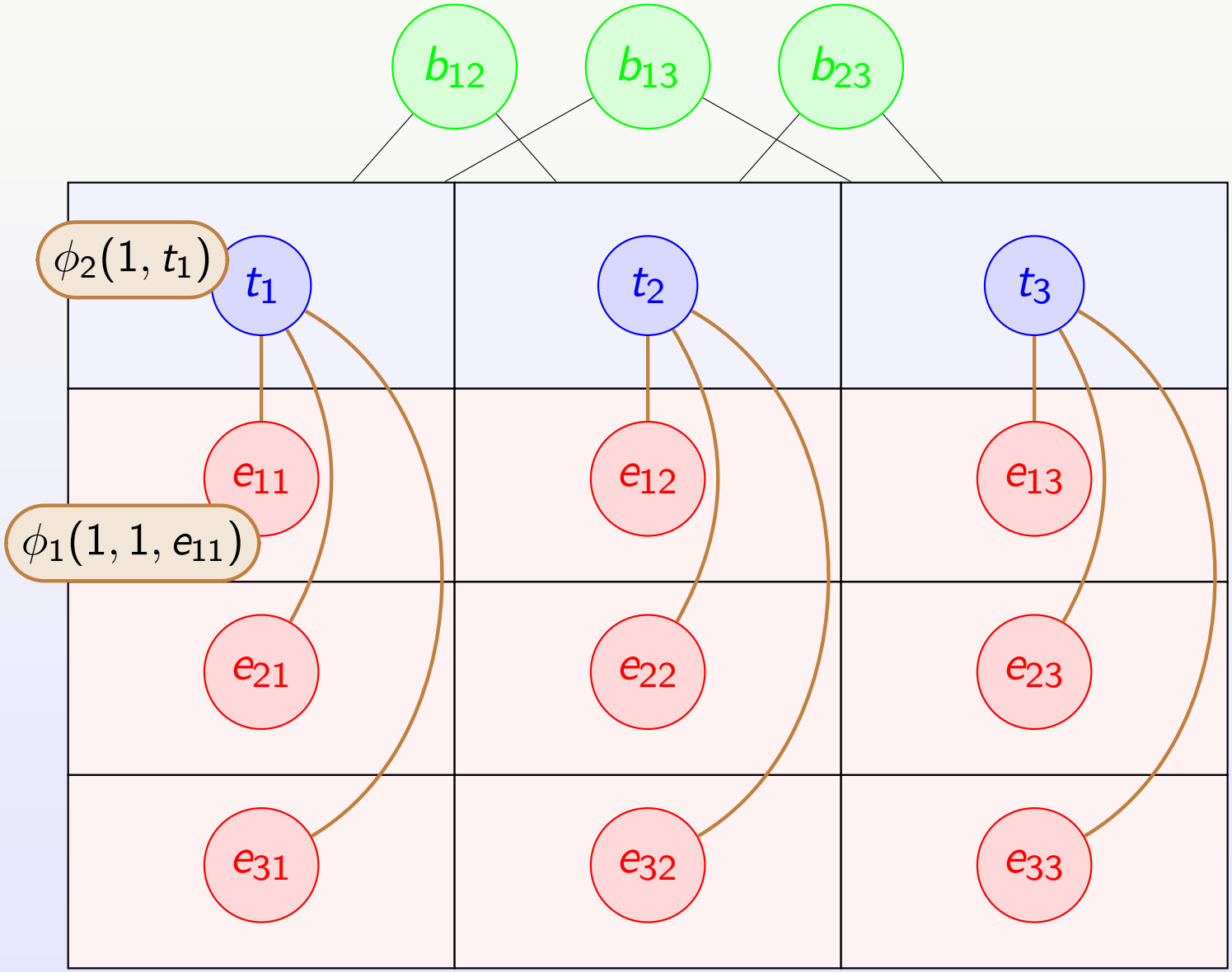
Potentials



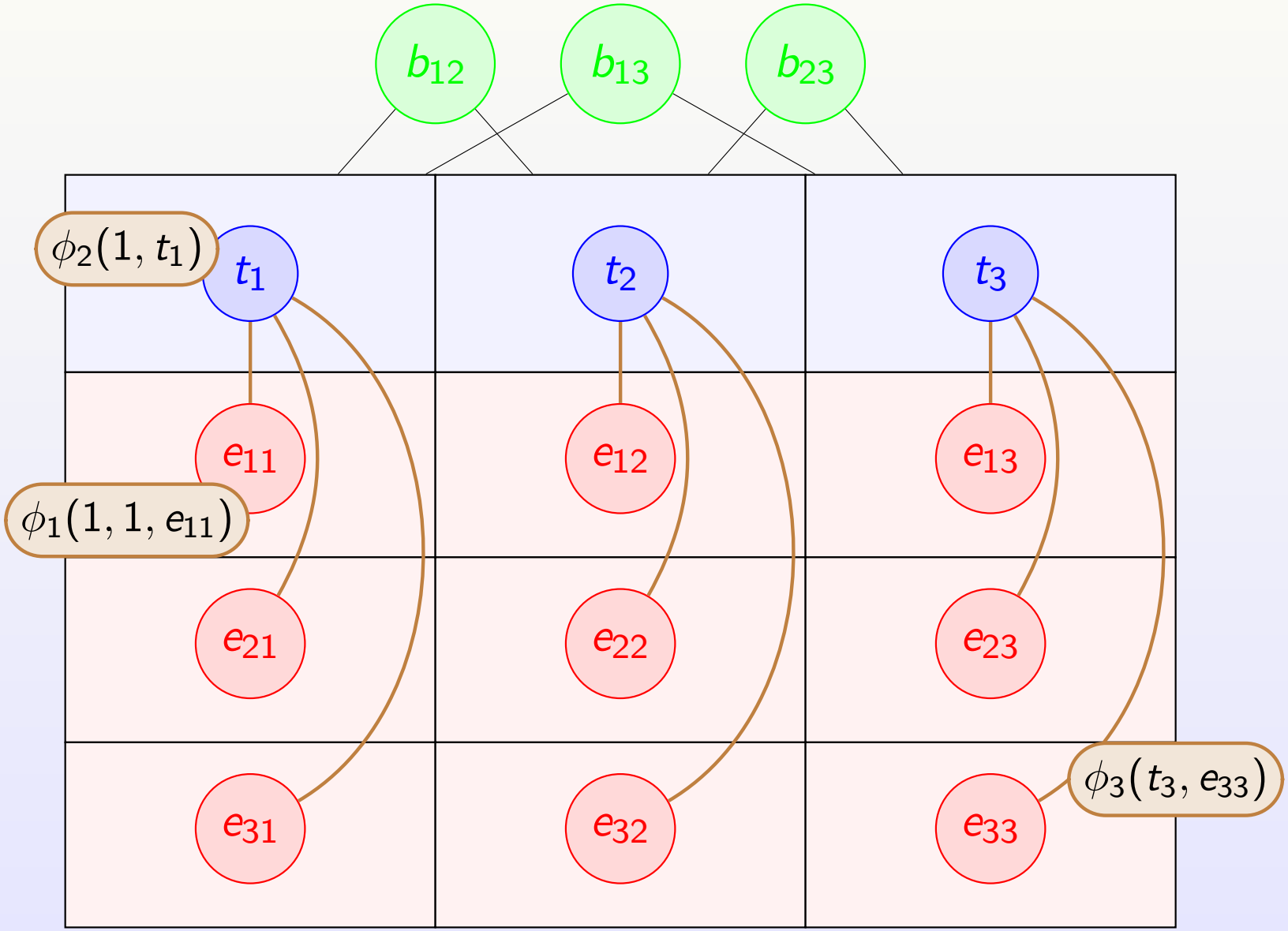
Potentials



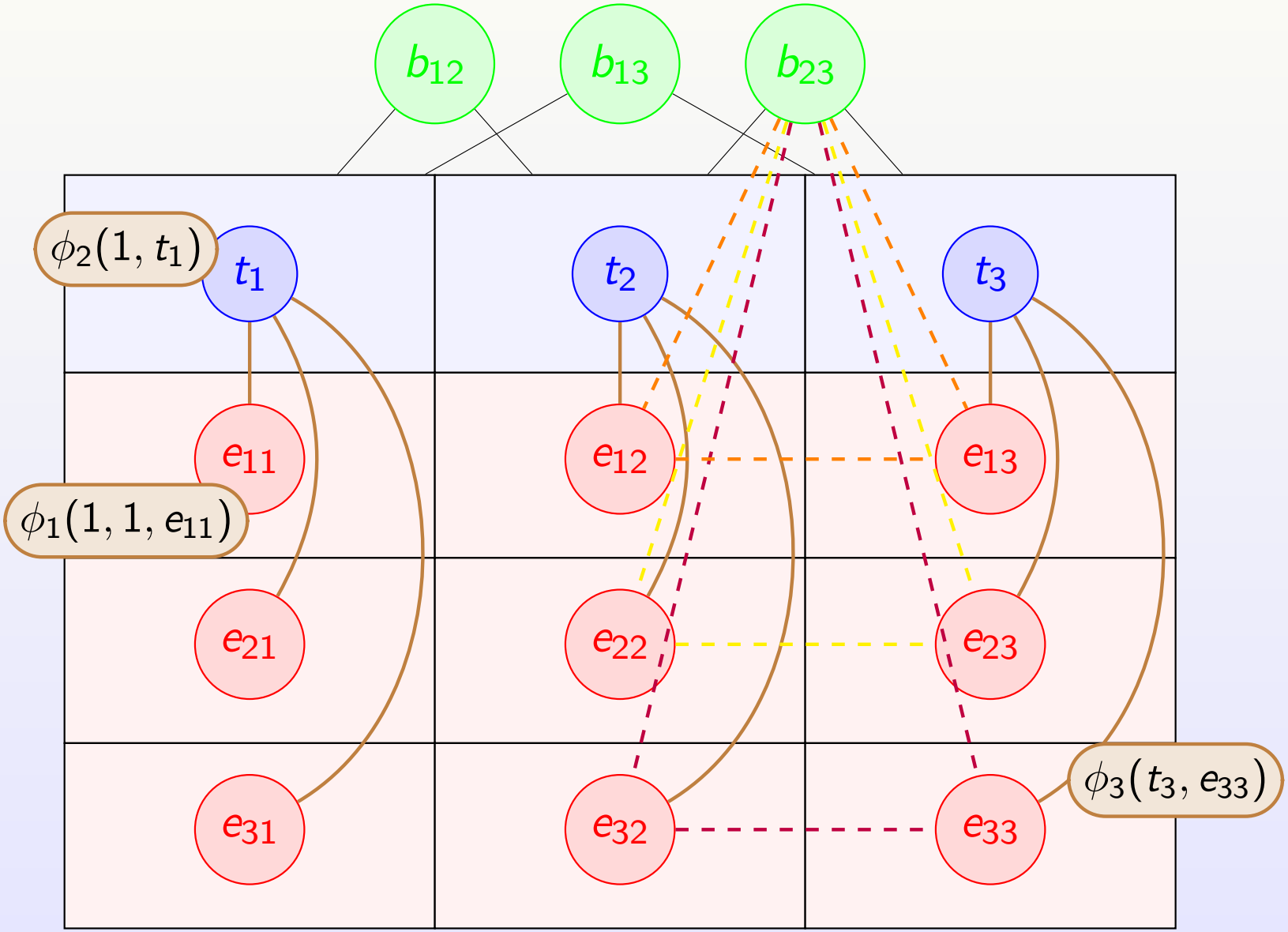
Potentials



