AI6122 Text Data Management & Analysis

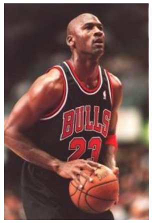
Topic: Entity Linking

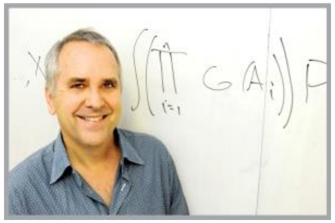
NER and EL

- Named-entity recognition (NER)
  - The task to locate and classify named entities in text into pre-defined categories
    - **names** of persons**,** organizations**,** locations,
    - expressions of times, quantities, monetary values, percentages, **etc**.
  - Example: [Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

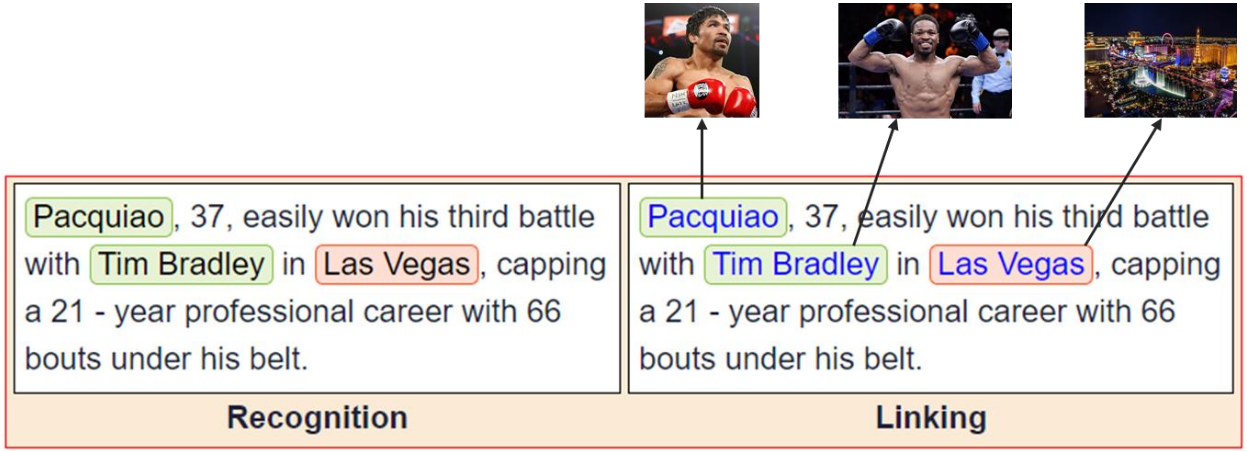
- Entity linking (EL)
  - The task of determining the ***identity of entities mentioned in text***, with reference to a knowledge base.

  - Example: Michael Jordan will give a talk at the conference





Entity Linking

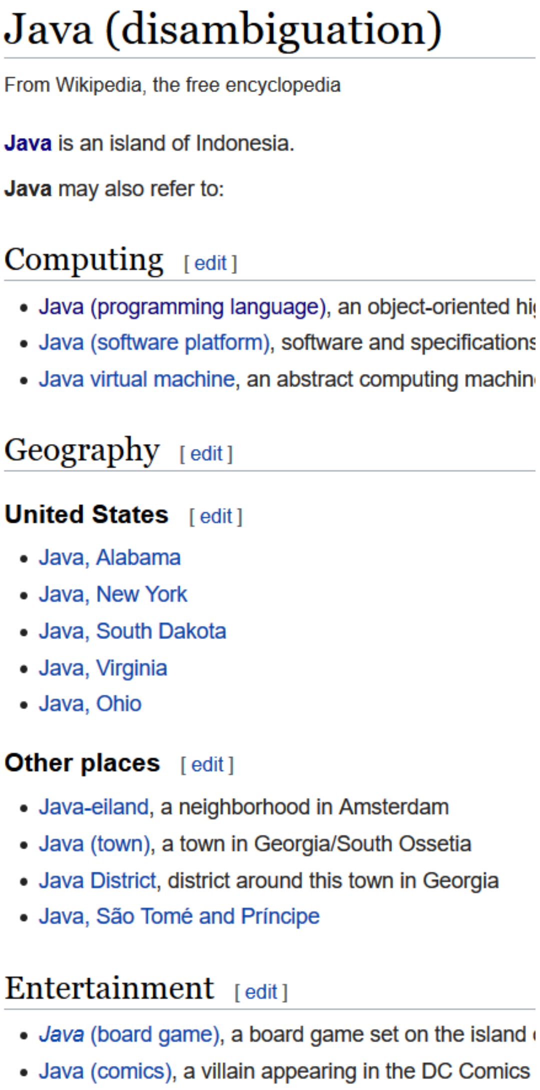


Pacquiao, 37, easily won his third battle with Tim Bradley in Las Vegas, capping a 21 - year professional career with 66 bouts under his belt.

Recognition

Pacquiao, 37, easily won his third battle with Tim Bradley in Las Vegas, capping a 21 - year professional career with 66 bouts under his belt.

Linking



Java (disambiguation)

From Wikipedia, the free encyclopedia

Java is an island of Indonesia.

Java may also refer to:

Computing [edit]

- Java (programming language), an object-oriented hi
- Java (software platform), software and specifications
- Java virtual machine, an abstract computing machin

Geography [edit]

United States [edit]

- Java, Alabama
- Java, New York
- Java, South Dakota
- Java, Virginia
- Java, Ohio

Other places [edit]

- Java-eiland, a neighborhood in Amsterdam
- Java (town), a town in Georgia/South Ossetia
- Java District, district around this town in Georgia
- Java, São Tomé and Príncipe

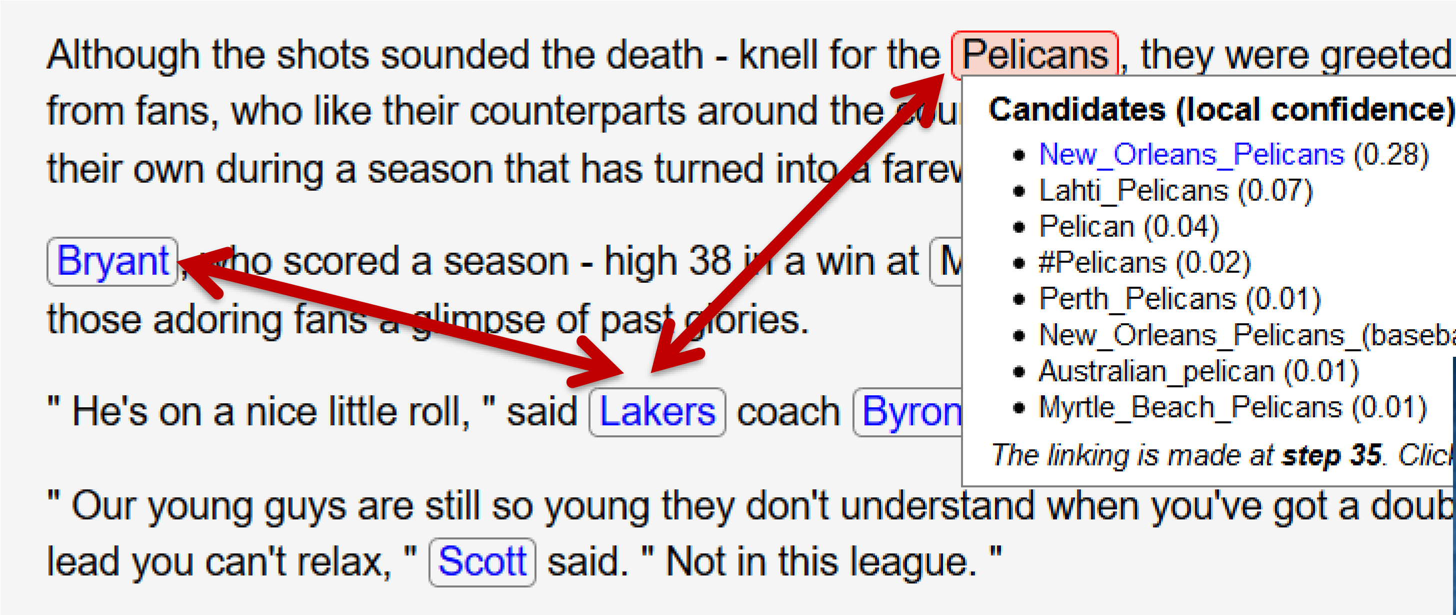
Entertainment [edit]

- *Java* (board game), a board game set on the island
- Java (comics), a villain appearing in the DC Comics

Local contexts:

- Probability of an entity given the mention’s surface form
- String similarity features between the mention’s surface form and the entity’s title (e.g., prefix, suffix, abbreviation…)
- Semantic similarity between the candidate entity and the mention’s surrounding context.

Collective Context



Although the shots sounded the death - knell for the Pelicans, they were greeted by cheers from fans, who like their counterparts around the cou their own during a season that has turned into a farew

Bryant, who scored a season - high 38 in a win at M those adoring fans a glimpse of past glories.

" He's on a nice little roll, " said Lakers coach Byron

" Our young guys are still so young they don't understand when you've got a doub lead you can't relax, " Scott said. " Not in this league. "

Candidates (local confidence):

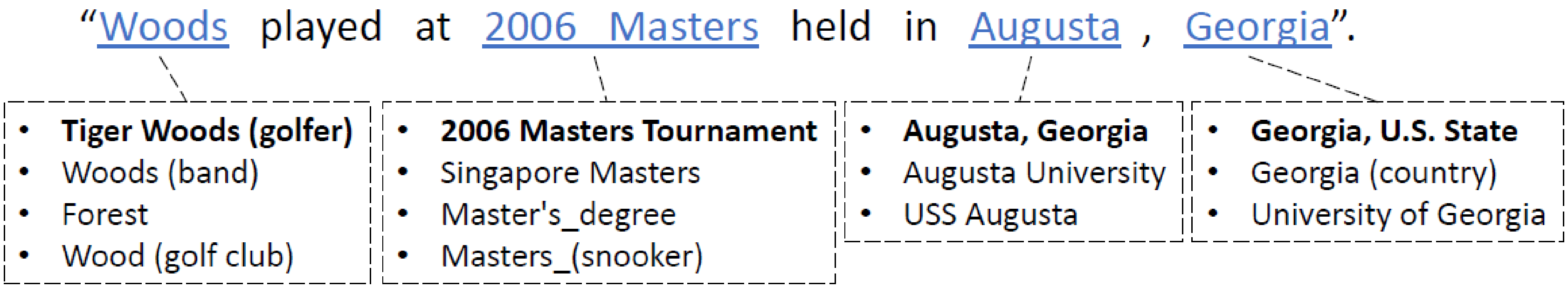
- New_Orleans_Pelicans (0.28)
- Lahti_Pelicans (0.07)
- Pelican (0.04)
- #Pelicans (0.02)
- Perth_Pelicans (0.01)
- New_Orleans_Pelicans_(baseball) (0.01)
- Australian_pelican (0.01)
- Myrtle_Beach_Pelicans (0.01)

The linking is made at ***step 35****. Clic*

Collective context:

- Coherence between **linked entities** in a document

Coherence between linked entities (in a document)



"Woods played at 2006 Masters held in Augusta , Georgia".

