Entity Linking with a Knowledge Base for Heterogeneous Data

Jianyong Wang Department of Computer Science and Technology Tsinghua University jianyong@tsinghua.edu.cn

Joint work with Wei Shen (Tsinghua), Ping Luo (HP), and Min Wang (HP)

Outline

- Introduction to entity linking with a knowledge base
 - Motivation & definition
- Entity linking for unstructured Web documents
- Entity linking for structured Web lists/tables
- Entity linking for Tweets
- Conclusion

- Knowledge base construction from heterogeneous data
 - Better user experience of information search and recommendation is always in great demand
 - Semantic search on the Web, Deep Q/A in NL, ...



ho was US president wh	en Barack Obama was born
ut interpretation: United States presi ut: ohn F. Kennedy	dent Barack Obama date of birth
is information.	
sic information: official position	President (35 th)
ic information: official position country	President (35 th) United States
official position country political affiliation	President (35 th) United States Democrat
official position country political affiliation start date	President (35 th) United States Democrat 20. January 1961 (49 years 3 months 26 days ago)
sic information: official position country political affiliation start date end date	President (35 th) United States Democrat 20. January 1961 (49 years 3 months 26 days ago) 22. November 1963 (46 years 5 months 23 days ago)

- Knowledge base construction from heterogeneous data
 - Better user experience of information search and recommendation is always in great demand
 - Semantic search on the Web, Deep Q/A in NL, ...

Free texts, Tables, Lists, Twitter, Weibo, ...

• Structured knowledge discovery from heterogeneous data

Entities, semantic categories, mutual relations, ...





- Existing Knowledge Bases are far from perfect
 - They are large, but their coverage is still low
 - Popular or well-known person, place or thing
 - E.g., Google's Knowledge graph

As of 2012, its semantic network contained over 570 million objects and more than 18 billion facts about the relationships between these different objects which are used to understand the meaning of the keywords entered for the search

Source: http://en.wikipedia.org/wiki/ Google_Knowledge_Graph



- Knowledge base population
 - Automatically populating and enriching the existing KB with the newly extracted facts
 - Why?
 - Limited coverage for existing KBs
 - As world evolves
 - New facts come into existence
- Entity linking is inherently considered as an important subtask for knowledge base population

Entity	Relation	Entity
"Michael Jordan"	isPlayerOf	"Bulls"

"Michael Jordan": Michael Jordan (NBA); Michael Jordan (Professor); Michael Jordan (footballer);
"Bulls": Chicago Bulls; Bulls, New Zealand; Bulls (rugby);



Figure : An example of YAGO

Challenges in Entity Linking

- Entity ambiguity problem
 - Name variations: a named entity may have multiple names
 - National Basketball Association → "NBA"
 - New York City \rightarrow "Big Apple"
 - Entity ambiguity: a name may refer to several different named entities

• "Michael Jordan"

Michael J. Jordan (NBA player)
 Michael I. Jordan (Berkeley professor)
 Michael W. Jordan (footballer)
 Michael Jordan (mycologist)

Outline

- Introduction to entity linking with a knowledge base
 - Motivation & definition
- Entity linking for unstructured Web documents
 - Entities detected from unstructured Web documents: a common scenario
 - The LINDEN framework (WWW'12)
- Entity linking for structured Web lists/tables
 - Entities detected from structured Web lists/tables
 - The LIEGE framework (KDD'12)
- Entity linking for Tweets
 - Entities detected from short and noisy Tweets
 - The KAURI framework (KDD'13)
- Conclusion

Entity linking for unstructured Data —Problem Definition

- Entity linking task
 - Input:
 - A textual named entity mention *m*, already recognized in the unstructured Web document
 - Output:
 - The corresponding real world entity *e* in the knowledge base
 - If the matching entity *e* for entity mention *m* does not exist in the knowledge base, we should return a NIL for *m*

Entity Linking for Unstructured Data —Previous Methods

- Essential step of entity linking
 - Define a similarity measure between the text around the entity mention and the document associated with the entity
- Bag of words model
 - Represent the context as a term vector
 - Measure the **co-occurrence** statistics of terms
 - Cannot capture the semantic knowledge
- Example:
 - Text: Michael Jordan wins NBA champion.
 - Entity name: Michael J. Jordan
 Description text: American basketbal player
 - Entity name: Michael I. Jordan Description text: Berkeley professor in AI

