ISWC 2013
Sydney, Australia
The 12th International Semantic Web Conference
and the 1st Australasian Semantic Web Conference
21-25 October 2013, Sydney, Australia

Discovering Missing Semantic Relations Between Entities in Wikipedia

Mengling Xu[1], **Zhichun Wang**[1], Rongfang Bie[1], Juanzi Li[2], Chen Zheng[1], Wantian Ke[1], and Mingquan Zhou[1]

[1]Beijing Normal University

[2]Tsinghua University

Motivation

- Wikipedia becomes one of the largest encyclopedia on the Web.
  - More than 20 million articles (March 2013)
  - 285 versions of different languages (March 2013)
  - Rich structured information: infoboxes, category system, links

Barack Obama

44th President of the United States

Incumbent

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Political party	Democratic
Spouse(s)	Michelle Robinson (m. 1992)
Children	Malia (b. 1998) Sasha (b. 2001)
Residence	White House (official) Chicago, Illinois (private)
Alma mater	Occidental College Columbia University (B.A.) Harvard Law School (J.D.)
Profession	Community organizer Lawyer Constitutional law professor Author
Religion	Christian
Awards	Nobel Peace Prize
Signature	
Website	barackobama.com

dbpedia-owl:activeYearsEndDate	2004-11-04 (xsd:date) 2008-11-16 (xsd:date)
dbpedia-owl:activeYearsStartDate	1997-01-08 (xsd:date) 2005-01-03 (xsd:date) 2009-01-20 (xsd:date)
dbpedia-owl:almaMater	dbpedia:Columbia_University dbpedia:Harvard_Law_School dbpedia:Occidental_College
dbpedia-owl:birthDate	1961-08-04 (xsd:date)
dbpedia-owl:birthPlace	dbpedia:Honolulu dbpedia:Hawaii
dbpedia-owl:child	dbpedia:Family_of_Barack_Obama
dbpedia-owl:office	President of the United States Member of the Illinois Senate from the 13th District
dbpedia-owl:orderInOffice	44th
dbpedia-owl:party	dbpedia:Democratic_Party_(United_States)
dbpedia-owl:profession	dbpedia:Author dbpedia:Constitutional_law dbpedia:Lawyer dbpedia:Community_organizing
dbpedia-owl:region	dbpedia:Illinois
dbpedia-owl:religion	dbpedia:Christian
dbpedia-owl:residence	dbpedia:White_House
dbpedia-owl:seniority	United States Senate
dbpedia-owl:spouse	dbpedia:Michelle_Obama
dbpedia-owl:successor	dbpedia:Kwame_Raoul dbpedia:Roland_Burris
dbpedia-owl:termPeriod	dbpedia:Barack_Obama__1 dbpedia:Barack_Obama__2
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/President
dbpedia-owl:vicePresident	dbpedia:Joe_Biden

Barack Obama

44th President of the United States

Incumbent

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Political party	Democratic
Spouse(s)	Michelle Robinson (m. 1992)
Children	Malia (b. 1998) Sasha (b. 2001)
Residence	White House (official) Chicago, Illinois (private)
Alma mater	Occidental College Columbia University (B.A.) Harvard Law School (J.D.)
Profession	Community organizer Lawyer Constitutional law professor Author
Religion	Christian
Awards	Nobel Peace Prize
Signature	
Website	barackobama.com

dbpedia-owl:activeYearsEndDate	2004-11-04 (xsd:date) 2008-11-16 (xsd:date)
dbpedia-owl:activeYearsStartDate	1997-01-08 (xsd:date) 2005-01-03 (xsd:date) 2009-01-20 (xsd:date)
dbpedia-owl:almaMater	dbpedia:Columbia_University dbpedia:Harvard_Law_School dbpedia:Occidental_College
dbpedia-owl:birthDate	1961-08-04 (xsd:date)
dbpedia-owl:birthPlace	dbpedia:Honolulu dbpedia:Hawaii
dbpedia-owl:child	dbpedia:Family_of_Barack_Obama
dbpedia-owl:office	President of the United States Member of the Illinois Senate from the 13th District
dbpedia-owl:orderInOffice	44th
dbpedia-owl:party	dbpedia:Democratic_Party_(United_States)
dbpedia-owl:profession	dbpedia:Author dbpedia:Constitutional_law dbpedia:Lawyer dbpedia:Community_organizing
dbpedia-owl:region	dbpedia:Illinois
dbpedia-owl:religion	dbpedia:Christian
dbpedia-owl:residence	dbpedia:White_House
dbpedia-owl:seniority	United States Senate
dbpedia-owl:spouse	dbpedia:Michelle_Obama
dbpedia-owl:successor	dbpedia:Kwame_Raoul dbpedia:Roland_Burris
dbpedia-owl:termPeriod	dbpedia:Barack_Obama__1 dbpedia:Barack_Obama__2
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/President
dbpedia-owl:vicePresident	dbpedia:Joe_Biden

dbpedia-owl:almaMater	dbpedia:The_Queen's_College,_Oxford
dbpedia-owl:birthDate	1955-06-08 (xsd:date)
dbpedia-owl:birthName	Timothy John Berners-Lee
dbpedia-owl:birthYear	1955-01-01 00:00:00 (xsd:date)
dbpedia-owl:occupation	dbpedia:Tim_Berners-Lee__1 dbpedia:Computer_scientist
dbpedia-owl:personFunction	dbpedia:Tim_Berners-Lee__2
dbpedia-owl:religion	dbpedia:Unitarian_Universalism
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/8/83/T
dbpedia-owl:title	Professor
dbpedia-owl:wikiPageExternalLink	http://www.w3.org/History/19921103-hypertext/hypertext/WWW/ http://computemagazine.com/man-who-invented-world-wide-w http://www.w3.org/People/Berners-Lee/ http://www.technologyreview.com/Infotech/13784/ http://info.cern.ch/hypertext/WWW/TheProject.html, http://identi.ca/timbl http://news.bbc.co.uk/2/hi/technology/4132752.stm http://videolectures.net/iswc06_lee_itbl/
dbpedia-owl:wikiPageID	30034 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	545216408 (xsd:integer)