- **Tiger Woods (golfer)**
- Woods (band)
- Forest
- Wood (golf club)

- **2006 Masters Tournament**
- Singapore Masters
- Master's_degree
- Masters_(snooker)

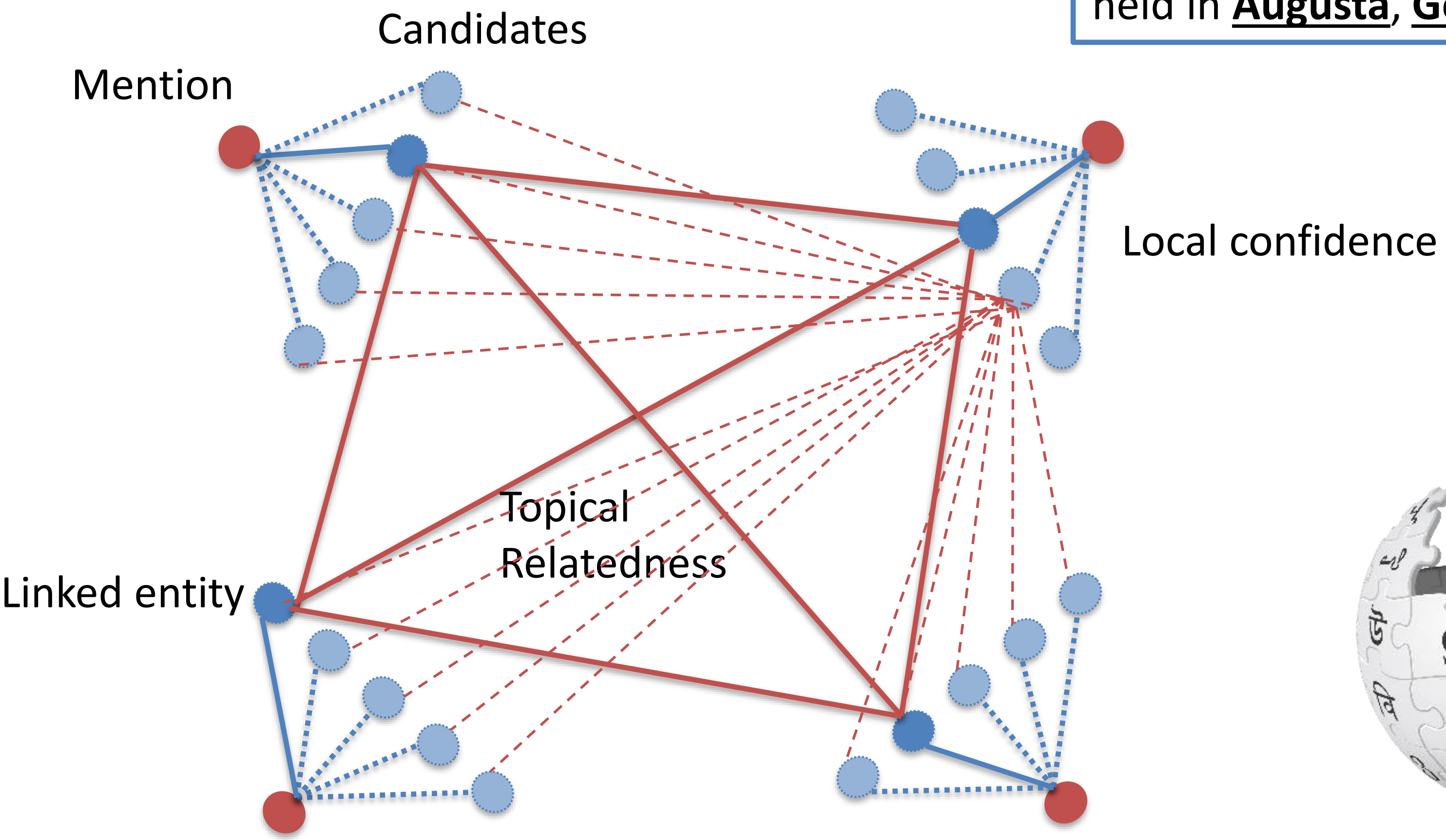
- **Augusta, Georgia**
- Augusta University
- USS Augusta

- **Georgia, U.S. State**
- Georgia (country)
- University of Georgia

What does coherence mean?

What does coherence mean?

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



Candidates

Mention

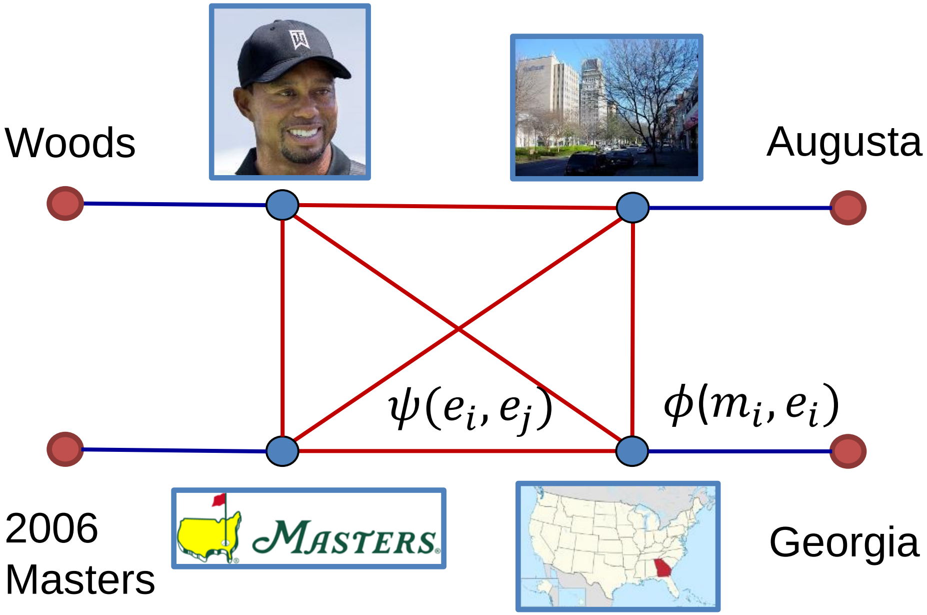
Local confidence

Topical Relatedness

Linked entity

Assumption: All pairs of linked entities are related

“**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**”



Woods

Augusta

$\psi(e_i, e_j)$

$\phi(m_i, e_i)$

2006 Masters

Georgia

- $\psi(e_i, e_j)$
  - – relatedness of linked entities
- $\phi(m_i, e_i)$
  - – local confidence score

Collective Linking: Assumption

- **All-Link**: all pairs of linked entities are related

$$\Gamma^* = \arg\max_{\Gamma} \left[\underbrace{\sum_{i=1}^{N} \phi(m_i, e_i)}_{\text{Local confidence}} + \underbrace{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \psi(e_i, e_j)}_{\text{Global coherence}} \right]$$

  - Utilize of **disambiguation context** Γ'

$$\Gamma^* = \arg\max_{\Gamma} \sum_{i=1}^{N} \left[\phi(m_i, e_i) + \sum_{e_j \in \Gamma'} \psi(e_i, e_j) \right]$$

Collective Linking: Assumption

- Disambiguation context Γ′ not always available
  - Contribution from both unambiguous and ambiguous mentions.
  - $S_{ij}\ (e_i)$: support for label e_i from mention m_j

$$S_{ij}(e_i) = \max_{e_j} \left[\phi(m_j, e_j) + \psi(e_i, e_j) \right]$$

$$e_i = \arg\max_{e_i} \left[\phi(m_i, e_i) + \sum_{j=1, j \neq i}^{N} S_{ij}(e_i) \right]$$

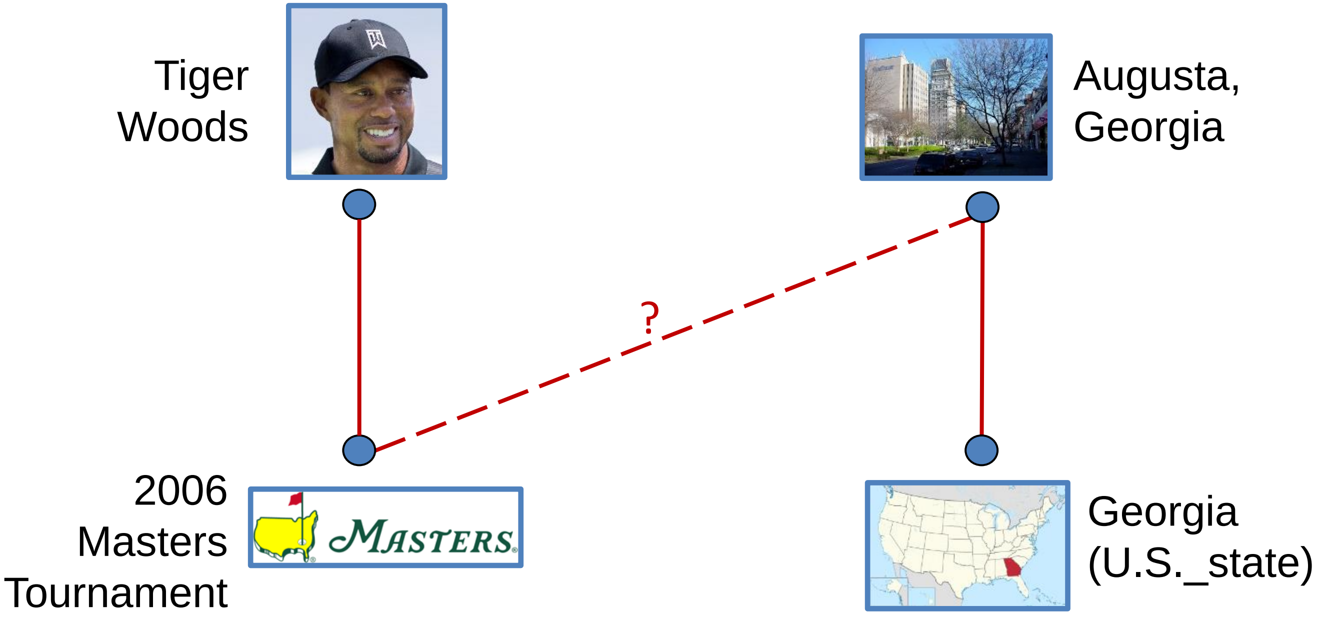
  - best performance is obtained by considering evidence from **<u>not all</u>** but only <u>top-k</u> supporting mentions

- **Single-Link**: consider only the most related evidence

$$\Gamma^* = \arg\max_{\Gamma} \sum_{i=1}^{N} \left[\phi(m_i, e_i) + \max_{j=1}^{N} \psi(e_i, e_j) \right]$$

Are mentioned entities densely connected?