Entity Linking for Unstructured Data —Our solution: The LINDEN Framework

Define four features

- Feature 1: Prior probability
 - Based on the count information
- Semantic network based features
 - Feature 2: Semantic associativity
 - Based on the Wikipedia hyperlink structure
 - Feature 3: Semantic similarity
 - Derived from the taxonomy of YAGO
 - Feature 4: Global coherence



- To rank the candidates, we compute a score by a linear combination of the four features
 - $Score_m(e) = \overrightarrow{w} \cdot F_m(e)$, where $F_m(e) = \langle LP(e|m), SA(e), SS(e), GC(e) \rangle$
 - Use a max-margin technique to automatically learn the weights

 $\overrightarrow{w} \cdot F_m(e^*) - \overrightarrow{w} \cdot F_m(e) \ge 1 - \xi_m \tag{12}$

• Minimize over $\xi_m \ge 0$ and the objective $||\vec{w}||_2^2 + \alpha \Sigma_m \xi_m$



Entity mentions: Michael Jordan, NBA

Entity Linking for Unstructured Data —Our solution: Experimental Study

Table 3: Experimental results over the CZ data set

	# of total	LINDEN		Cucerzan		
	mentions	#	Accu.	#	Accu.	
All	614	581	0.9463	549	0.8941	
Linkable	522	493	0.9444	466	0.8927	
Unlinkable	92	88	0.9565	83	0.9022	

Outline

- Introduction to entity linking with a knowledge base
 - Motivation & definition
- Entity linking for unstructured Web documents
 - Entities detected from unstructured Web documents: a common scenario
 - The LINDEN framework (WWW'12)
- Entity linking for structured Web list/tables
 - Entities detected from structured Web lists/tables
 - The LIEGE framework (KDD'12)
- Entity linking for Tweets
 - Entities detected from short and noisy Tweets
 - The KAURI framework (KDD'13)
- Conclusion

Entity Linking for Structured Data —Motivation

A list of famous football players

Mahdi Abdul-Rahman	A list of best	_		
Mahmoud Abdul-Rauf	selling albun	าร	TATAL	Tiet
Tariq Abdul-Wahad		A list of	web) LIST
Shareef Abdur-Rahim	Thriller	famous ar	tists	
Tom Abernethy	Their Greatest Hits (1971-1	Michael Jackson	A list of NBA	
Forest Able	Led Zeppelin IV		— plavers	
John Abramovic	The Wall	Eagles	• Wil Jones	
Alex Acker	Greatest Hits Volume I & V	Led Zeppelin		
Don Ackerman	Back in Black	Pink Floyd	• While Jones	
Mark Acres	Double Live	Billy Joel	 Adonis Jordan 	
Bud Actor	Come On Over	AC/DC	 Charles Jordan 	
Dud Acton		Garth Brooks	DoAndro Jordan	
Alvan Adams		Shania Twain	DeAndre Sordan	
			 Eddie Jordan 	
			• Jerome Jordan	
			• Ken Jordan 🛛 🧷	NBA player
			Michael Jordan	Berkeley professor
			- Baggia Jordan	
				••• •••

.

Thomas Jordan

Entity Linking for Structured Data —Problem Definition

- List linking task
 - Link the entity mentions that appear in the Web lists with the corresponding real world entities in the knowledge base



Figure: An illustration of the list linking task. The Web list enumerates some best-selling single volume books. Candidate mapping entities from knowledge base for each list item are shown on the right of the figure; true mapping entity for each list item is underlined.

Entity Linking for Structured Data —List Linking

- The list linking task is practically important
 - Knowledge base population and table annotation
- Challenge
 - No textual context
 - Different from the task of linking entities in free text
- Assumption
 - Entities mentioned in a Web list can be any collection of entities that have the same conceptual type

Entity Linking for Structured Data —Our solution: Linking Quality Metric

Prior probability

• Define the popularity of an entity based on the link count information from Wikipedia

Coherence

- The type of the candidate mapping entity should be coherent with the types of the other mapping entities in the same Web list
- Type hierarchy based similarity
- Distributional context similarity

Entity Linking for Structured Data —Our solution: Linking Quality



• We utilize the **max-margin** technique to automatically learn the weight vector which gives proper weights for different features (Details in paper)