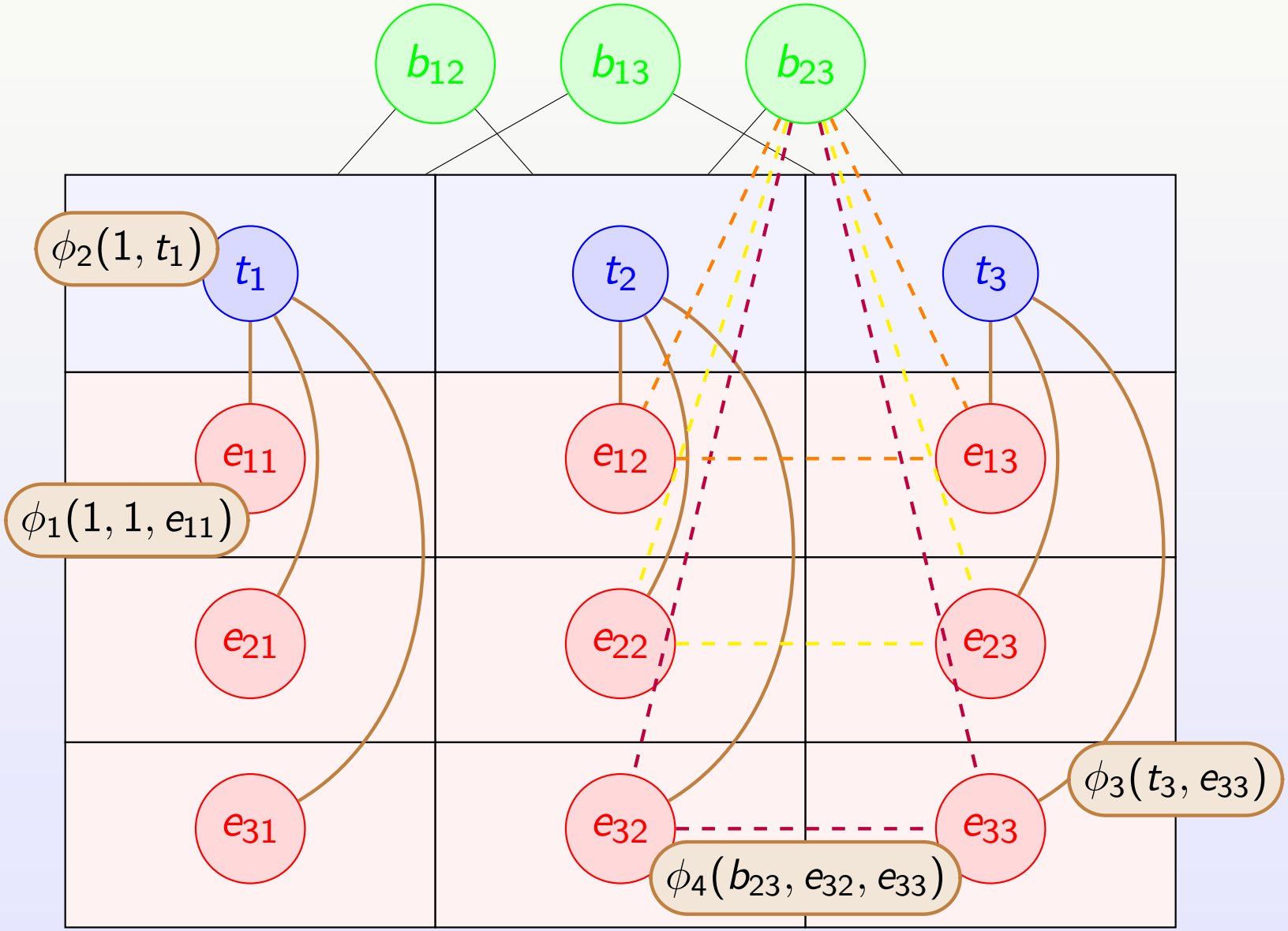
Potentials



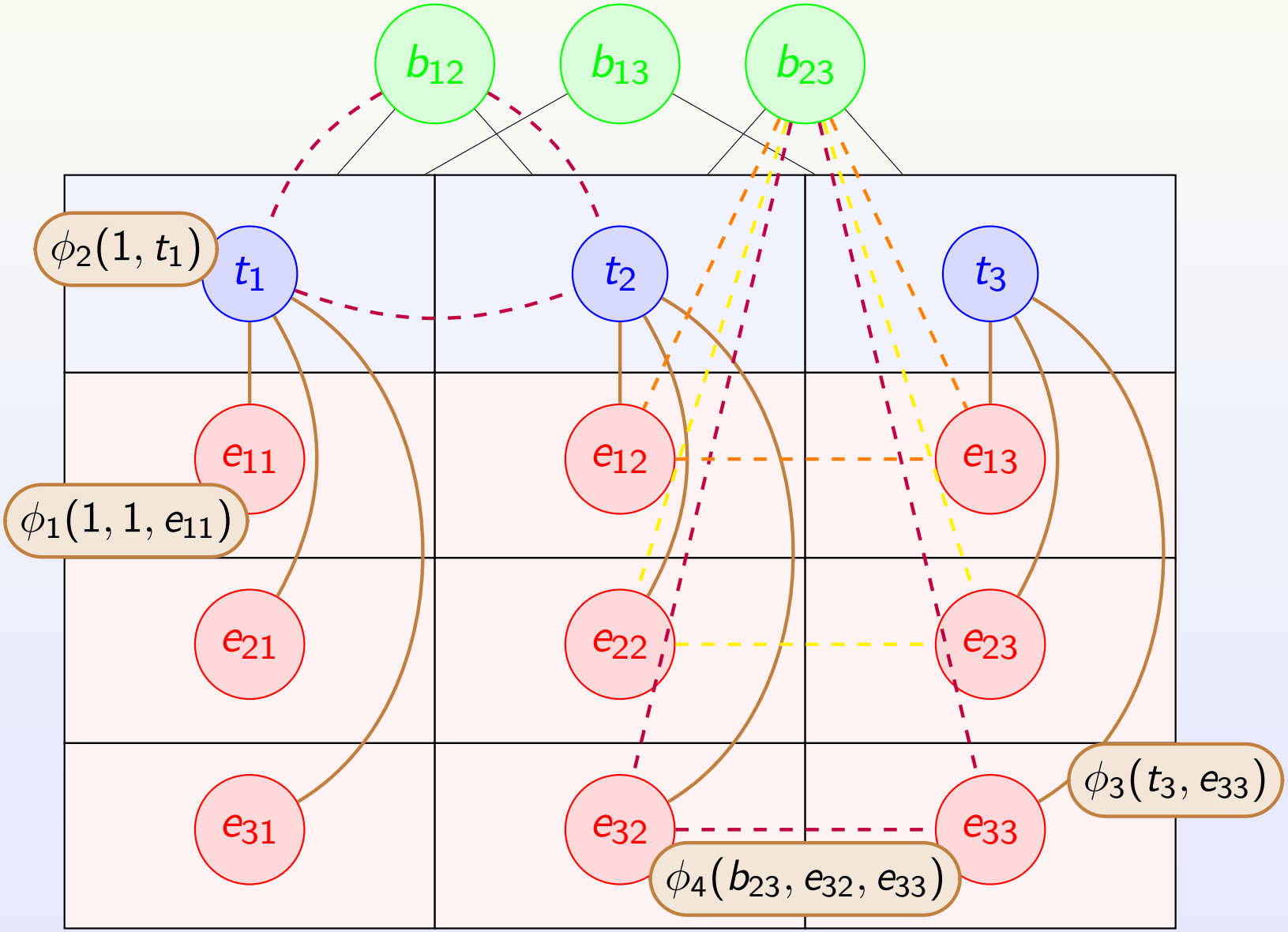
Potentials



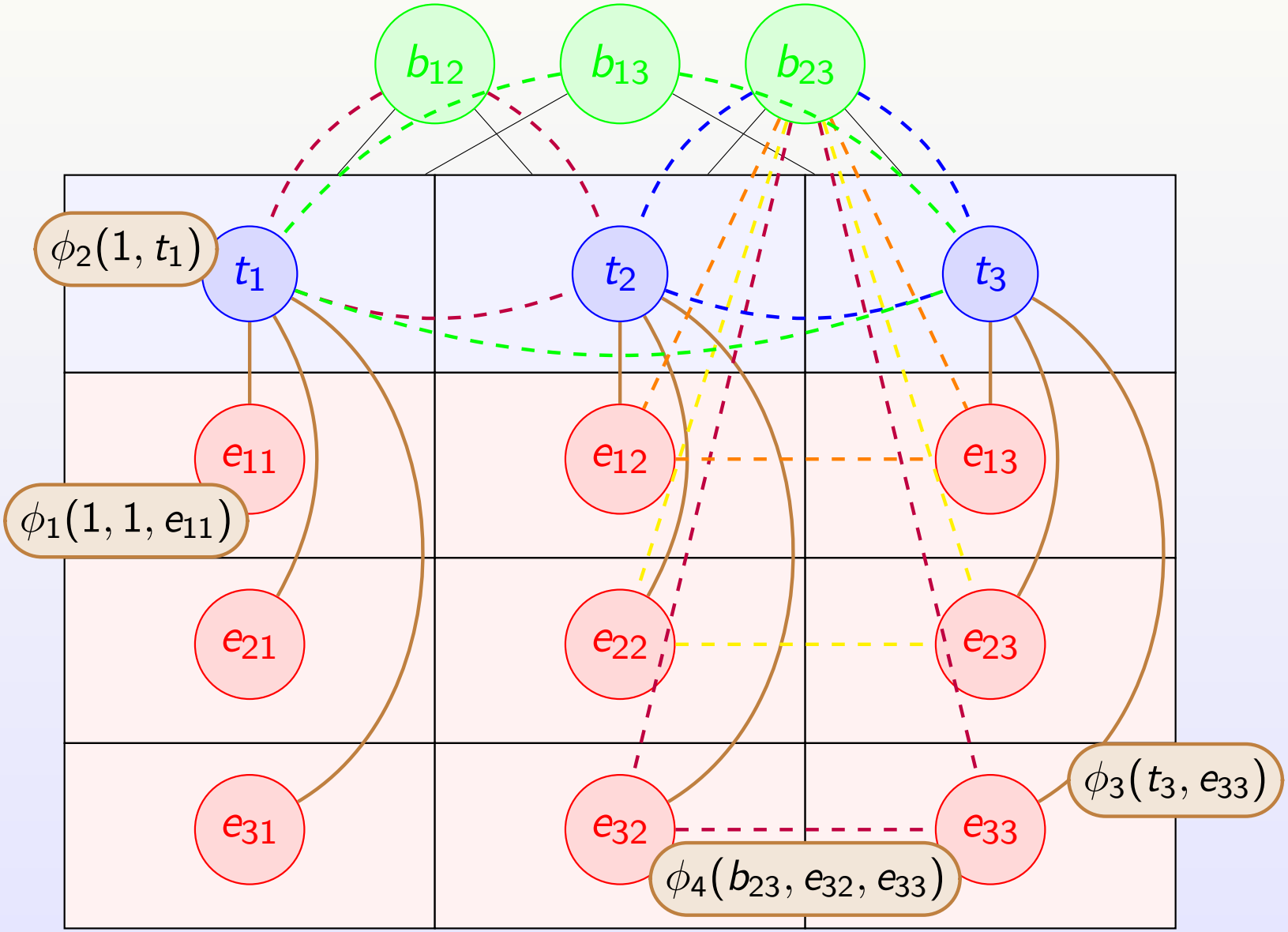