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



Tiger Woods

Augusta, Georgia

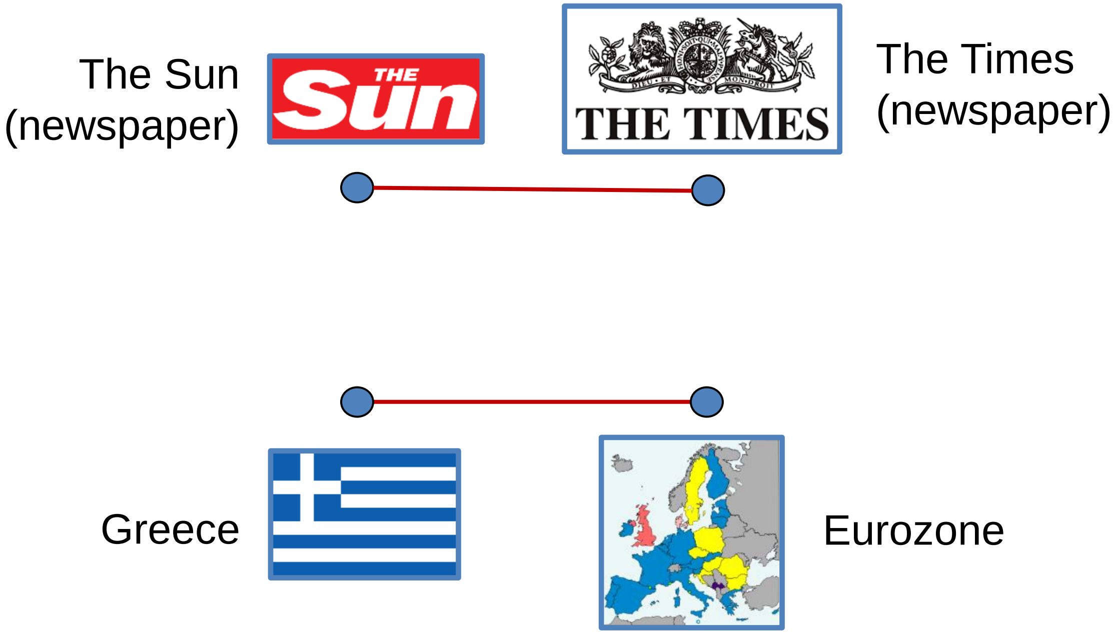
?

2006 Masters Tournament

Georgia (U.S._state)

Are mentioned entities densely connected?

“**The Sun** and **The Times** reported that **Greece** will have to leave the **Euro** soon”.



The Sun
(newspaper)

The Times
(newspaper)

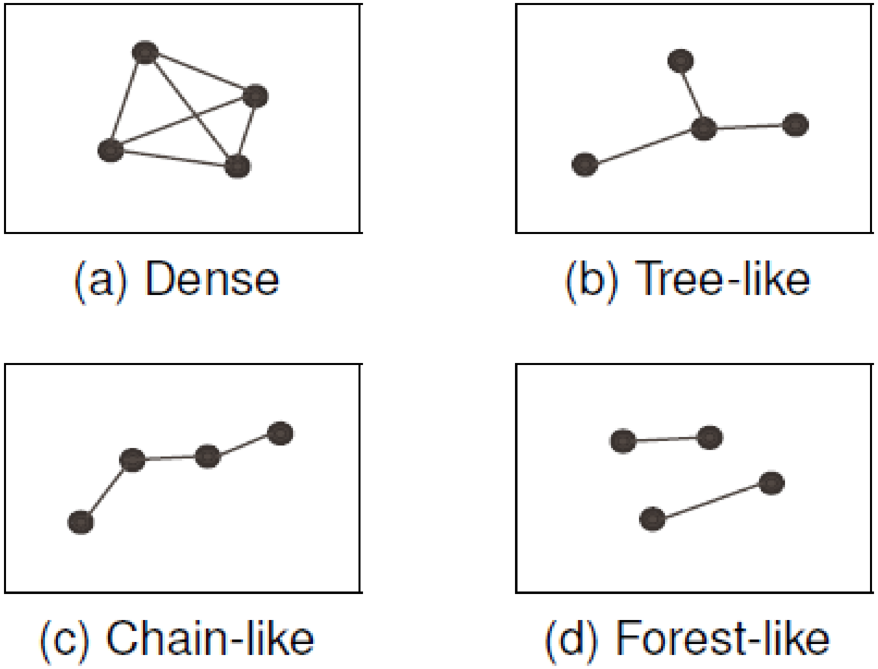
Greece

Eurozone

Complete-pairwise coherence is not always necessary

Complete-pairwise coherence is not always necessary?

- Measure the degree of coherence in real datasets
  - Average degree of entity relatedness graph which consists of high-weighted edges.
  - Possible connection patterns



$N-1$

(a) Dense

$2\frac{N-1}{N}$

(b) Tree-like

$2\frac{N-1}{N}$

(c) Chain-like

1

(d) Forest-like

Pairwise Coherence (Relatedness) Measure

- Wikipedia Link-based Measure

$$WLM(e_1, e_2) = 1 - \frac{\log(\max(|U_1|, |U_2|) + 1) - \log(|U_1 \cap U_2| + 1)}{\log(|W| + 1) - \log(\min(|U_1|, |U_2|) + 1)}$$

- Normalized Jaccard Similarity

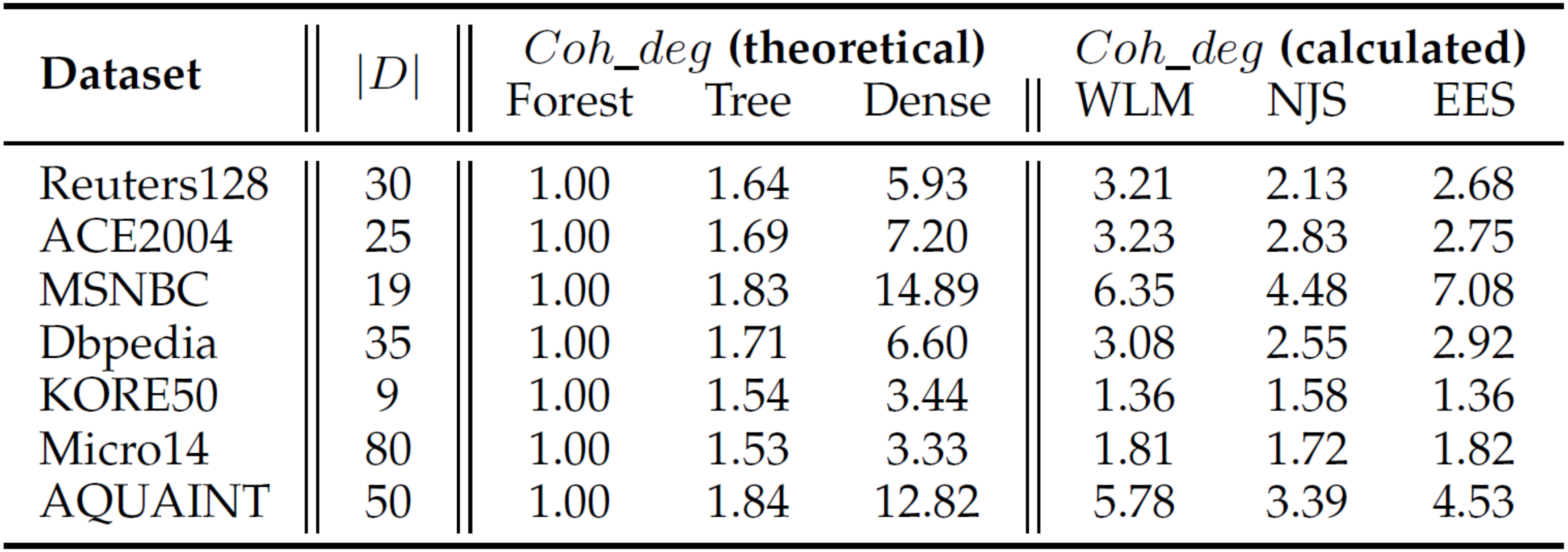
$$NJS(e_1, e_2) = \frac{\log(|U_1 \cap U_2| + 1)}{\log(|U_1 \cup U_2| + 1)}$$

- Embedding Similarity

$$EES(e_1, e_2) = cos(embeding(e_1), embeding(e_2))$$

More About Coherence Analysis

Filtered graph by edge weight: the maximum value such that every node has at least one edge



Dataset	$\lvert D\rvert$	Coh_deg **(theoretical)**			Coh_deg **(calculated)**		
		Forest	Tree	Dense	WLM	NJS	EES
Reuters128	30	1.00	1.64	5.93	3.21	2.13	2.68
ACE2004	25	1.00	1.69	7.20	3.23	2.83	2.75
MSNBC	19	1.00	1.83	14.89	6.35	4.48	7.08
Dbpedia	35	1.00	1.71	6.60	3.08	2.55	2.92
KORE50	9	1.00	1.54	3.44	1.36	1.58	1.36
Micro14	80	1.00	1.53	3.33	1.81	1.72	1.82
AQUAINT	50	1.00	1.84	12.82	5.78	3.39	4.53