Entity Linking for Structured Data —Our solution: Iterative Substitution Algorithm

Algorithm 1 Iterative Substitution Algorithm **Input:** Web list *L*, candidate mapping entity sets *R*. **Output:** mapping entity list M. 1: for each $l_i \in L$ do 2: $m_i^{(0)} = \arg \max_{r_{i,j}} P_{pr}(r_{i,j}), r_{i,j} \in R_i$ 3: end for 4: $M^{(0)} = \{m_i^{(0)} | l_i \in L\}$ 5: iter = 16: while true do for each $l_i \in L$ do 7: for each $r_{i,j} \neq m_i^{(iter-1)} \in R_i$ do 8: $M_{r_{i,j}}^{(iter)} = (M^{(iter-1)} - \{m_i^{(iter-1)}\}) \bigcup \{r_{i,j}\}$ 9: $IncreLQ_{r_{i,j}} = LQ(M_{r_{i,j}}^{(iter)}) - LQ(M^{(iter-1)})$ 10:end for 11: 12: end for $r_{i,j}^{max} = \arg \max_{r_{i,j}} IncreLQ_{r_{i,j}}, \quad r_{i,j} \in R_i, R_i \in R$ 13: 4: if $IncreLQ_{r_{i,i}^{max}} > 0$ then $M^{(iter)} = (M^{(iter-1)} - \{m_i^{(iter-1)}\}) \bigcup \{r_{i,i}^{max}\}$ 15:16: iter + +17: else 18: break end if 9: 20: end while 21: $M = M^{(iter-1)}$

Initialization:

• Pick the candidate entity that has the maximum *prior probability* as the initial estimate of the mapping entity for the list item

Iterative substitution:

• Iteratively refine the mapping entity list to improve its *linking quality*

- When the maximum improvement is smaller than zero, we stop the iteration.
- We prove that this algorithm is guaranteed to **converge**.

Entity Linking for Structured Data —Our solution: Experimental Study

Table 3: Experimental results over Wiki_Manual

Approach	# correctly linked	Accuracy
TableAnno	1419	0.8392
$\text{LIEGE}_{\beta=0,\gamma=0}$	1461	0.8640
$\text{LIEGE}_{\beta=0}$	1519	0.8983
$\text{LIEGE}_{\gamma=0}$	1498	0.8859
$LIEGE_{full}$	1536	0.9083

$$LQ(r_{i,j}) = \alpha * P_{pr}(r_{i,j}) + \beta * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{hr}(r_{i,j}, m_u)$$
$$+ \gamma * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{ds}(r_{i,j}, m_u)$$

Outline

- Introduction to entity linking with a knowledge base
 - Motivation & definition
- Entity linking for unstructured Web documents
 - Entities detected from unstructured Web documents: a common scenario
 - The LINDEN framework (WWW'12)
- Entity linking for structured Web list/tables
 - Entities detected from structured Web lists/tables
 - The LIEGE framework (KDD'12)
- Entity linking for Tweets
 - Entities detected from short and noisy Tweets
 - The KAURI framework (KDD'13)
- Conclusion

Entity Linking for Tweets —Motivation

- Twitter: important information source
- Beneficial for exploiting and understanding this huge corpus of valuable text data on the Web, and also helps populate and enrich the existing knowledge bases.

....

Tweets

Bulls should still aim for a title, even through the horrible news.

² McNealy finished, he was pretty much squarely in Sun's camp. @jniccolai

Scott explains what open means...

Tyson Chandler says **Tony Allen** is the best on-ball defender in the **#NBA** http://t.co/YGmByJMC Sun: the star at the center of the Solar System Sun Microsystems: a multinational computer company Sun-Hwa Kwon: a fictional character named "Sun-Hwa Kwon"

 t_1

Entity Linking for Tweets —Problem Definition

Tweet entity linking

link the textual named entity mentions detected from tweets
 with their mapping entities existing in a knowledge base



Figure: An illustration of the tweet entity linking task. Named entity mentions detected in tweets are in bold; candidate mapping entities for each entity mention are generated by a dictionary-based method and ranked by their prior probabilities in decreasing order; true mapping entities are underlined.