Potentials



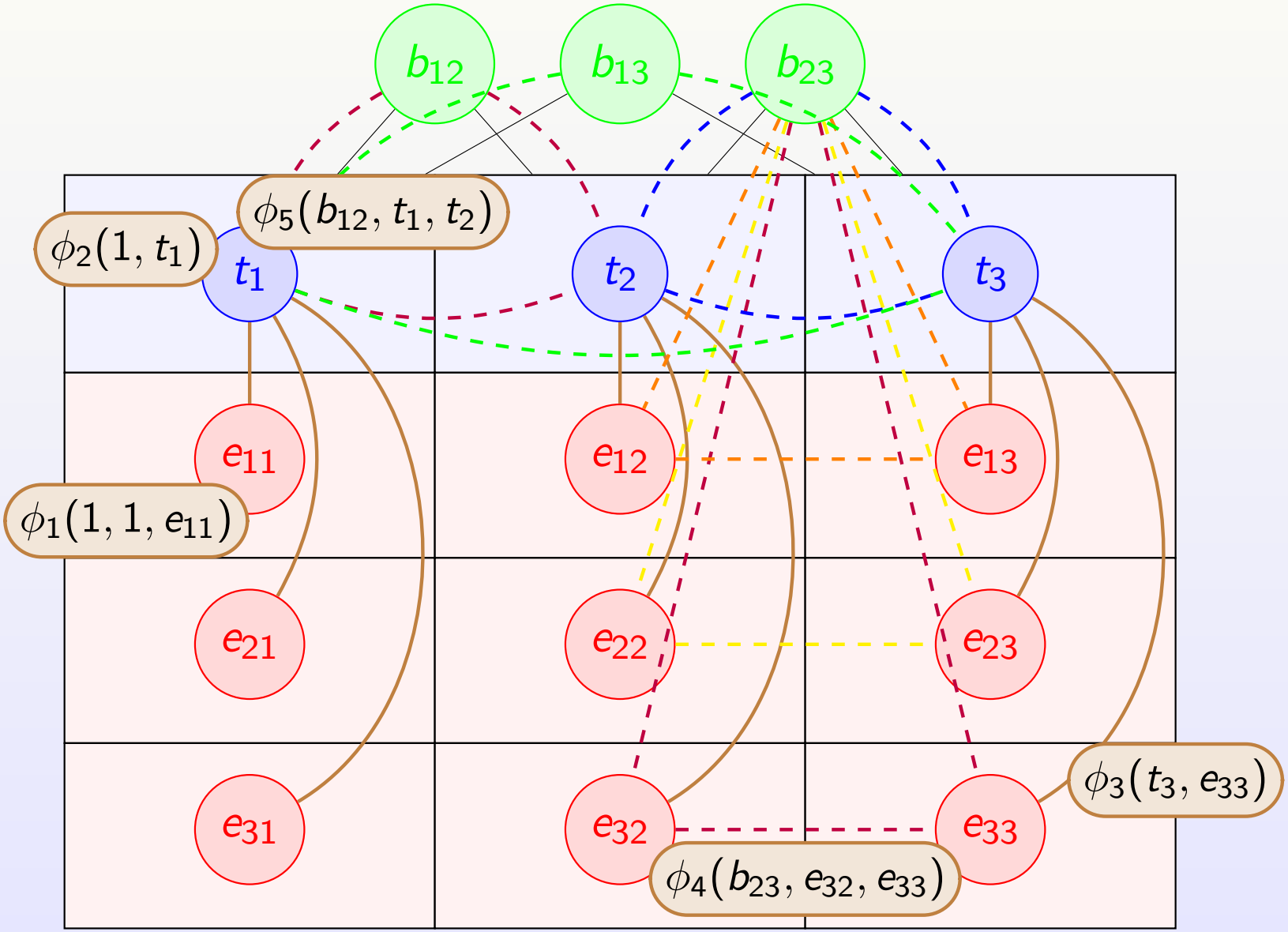
Potentials



Potentials



Potentials



Optimization objective

No relationships

$$\max_{\mathbf{e},\mathbf{t}} \underbrace{\prod_{c} \phi_2(c, t_c)}_{\text{columns}} \underbrace{\prod_{r,c} \phi_1(r, c, e_{r,c})\, \phi_3(t_c, e_{r,c})}_{\text{cells}} \tag{1}$$