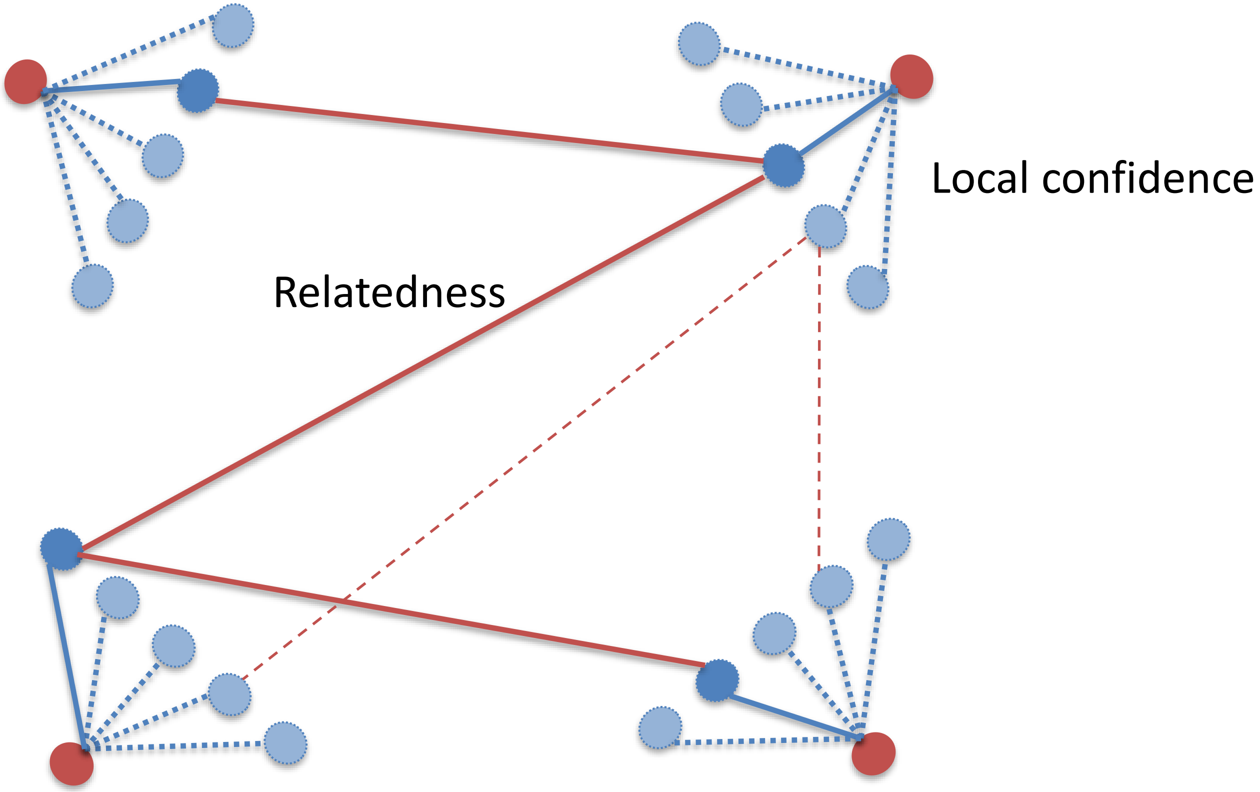
In general, the calculated values lie closer to tree (or chain) form's expected values rather than that of the dense form.

Tree-based Objective for Collective Linking

- **MINTREE** Coherence Measure**.**
  - Given a set of entities V and its associated entity relatedness graph $G(V;E)$, the edges connecting all pairs of entities are weighted by a semantic distance.
  - The coherence of the graph G is defined as the weight of the minimum-spanning tree that can be formed in G.
  - Semantic distance

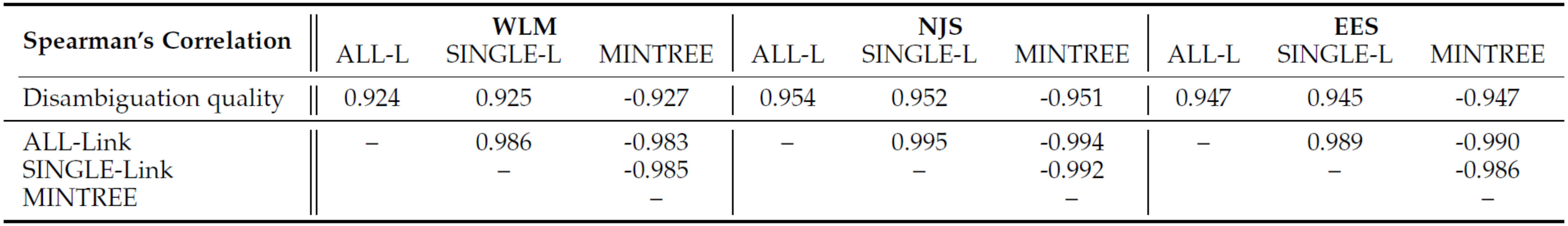
$$d(e_i, e_j) = 1 - \frac{\phi(m_i, e_i) + \psi(e_i, e_j) + \phi(m_j, e_j)}{3}$$

MINTREE coherence



$$d(e_i, e_j) = 1 - \frac{\phi(m_i, e_i) + \psi(e_i, e_j) + \phi(m_j, e_j)}{3}$$

ALL-Link, SINGLE-Link, and MINTREE

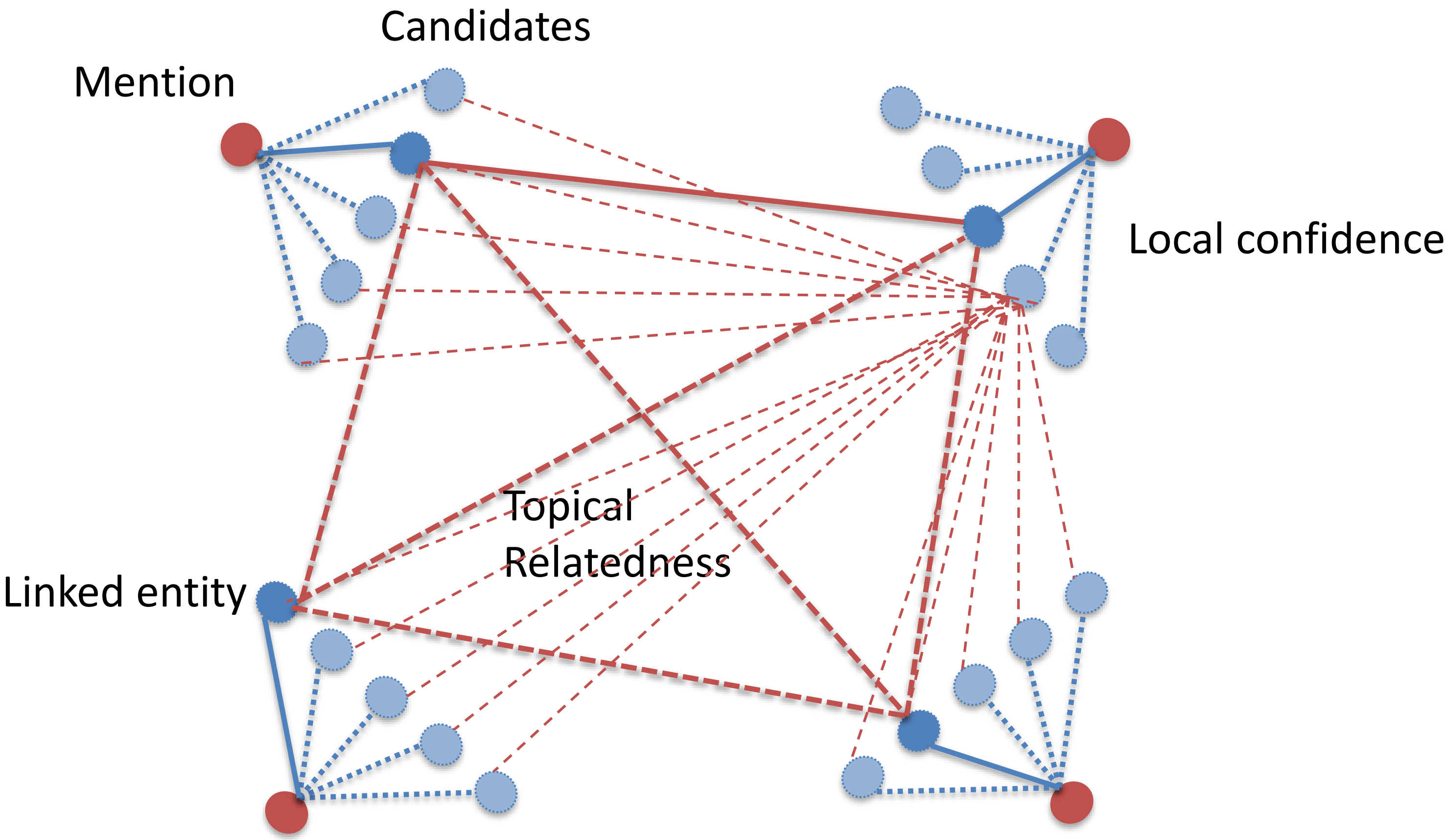


Spearman's Correlation		WLM			NJS			EES	
	ALL-L	SINGLE-L	MINTREE	ALL-L	SINGLE-L	MINTREE	ALL-L	SINGLE-L	MINTREE
Disambiguation quality	0.924	0.925	-0.927	0.954	0.952	-0.951	0.947	0.945	-0.947
ALL-Link	–	0.986	-0.983	–	0.995	-0.994	–	0.989	-0.990
SINGLE-Link		–	-0.985		–	-0.992		–	-0.986
MINTREE			–			–			–

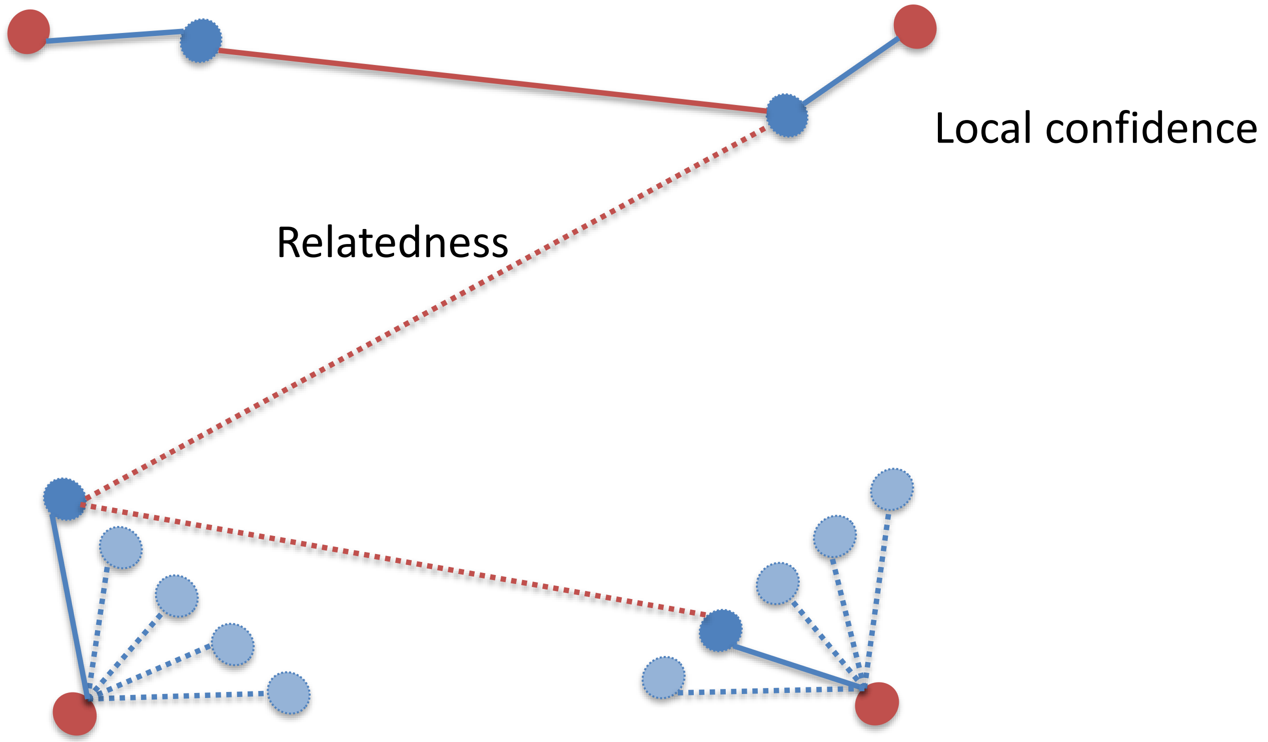
- Given a document with a set of mentions
- Start with all mentions assigned to wrong entities
- At each step, make one mention links to its current entity
  - Increase number of correct decision by one
  - Compute the objective score