Entity Linking for Tweets —Challenge

- Challenge
 - noisy, short, and informal nature of tweets
- Previous entity linking methods (EACL'06, EMNLP'07, KDD'09, SIGIR'11, EMNLP'11, and WWW'12)
 - focus on linking entities in Web documents
 - Context Similarity
 - Topical Coherence



Entity Linking for Tweets —Our Solution: The KAURI Framework

- We can increase the linking accuracy, if we
 - combine intra-tweet local information
 - with inter-tweet user interest information



Entity Linking for Tweets —Our Solution: Intra-tweet Local Information



• We utilize the **max-margin** technique to automatically learn the weight vector which gives proper weights for those three intra-tweet local features.

W. Shen, J. Wang, P. Luo, and M. Wang. Liege: Link entities in web lists with knowledge base. In SIGKDD'12.

2015-11-12

Entity Linking for Tweets —Our Solution: User interest propagation alg'



• Initialization: $\overrightarrow{\mathbf{s}} = \overrightarrow{\mathbf{p}}$

 Then apply this formula iteratively until s stabilizes within some threshold

Entity Linking for Tweets —Our Solution: Experimental Study

Method	Linkable		Unli	nkable	All	
	#	Accu.	#	Accu.	#	Accu.
LINDEN	1852	0.827	353	0.808	2205	0.824
$LOCAL_{\beta=0,\gamma=0}$	1784	0.796	355	0.812	2139	0.799
$LOCAL_{\gamma=0}$	1795	0.801	355	0.812	2150	0.803
$LOCAL_{\beta=0}$	1862	0.831	355	0.812	2217	0.828
$LOCAL_{full}$	1863	0.832	355	0.812	2218	0.829
$KAURI_{\beta=0,\gamma=0}$	1882	0.840	356	0.815	2238	0.836
$KAURI_{\gamma=0}$	1894	0.846	357	0.817	2251	0.841
$KAURI_{\beta=0}$	1913	0.854	371	0.849	2284	0.853
$KAURI_{full}$	1923	0.858	373	0.854	2296	0.858

Table: Experimental results over the data set

$$p_{j,q}^{i} = \alpha * Pp(r_{j,q}^{i}) + \beta * Sim(r_{j,q}^{i}) + \gamma * Coh(r_{j,q}^{i})$$

LINDEN is our model proposed to address the task of linking entities in Web documents.

W. Shen, J. Wang, P. Luo, and M. Wang. Linden: linking named entities with knowledge base via semantic knowledge. In *WWW'12.* 2015-11-12

29

Outline

- Introduction to entity linking with a knowledge base
 - Motivation & definition
- Entity linking for unstructured Web documents
- Entity linking for structured Web lists/tables
- Entity linking for Tweets
- Conclusion

Conclusion

- Entity linking is an interesting and challenging task
- Entity linking is very important for knowledge base population
- Recent progress
 - Entity linking for Web documents (many existing work)
 - Popularity + semantic knowledge
 - Entity linking for Web lists (LIEGE is the first one)
 - Coherence + iterative refining
 - Entity linking for Tweets (a few papers)
 - Global user interest propagation
- Future directions
 - Efficient, large-scale entity linking
 - Entity linking with domain-specific knowledge bases
 - E.g., in the domains of computer science, biomedicine, entertainment, products, finance, tourism, etc.
 - Crowdsourcing-based entity linking

References

- L. Jiang, J. Wang, N. An, et al. GRAPE: A Graph-Based Framework for Disambiguating People Appearances in Web Search. ICDM'09
- X. Fan, J. Wang, X. Pu, L. Zhou, B. Lv. On Graph-based Name Disambiguation. ACM JDIQ, 2011
- W. Shen, J. Wang, P. Luo, and M. Wang. "Linden: linking named entities with knowledge base via semantic knowledge," *WWW* '12
- W. Shen, J. Wang, P. Luo, and M. Wang. "Liege: Link entities in web lists with knowledge base," *SIGKDD* '12
- W. Shen, J. Wang, P. Luo, and M. Wang. "A graph-based approach for ontology population with named entities," *CIKM* '12
- W. Shen, J. Wang, P. Luo, and M. Wang. "Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling," *SIGKDD*'13
- L. Jiang, P. Luo, J. Wang, et al. GRIAS: an Entity-Relation Graph based Framework for Discovering Entity Aliases. **ICDM'13**
- W. Shen, J. Han, J. Wang. A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks. *SIGMOD*'14

Thanks for your attention!