Optimization objective

No relationships

$$\max_{\mathbf{e},\mathbf{t}} \underbrace{\prod_{c} \phi_2(c, t_c)}_{\text{columns}} \underbrace{\prod_{r,c} \phi_1(r, c, e_{r,c})\, \phi_3(t_c, e_{r,c})}_{\text{cells}} \tag{1}$$

With relationships

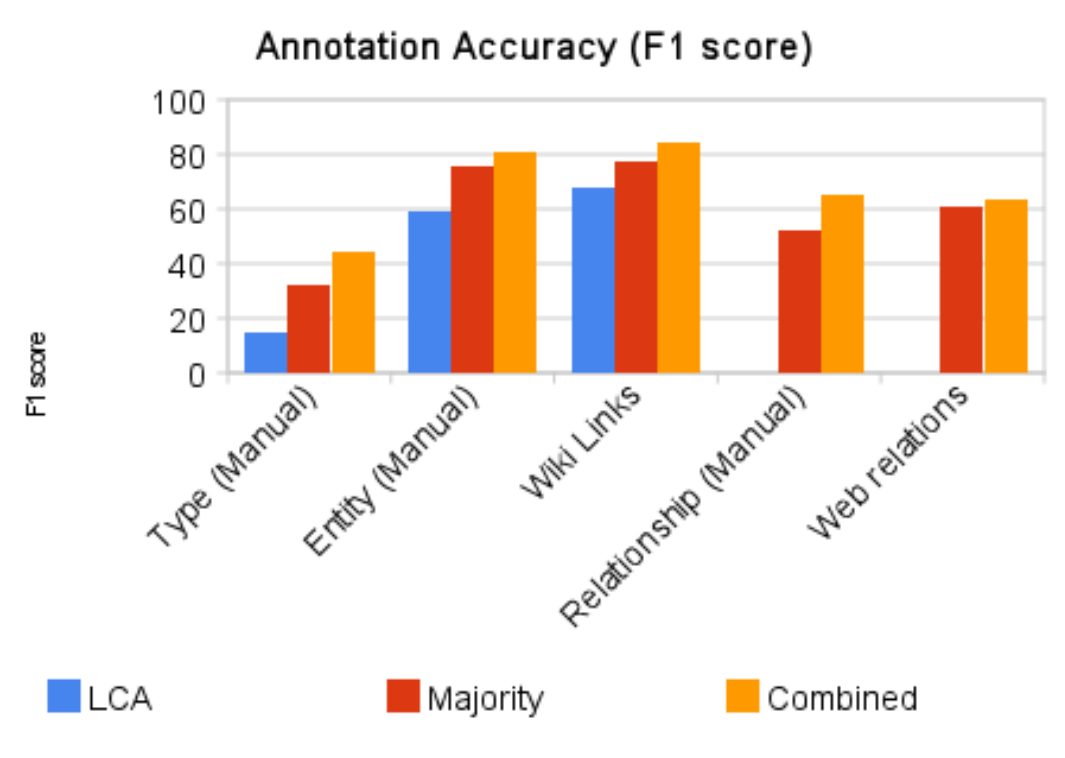
$$\max_{\mathbf{e},\mathbf{t},\mathbf{b}} \underbrace{\prod_{c,c'} \phi_4(b_{c,c'}, t_c, t_{c'}) \prod_{r} \phi_5(b_{c,c'}, e_{rc}, e_{r,c'})}_{\text{relation}} \underbrace{\prod_{c} \phi_2(c, t_c)}_{\text{columns}} \underbrace{\prod_{r,c} \phi_1(r, c, e_{r,c})\, \phi_3(t_c, e_{r,c})}_{\text{cells}} \tag{2}$$