- Spearman's Correlation
  - The number of correct decisions
  - The objective scores

MINTREE coherence



MINTREE coherence

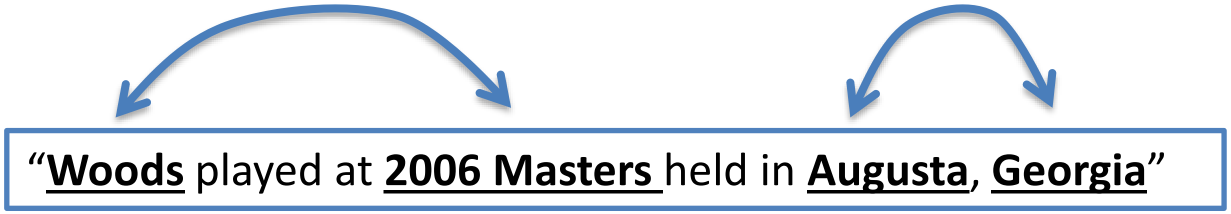


- Existing algorithms for minimum spanning tree cannot be applied directly

Pair-Linking

- We do not need to look at all other entity when deriving linking decisions.

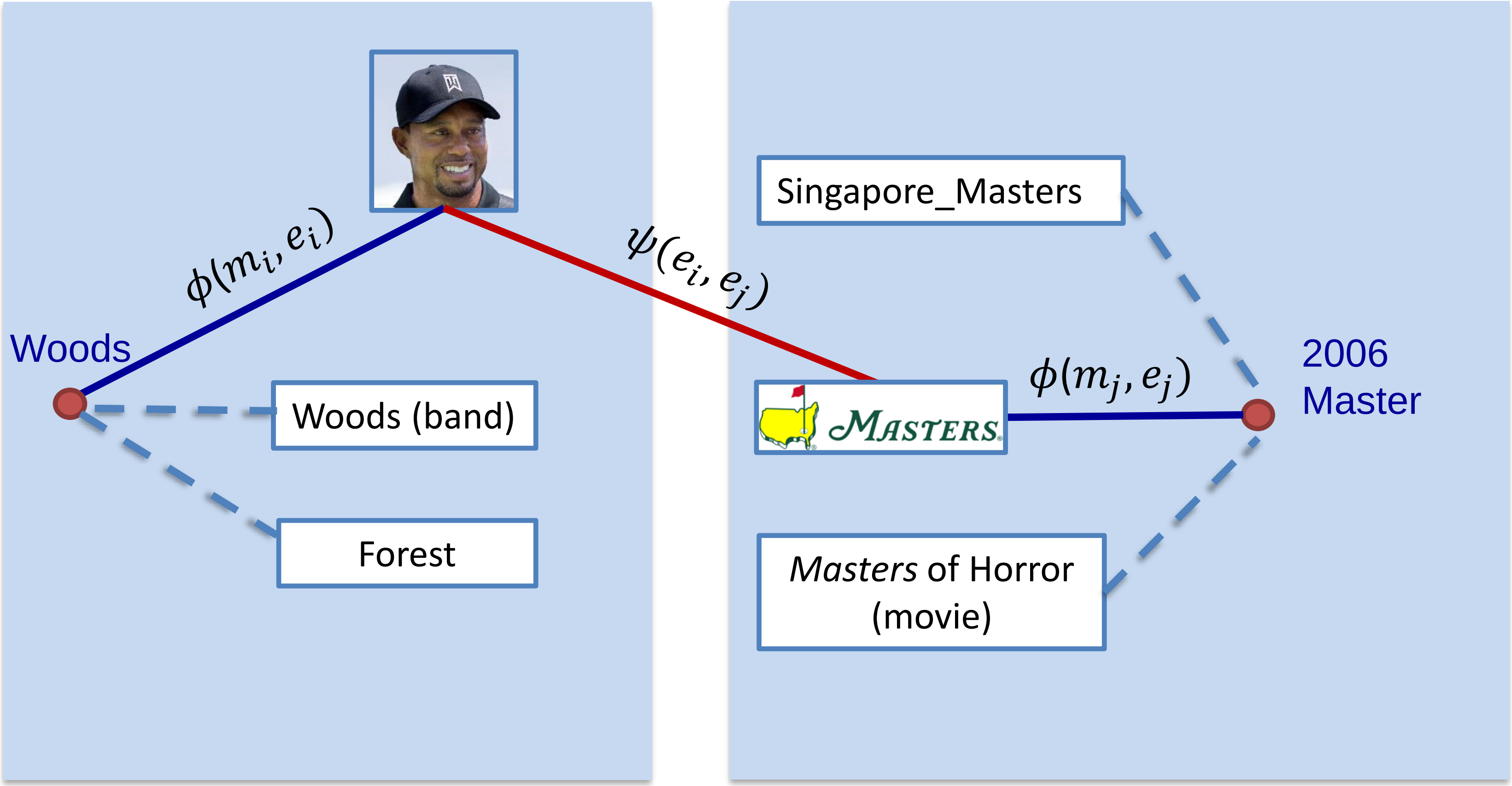
- Interactively resolve a pair of mention at each step, from the more confident pairs to less confident pairs.



“**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**”

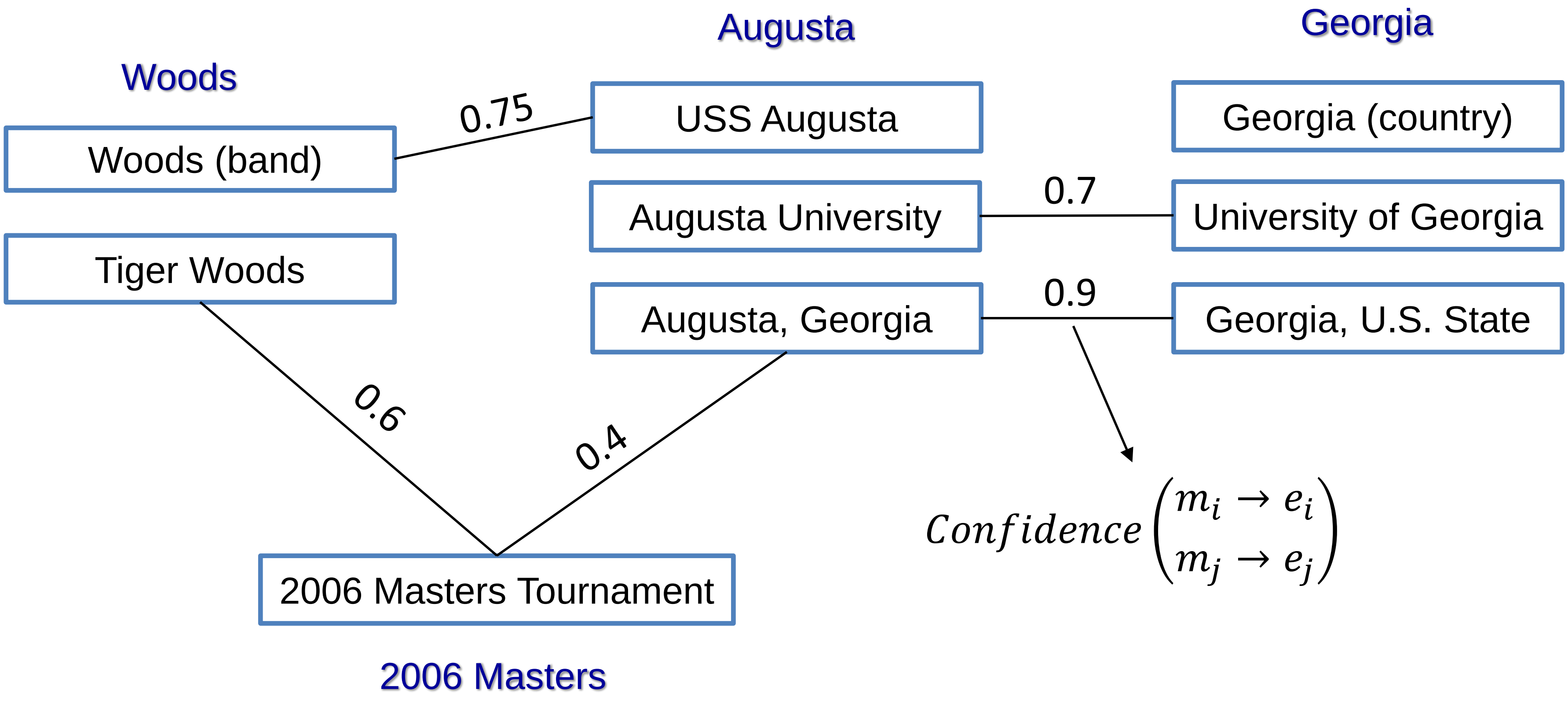
Pair-Linking: Local confidence + Coherence

- Pairwise confidence



Pair-Linking Example

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



Woods

Woods (band)

Tiger Woods

Augusta

USS Augusta

Augusta University

Augusta, Georgia

Georgia

Georgia (country)

University of Georgia

Georgia, U.S. State

2006 Masters Tournament

2006 Masters

0.75

0.7

0.9

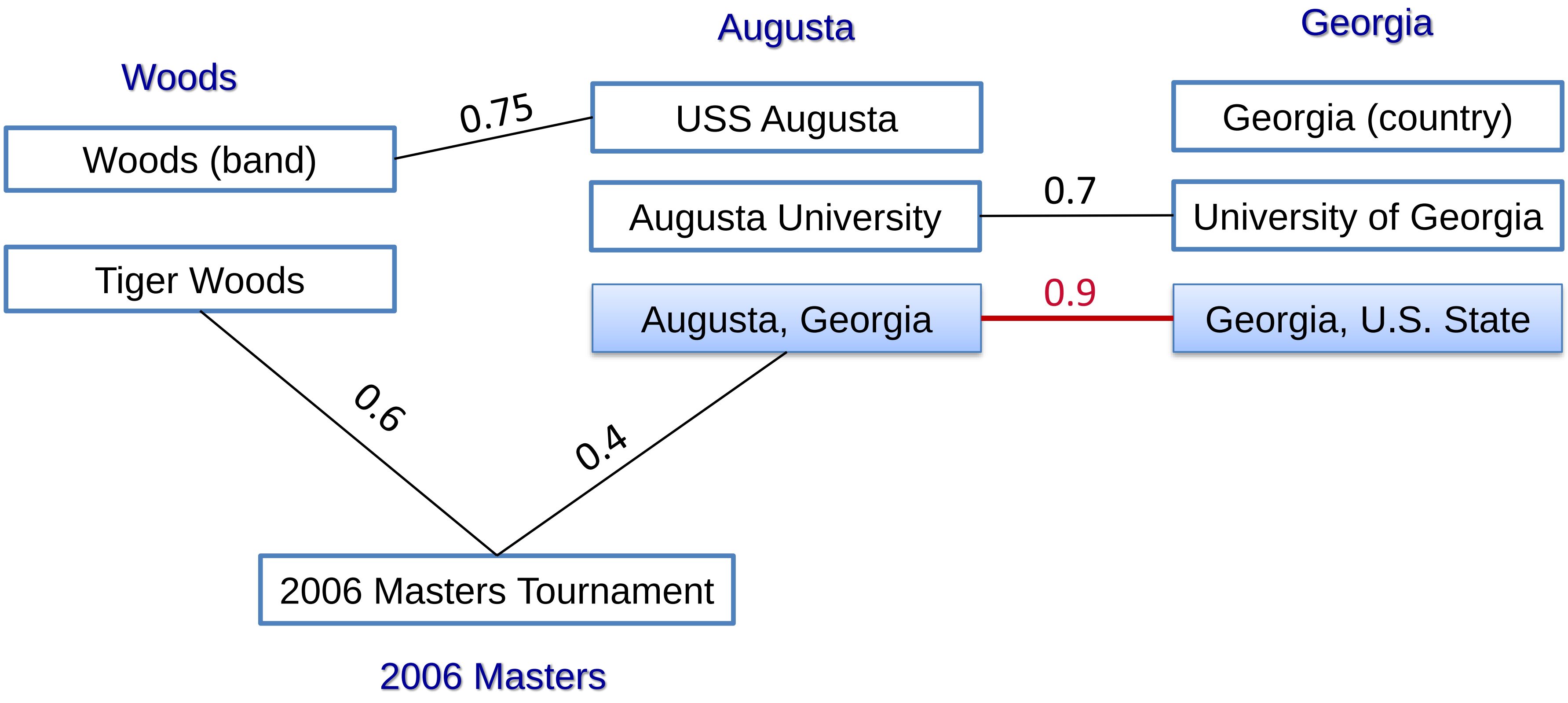
0.6

0.4

$$Confidence\begin{pmatrix} m_i \rightarrow e_i \\ m_j \rightarrow e_j \end{pmatrix}$$