Approximate Inference using Message Passing algorithm

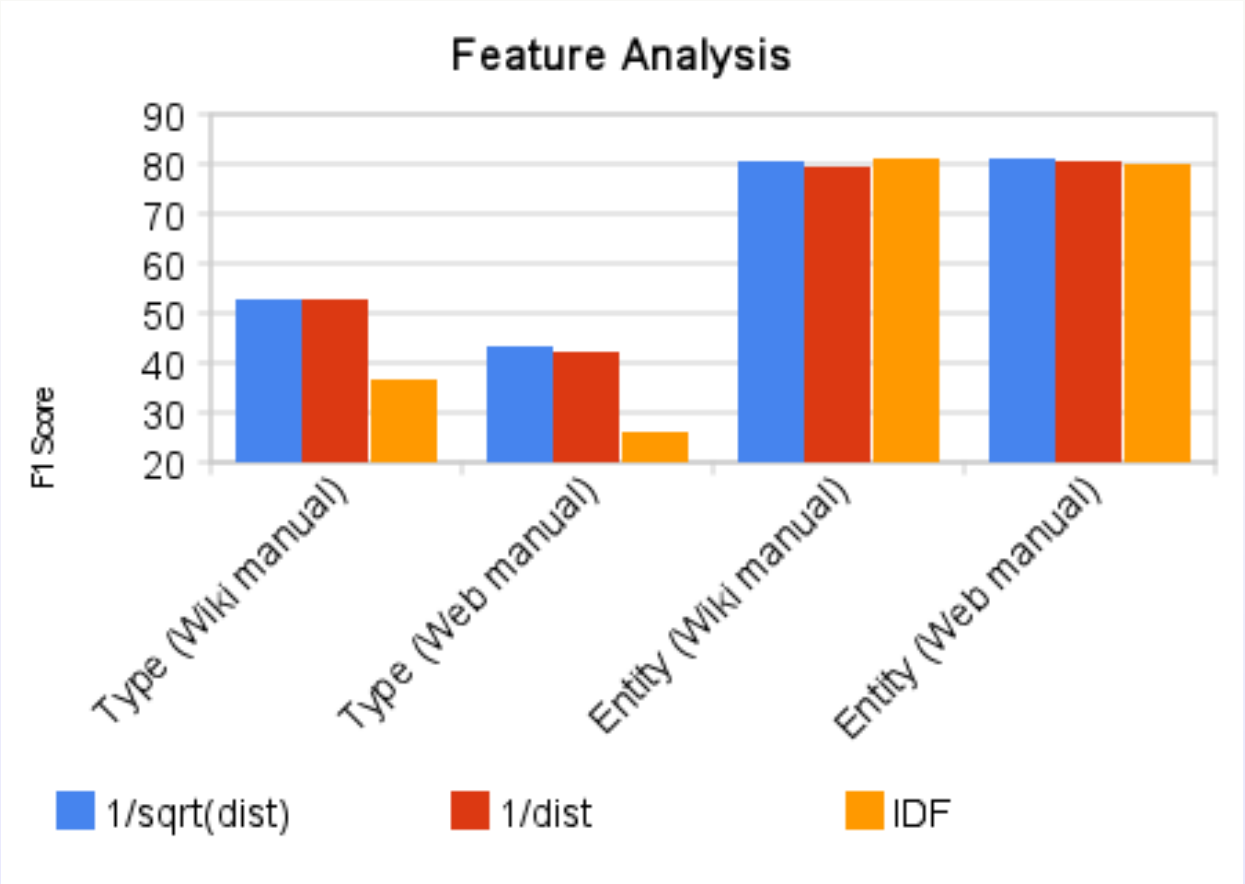
Comparison with baseline algorithm

Improves Entity annotation accuracy

Improves precision in Relationship and Type annotations

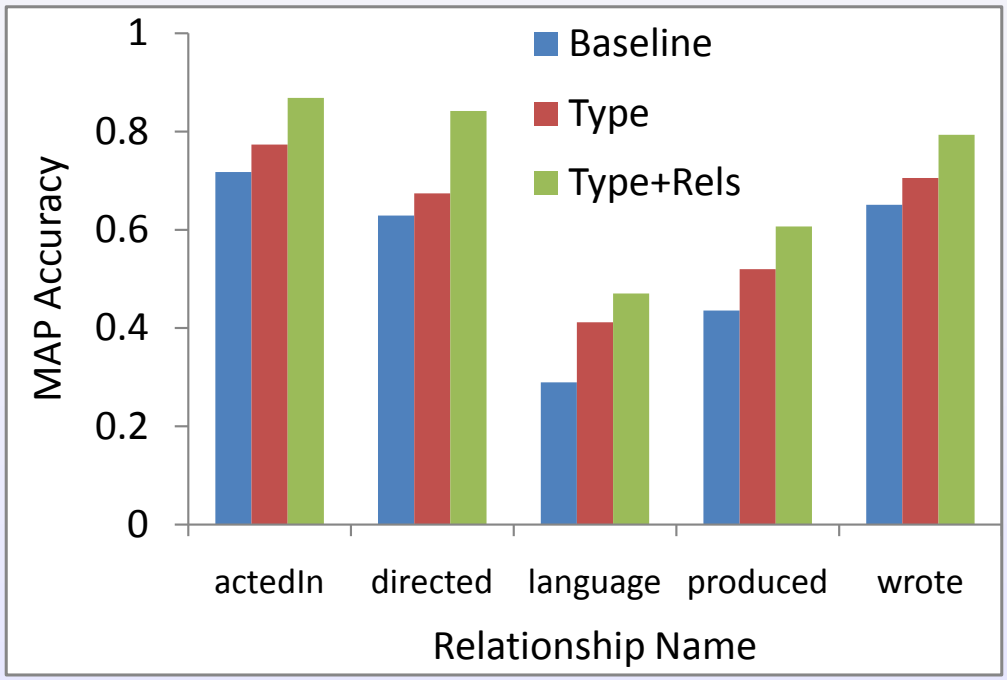


Analysis of Type-Entity compatibility feature

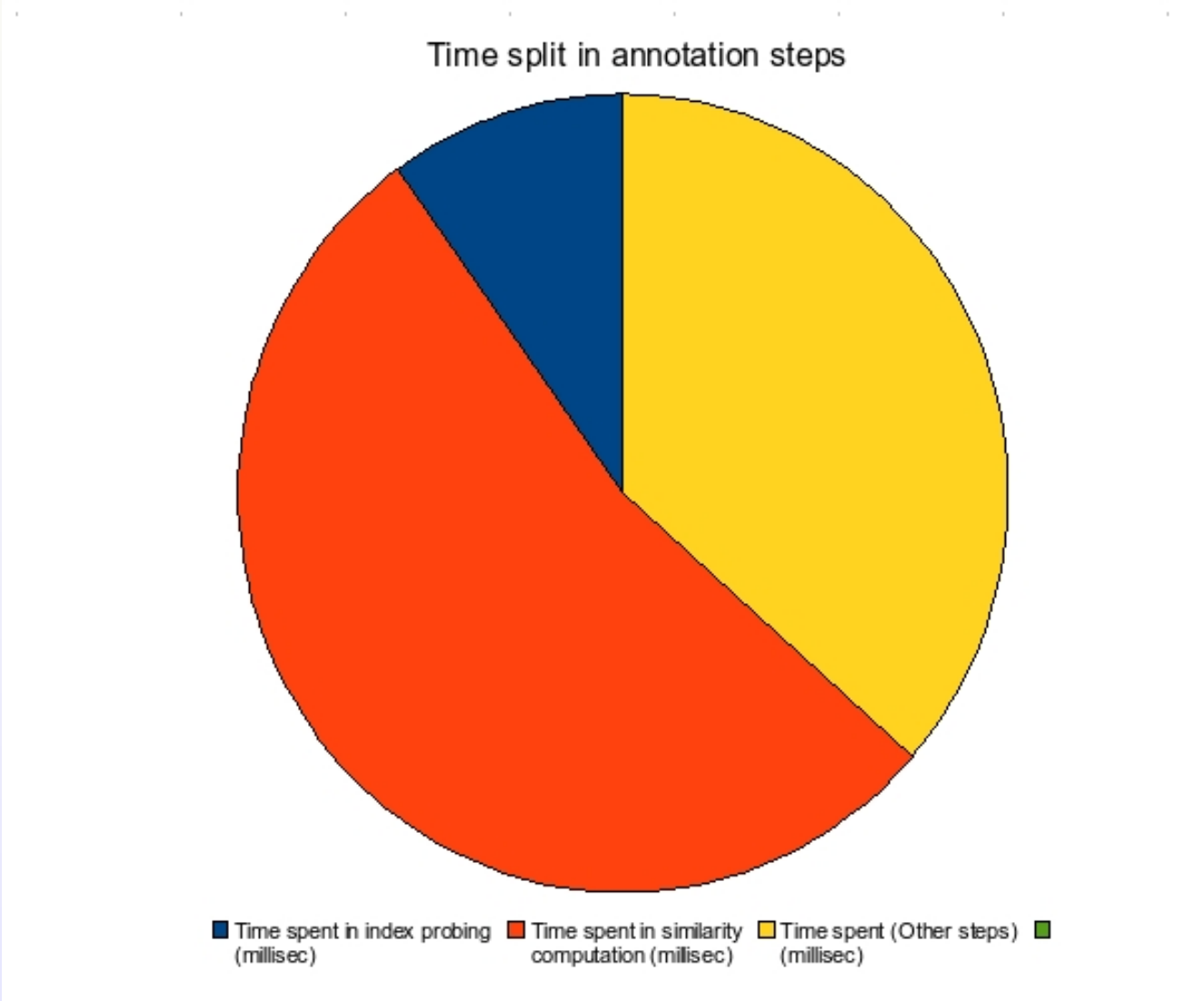


Entity Search

Relations from YAGO
Ground truth from dbPedia
Baseline (without annotation) With type annotation With type and relationship annotation