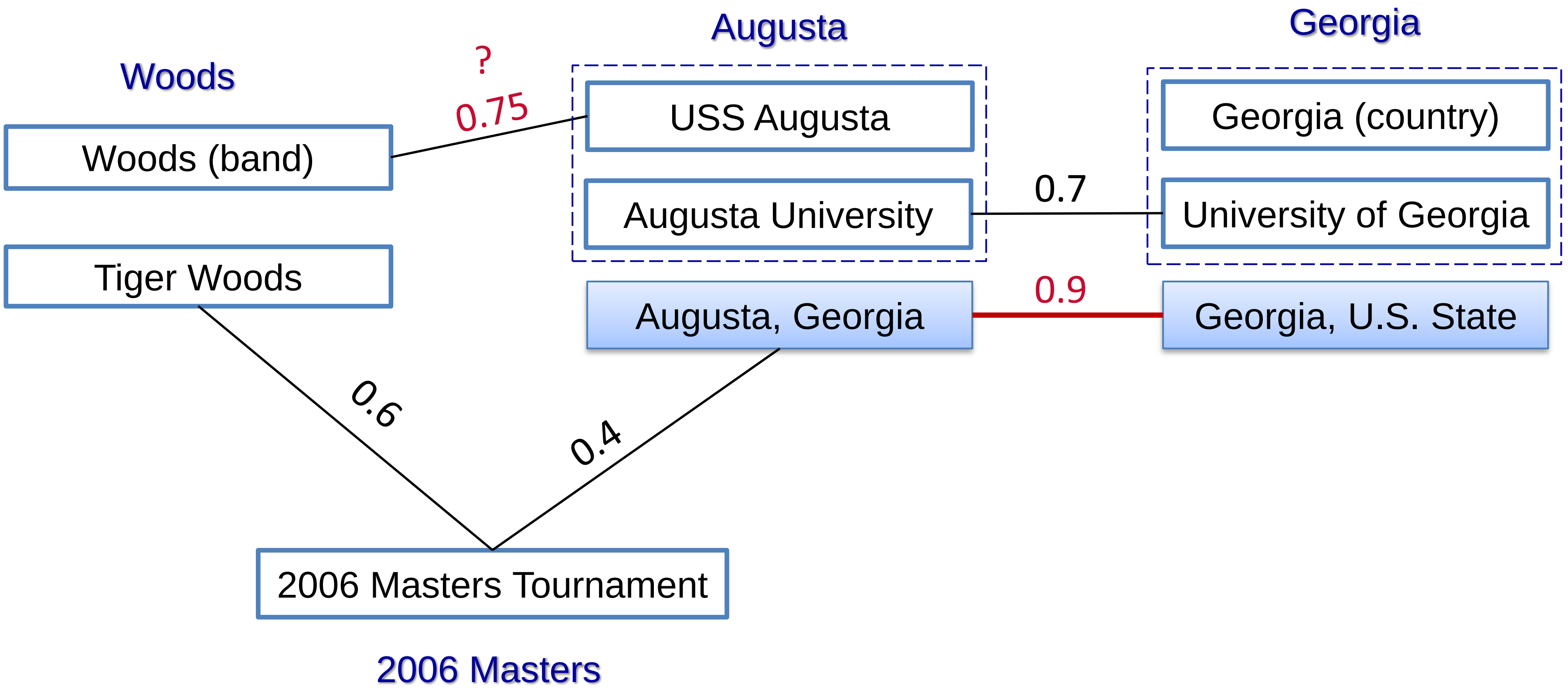
Pair-Linking Example

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



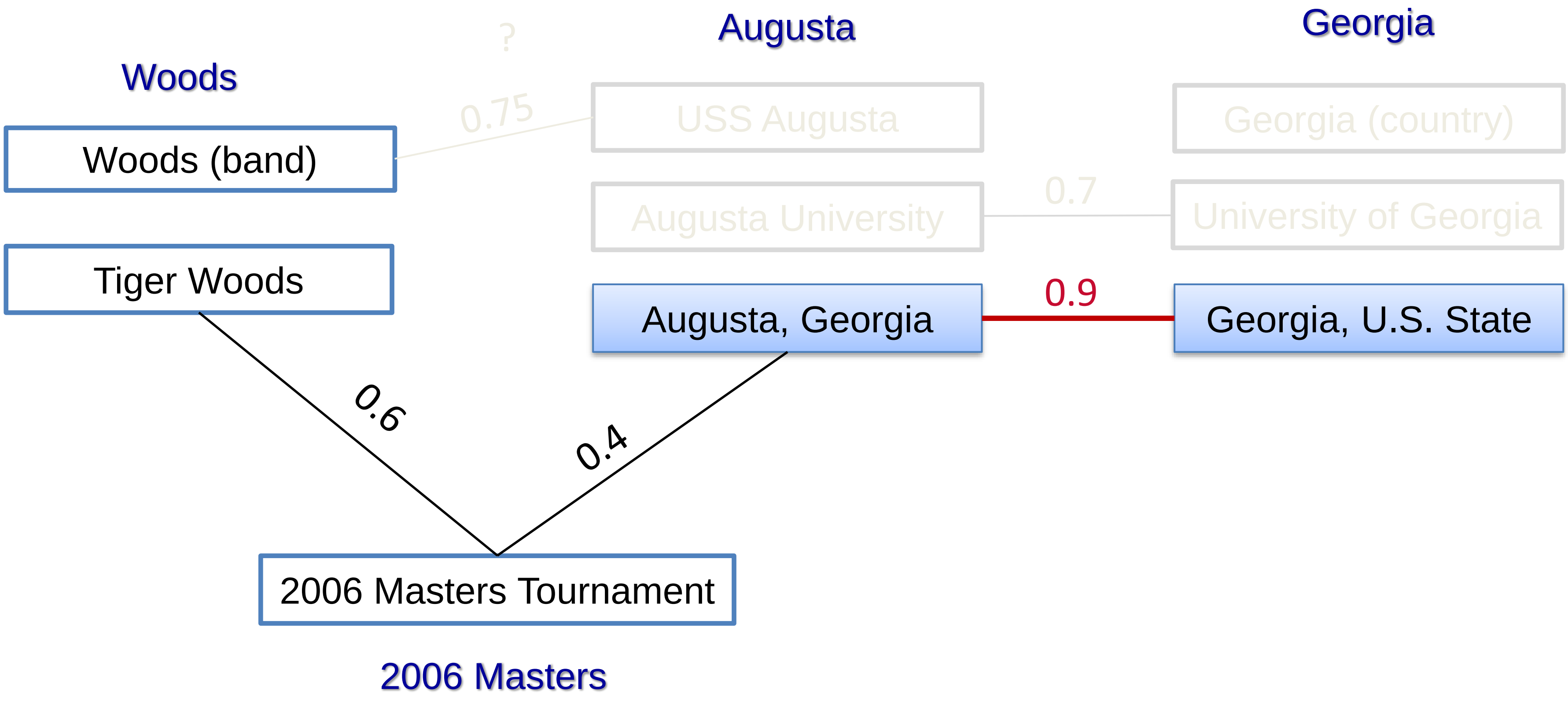
Pair-Linking Example

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



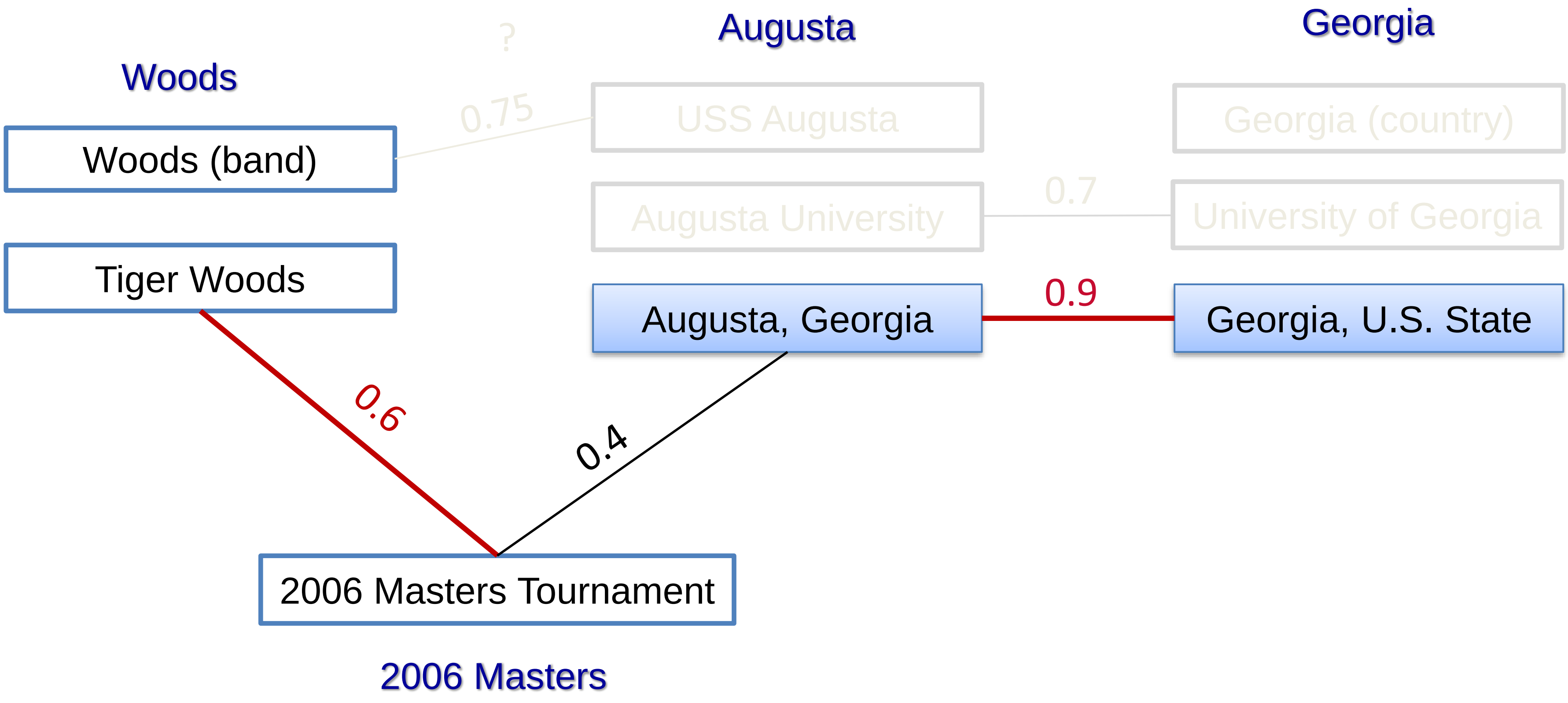
Pair-Linking Example

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



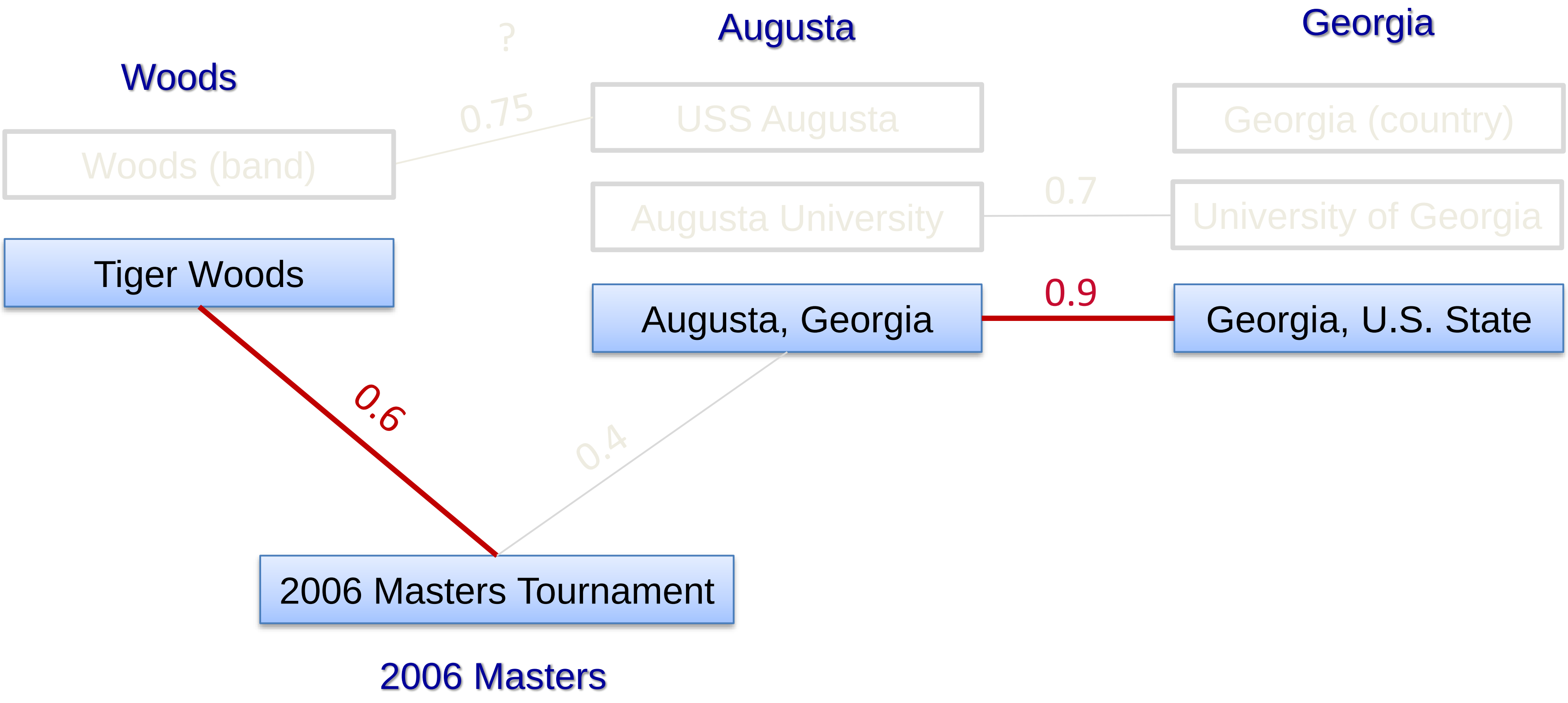
Pair-Linking Example

"**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**"



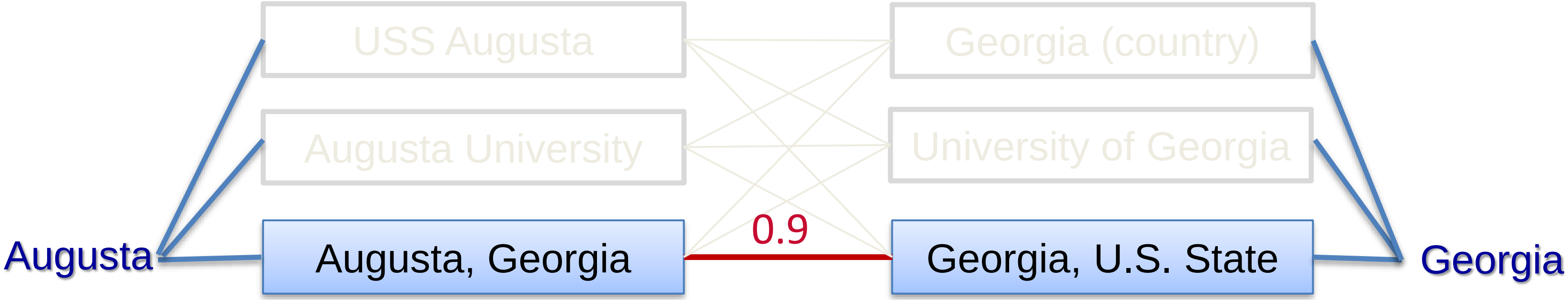
Pair-Linking Example

“**Woods** played at **2006 Masters** held in **Augusta**, **Georgia**”

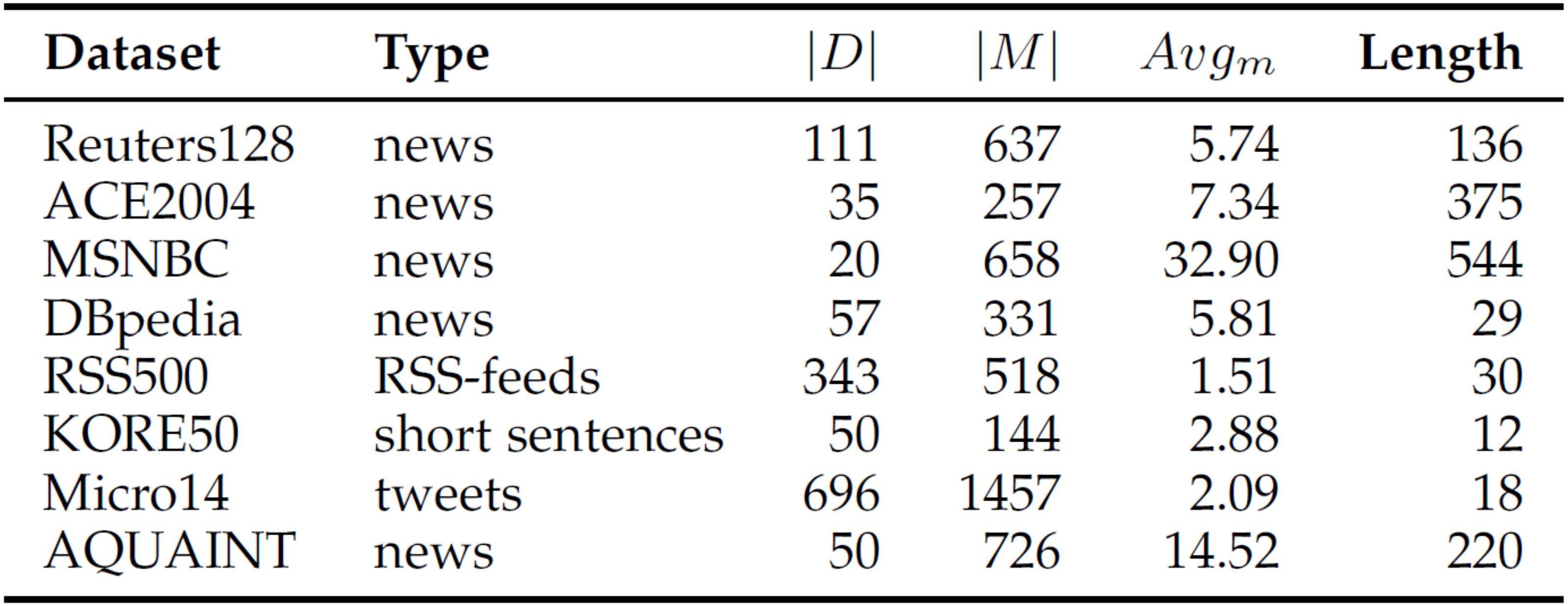


Pair-Linking is Super Fast

- Pair-Linking cares about the pair with highest confidence score.
  - Use priority queue to store and retrieve the pair.
  - Utilize early stop to avoid scanning all possible pair of candidates.



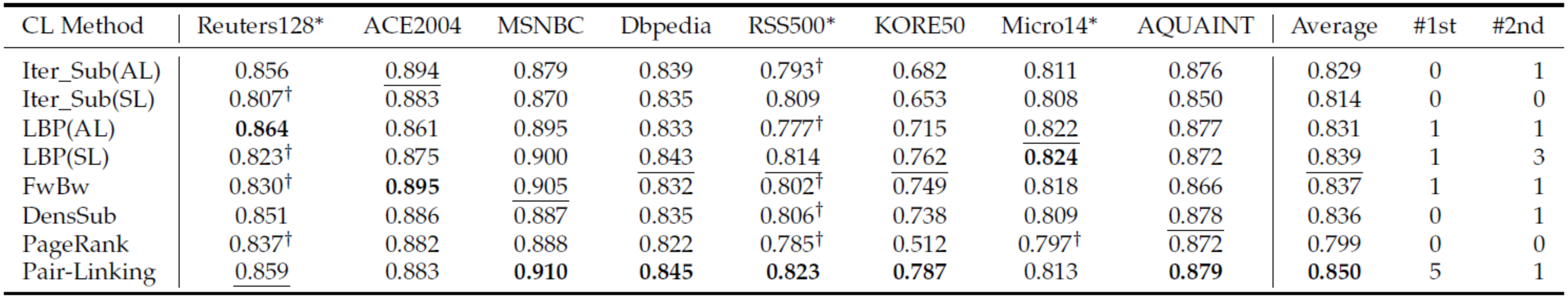
Experiment: 8 benchmark datasets



Dataset	Type	$\|D\|$	$\|M\|$	Avg_m	Length
Reuters128	news	111	637	5.74	136
ACE2004	news	35	257	7.34	375
MSNBC	news	20	658	32.90	544
DBpedia	news	57	331	5.81	29
RSS500	RSS-feeds	343	518	1.51	30
KORE50	short sentences	50	144	2.88	12
Micro14	tweets	696	1457	2.09	18
AQUAINT	news	50	726	14.52	220

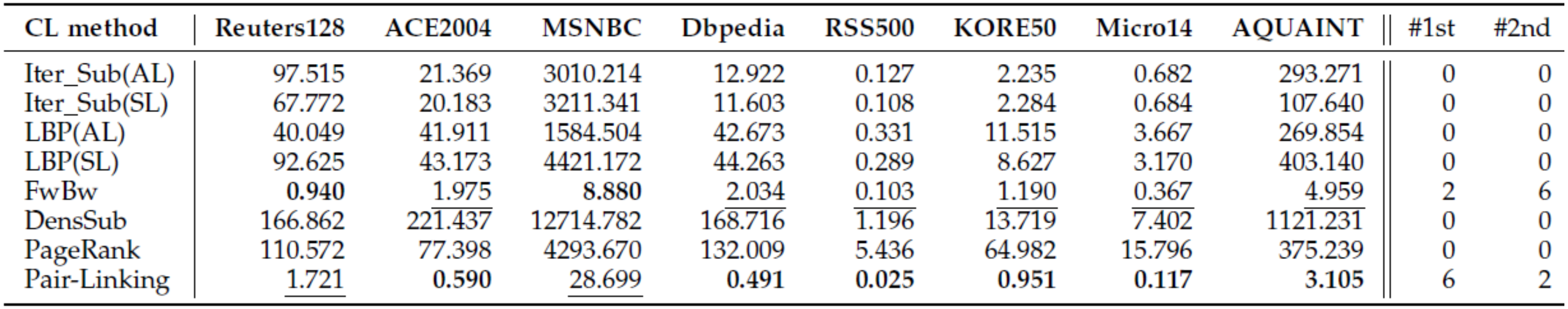
Pair-Linking Performance

- Linking accuracy (F_1) Normalized Jaccard + Embedding Sim



CL Method	Reuters128*	ACE2004	MSNBC	Dbpedia	RSS500*	KORE50	Micro14*	AQUAINT	Average	#1st	#2nd
Iter_Sub(AL)	0.856	0.894	0.879	0.839	0.793†	0.682	0.811	0.876	0.829	0	1
Iter_Sub(SL)	0.807†	0.883	0.870	0.835	0.809	0.653	0.808	0.850	0.814	0	0
LBP(AL)	**0.864**	0.861	0.895	0.833	0.777†	0.715	0.822	0.877	0.831	1	1
LBP(SL)	0.823†	0.875	0.900	0.843	0.814	0.762	**0.824**	0.872	0.839	1	3
FwBw	0.830†	**0.895**	0.905	0.832	0.802†	0.749	0.818	0.866	0.837	1	1
DensSub	0.851	0.886	0.887	0.835	0.806†	0.738	0.809	0.878	0.836	0	1
PageRank	0.837†	0.882	0.888	0.822	0.785†	0.512	0.797†	0.872	0.799	0	0
Pair-Linking	0.859	0.883	**0.910**	**0.845**	**0.823**	**0.787**	0.813	**0.879**	**0.850**	5	1

- Speed: average number of milli-seconds per document

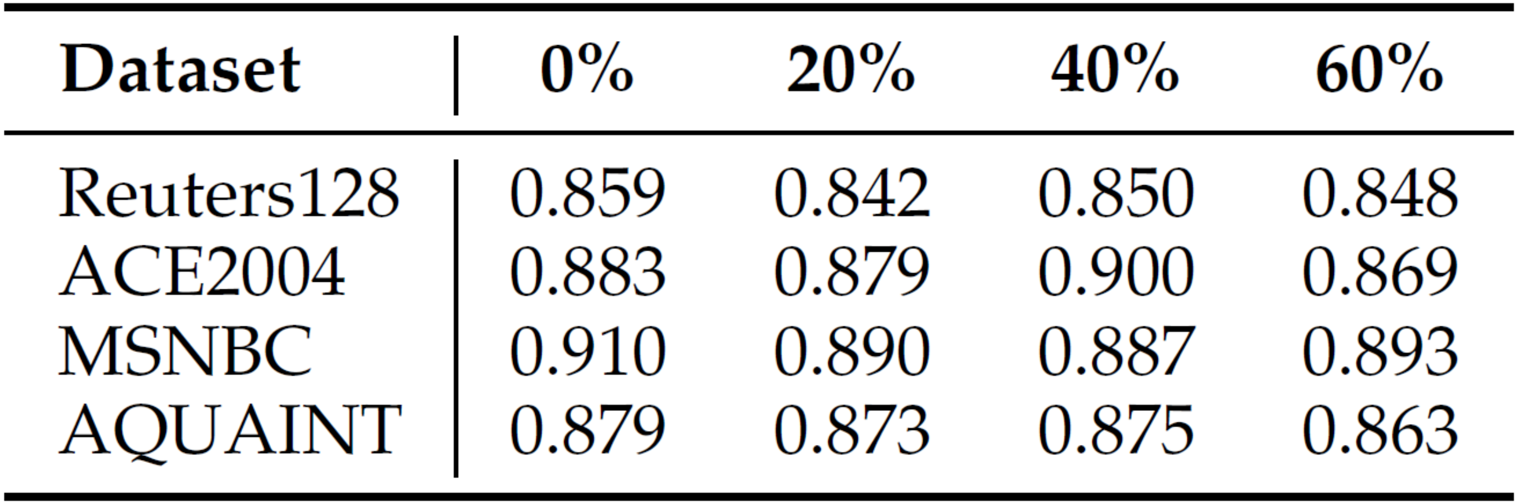


CL method	Reuters128	ACE2004	MSNBC	Dbpedia	RSS500	KORE50	Micro14	AQUAINT	#1st	#2nd
Iter_Sub(AL)	97.515	21.369	3010.214	12.922	0.127	2.235	0.682	293.271	0	0
Iter_Sub(SL)	67.772	20.183	3211.341	11.603	0.108	2.284	0.684	107.640	0	0
LBP(AL)	40.049	41.911	1584.504	42.673	0.331	11.515	3.667	269.854	0	0
LBP(SL)	92.625	43.173	4421.172	44.263	0.289	8.627	3.170	403.140	0	0
FwBw	**0.940**	1.975	**8.880**	2.034	0.103	1.190	0.367	4.959	2	6
DensSub	166.862	221.437	12714.782	168.716	1.196	13.719	7.402	1121.231	0	0
PageRank	110.572	77.398	4293.670	132.009	5.436	64.982	15.796	375.239	0	0
Pair-Linking	1.721	**0.590**	28.699	**0.491**	**0.025**	**0.951**	**0.117**	**3.105**	6	2

(*) Performances on ACE2004, RSS500 and Micro2014 are not shown here.

NIL mention: cannot link to any entity in knowledge base

- How robust is Pair-Linking if NIL mentions are presenting in a document?
- Randomly remove some ground truths from candidate entities
- F_1 score vs percentage of NIL mentions (as noises)



Dataset	0%	20%	40%	60%
Reuters128	0.859	0.842	0.850	0.848
ACE2004	0.883	0.879	0.900	0.869
MSNBC	0.910	0.890	0.887	0.893
AQUAINT	0.879	0.873	0.875	0.863

Summary

- Relook at the assumption of ALL-Link in collective linking
- Study the average degree of coherence graph for collective linking
  -
- Propose MINTREE objective and design Pair-Linking
  - High accuracy
  - Low computational cost

Pair-Linking for Collective Entity Disambiguation: Two Could Be Better Than All. IEEE TKDE. 31(7): 1383-1396 (2019)