Annotator performance



Rate of annotation

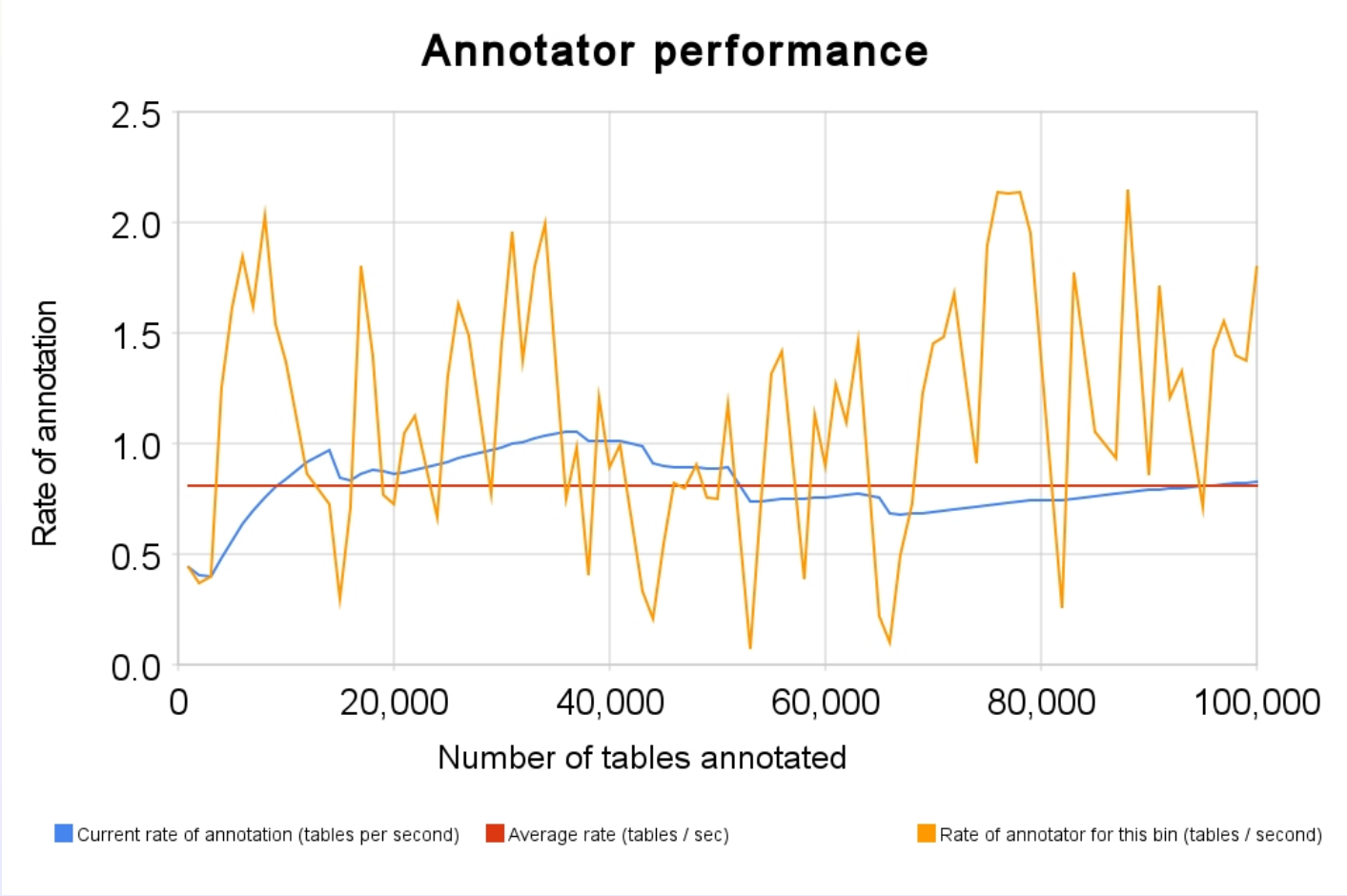


Figure: Rate of annotation (Tables per second)

Conclusion

- Web tables = wealth of information, useful for structured queries
- Annotations help mapping them to entities, types and relationships
- Challenges due to ambiguities and ontology structure
- Combined model outperforms baseline algorithms
- To lead to efficient and precise answering to structured queries

Thank you!


Satanjeev Banerjee and Ted Pedersen.
An adapted lesk algorithm for word sense disambiguation using wordnet.
In *Computational Linguistics and Intelligent Text Processing*, 2002.


Michael J. Cafarella, Alon Halevy, Zhe Daisy Wang, Eugene Wu, and Yang Zhang.
Webtables: Exploring the power of tables on the web.
In *International Conference on Very Large Data Bases (VLDB)*, 2008.


Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti.
Collective annotation of wikipedia entities in web text.
In *Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2009.

Michael Lesk.
Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.
In *International conference on Systems documentation (SIGDOC)*, 1986.

David Milne and Ian H. Witten.
Learning to link with wikipedia.
In *Conference on Information and Knowledge Management (CIKM)*, 2008.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum.
Yago - a core of semantic knowledge.
In *World Wide Web Conference (WWW)*, 2007.

- Parameter Setting
- Scope for improvement
- Stage I and II work split
- Analysis of annotations made
- Pseudo-code for algorithms

Demo@ `http://10.129.125.1:8090/demo/scripts/main.jsp`

Inference in special case

for each column c **do**
  for each type $T \in \mathcal{T}_c$ **do**
    $A_T \leftarrow \phi_2(c, T)$
    for each cell r, c in column c **do**
      choose $e^*_{r,c} = \arg\max_{E \in \mathcal{E}_{r,c}} \phi_1(r, c, E)\phi_3(T, E)$
      $A_T \leftarrow A_T \cdot \phi_1(r, c, e^*_{r,c}) \cdot \phi_3(T, e^*_{r,c})$
    end for
  end for
  finalize $t^*_c = \arg\max_T A_T$
  recall and finalize cell assignments $e^*_{r,c}$
end for
return $\mathbf{t}^*, \mathbf{e}^*$

Messages

$M(e_{rc} \rightarrow \phi_3(t_c, e_{rc})) = \phi_1(r, c, e_{rc}) \prod_{c'} M(\phi_5(b_{cc'}, e_{rc}, e_{rc'}) \rightarrow e_{rc})$

$M(\phi_3(t_c, e_{rc}) \rightarrow t_c) = \max_{e_{rc}} \phi_3(t_c, e_{rc}) M(e_{rc} \rightarrow \phi_3)$

$M(t_c \rightarrow \phi_3(t_c, e_{rc})) = \phi_2(c, t_c) \prod_{c'} M(\phi_4(b_{cc'}, t_c, t_{c'}) \rightarrow e_{rc}) \prod_{r' \neq r} M(\phi_3(t_c, e_{r'c}) \rightarrow t_c)$

$M(\phi_3(t_c, e_{rc}) \rightarrow e_{rc}) = \max_{t_c} \phi_3(t_c, e_{rc}) M(t_c \rightarrow \phi_3).\ M(e_{rc} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'})) =$

$\phi_1(r, c, e_{rc}) \prod_{c'' \neq c'} M(\phi_5(b_{cc''}, e_{rc}, t_{c''}) \rightarrow e_{rc}) \prod_{r} M(\phi_3(t_c, e_{r'c}) \rightarrow e_{rc})$

$M(e_{rc'} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'})) =$ similar to above.

$M(\phi_5(b_{cc'}, e_{rc}, e_{rc'}) \rightarrow b_{cc'}) = M(t_c \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'})) M(e_{rc'} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'}))$

$M(b_{cc'} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'})) = \prod_{r \neq r'} M(\phi_5(b_{cc'}, e_{r'c}, e_{r'c'}) \rightarrow b_{cc'}) M(\phi_4(b_{cc'}, t_c, t_{c'}) \rightarrow b_{cc'})$

Messages - Continued

$M(\phi_5(b_{cc'}, e_{rc}, e_{rc'}) \rightarrow e_{rc}) = \max_{e_{rc'}, b_{cc'}} \phi_5(b_{cc'}, e_{rc}, e_{rc'}) M(e_{rc'} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'})) M(b_{cc'} \rightarrow \phi_5(b_{cc'}, e_{rc}, e_{rc'}))$ $M(\phi_5(b_{cc'}, e_{rc}, e_{rc'} \rightarrow e_{rc'}) =$ similar to above.

$M(t_c \rightarrow \phi_4(b_{cc'}, t_c, t_{c'})) = \phi_2(c, t_c) \prod_{c'' \neq c'} M(\phi_4(b_{cc''}, t_c, t_{c''}) \rightarrow e_{rc}) \prod_r M(\phi_3(t_c, e_{r'c}) \rightarrow t_c)$

$M(t_{c'} \rightarrow \phi_4(b_{cc'}, t_c, t_{c'})) =$ similar to above.

$M(\phi_4(b_{cc'}, t_c, t_{c'}) \rightarrow b_{cc'}) = M(t_c \rightarrow \phi_4(b_{cc'}, t_c, t_{c'})) M(t_{c'} \rightarrow \phi_4(b_{cc'}, t_c, t_{c'}))$

$M(b_{cc'} \rightarrow \phi_4(b_{cc'}, t_c, t_{c'})) = \prod_r M(\phi_5(b_{cc'}, e_{rc}, e_{rc'}) \rightarrow b_{cc'}) M(\phi_4(b_{cc'}, t_c, t_{c'}) \rightarrow t_c) =$

$\max_{t_{c'}, b_{cc'}} \phi_4(b_{cc'}, t_c, t_{c'}) M(t_{c'} \rightarrow \phi_4(b_{cc'}, t_c, t_{c'})) M(b_{cc'} \rightarrow \phi_4(b_{cc'}, t_c, t_{c'}))$

$M(\phi_4(b_{cc'}, t_c, t_{c'} \rightarrow t_{c'}) =$ similar to above.

Parameter setting

Parameter	**Setting**
Type-Entity compatibility	$1/\sqrt{dist(T)}$
Features used	$\mathbf{f}_1, \ldots, \mathbf{f}_5$
Type and Entity accuracy	ϕ_1, ϕ_2, ϕ_3
Relationship accuracy	$\phi_1, \ldots, \phi_5$

Parameter	**Value**	**Parameter**	**Value**
max_hits	150	score_percent	0.1
USE_SQRT_LEVEL	true	consider_overlap	true
jazzy	false	useidf	true

Scope for improvement

- Better training of parameters
- Candidate label generation
  - Most time consuming task
  - Better data structures to represent YAGO
- Join queries over tables

Work completed - Stage I

- Literature Survey
- YAGO processing
- Basic feature design
- Trying out simple ideas
- Tool for collecting ground truth
- Gathering ground truth and basic evaluation

Work completed - Stage II

- Read few more papers
- Combined model creation and integration with inference module
- More YAGO processing
- Trials with sampled ground truth
- Integrating with StructLearn for parameter training
- Generating ground truth for Wiki_Link dataset
- More experiments on annotator efficiency and performance
- Annotating Web corpus of tables
- Adding context related features
- Trying out MLE training

Top 50 types in a subset of annotated corpus

	Total types annotated	5443406	-
	Annotated with NA	1966964	36%
	Total types in top K	1799671	33%
	Number of distinct types	248992	-
Rank	**Few interesting types in top 50**		
5	wordnet_village_108672738	73349	1.3%
7	wordnet_city_108524735	57439	1.05%
10	wordnet_movie_106613686	37218	0.68%
17	wordnet_administrative_district_108491826	17349	0.31%
19	wikicategory_States_of_the_United_States	15407	0.28%
21	wordnet_university_108286163	15274	0.28%
31	wordnet_company_108058098	10271	0.18%
32	wordnet_organization_108008335	10057	0.18%
33	wikicategory_American_films	10001	0.18%
34	wordnet_country_108544813	9985	0.18%
35	wikicategory_English-language_films	9714	0.178%
38	wikicategory_Albums_produced_by_Marti_Frederiksen	9282	0.17%
40	wordnet_player_110439851	9135	0.168%
41	wordnet_town_108665504	9119	0.167%
43	wordnet_municipality_108626283	8740	0.1605%
47	wikicategory_Susan_Herndon_albums	8309	0.152%
48	wikicategory_Physics_journals	7643	0.14%
50	wordnet_magazine_106595351	7235	0.1329%