Sir Tim Berners-Lee

Berners-Lee in 2008

Born	Timothy John Berners-Lee 8 June 1955 (age 58) London, England United Kingdom
Residence	United States and United Kingdom[1]
Nationality	British
Alma mater	Queen's College, Oxford
Occupation	Computer scientist
Employer	World Wide Web Consortium University of Southampton Massachusetts Institute of Technology
Known for	Inventing the World Wide Web Holder of the 3Com Founders Chair at MIT's Computer Science and Artificial Intelligence Laboratory
Title	Professor
Religion	Unitarian (raised as Anglican)[2]
Parents	Conway Berners-Lee Mary Lee Woods

Website

www.w3.org/People/Berners-Lee/

Barack Obama

44th President of the United States

Incumbent

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Political party	Democratic
Spouse(s)	Michelle Robinson (m. 1992)
Children	Malia (b. 1998) Sasha (b. 2001)
Residence	White House (official) Chicago, Illinois (private)
Alma m	
Professi	
	Author
Religion	Christian
Awards	Nobel Peace Prize
Signature	
Website	barackobama.com

dbpedia-owl:activeYearsEndDate	2004-11-04 (xsd:date) 2008-11-16 (xsd:date)
dbpedia-owl:activeYearsStartDate	1997-01-08 (xsd:date) 2005-01-03 (xsd:date) 2009-01-20 (xsd:date)
dbpedia-owl:almaMater	dbpedia:Columbia_University dbpedia:Harvard_Law_School dbpedia:Occidental_College
dbpedia-owl:birthDate	1961-08-04 (xsd:date)
dbpedia-owl:birthPlace	dbpedia:Honolulu dbpedia:Hawaii
dbpedia-owl:child	dbpedia:Family_of_Barack_Obama
dbpedia-owl:office	President of the United States Member of the Illinois Senate from the 13th District
dbpedia-owl:orderInOffice	44th
dbpedia-owl:party	dbpedia:Democratic_Party_(United_States)
dbpedia-owl:profession	dbpedia:Author dbpedia:Constitutional_law dbpedia:Lawyer dbpedia:Community_organizing
dbpedia-owl:region	dbpedia:Illinois
dbpedia-owl:religion	dbpedia:Christian
dbpedia-owl:residence	dbpedia:White_House
dbpedia-owl:seniority	United States Senate
dbpedia-owl:spouse	dbpedia:Michelle_Obama
dbpedia-owl:successor	dbpedia:Kwame_Raoul dbpedia:Roland_Burris
dbpedia-owl:termPeriod	dbpedia:Barack_Obama__1 dbpedia:Barack_Obama__2
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/President
dbpedia-owl:vicePresident	dbpedia:Joe_Biden

dbpedia-owl:almaMater	dbpedia:The_Queen's_College,_Oxford
dbpedia-owl:birthDate	1955-06-08 (xsd:date)
dbpedia-owl:birthName	Timothy John Berners-Lee
dbpedia-owl:birthYear	1955-01-01 00:00:00 (xsd:date)
dbpedia-owl:occupation	dbpedia:Tim_Berners-Lee__1
	http://www.w3.org/People/Berners-Lee/ http://www.technologyreview.com/Infotech/13784/ http://info.cern.ch/hypertext/WWW/TheProject.html, http://identi.ca/timbl http://news.bbc.co.uk/2/hi/technology/4132752.stm http://videolectures.net/iswc06_lee_itbl/
dbpedia-owl:wikiPageID	30034 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	545216408 (xsd:integer)

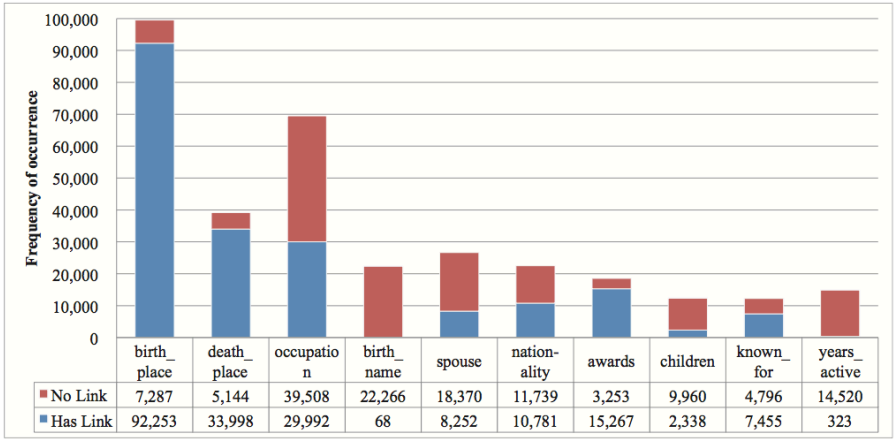
Sir Tim Berners-Lee

Berners-Lee in 2008

Born	Timothy John Berners-Lee 8 June 1955 (age 58) London, England United Kingdom
Residence	United States and United Kingdom[1]
Nationality	British
Alma mater	Queen's College, Oxford
Occupation	Computer scientist
Employer	World Wide Web Consortium University of Southampton Massachusetts Institute of Technology
Known for	Inventing the World Wide
ders	
Religion	Unitarian (raised as Anglican)[2]
Parents	Conway Berners-Lee Mary Lee Woods
Website	www.w3.org/People/Berners-Lee/

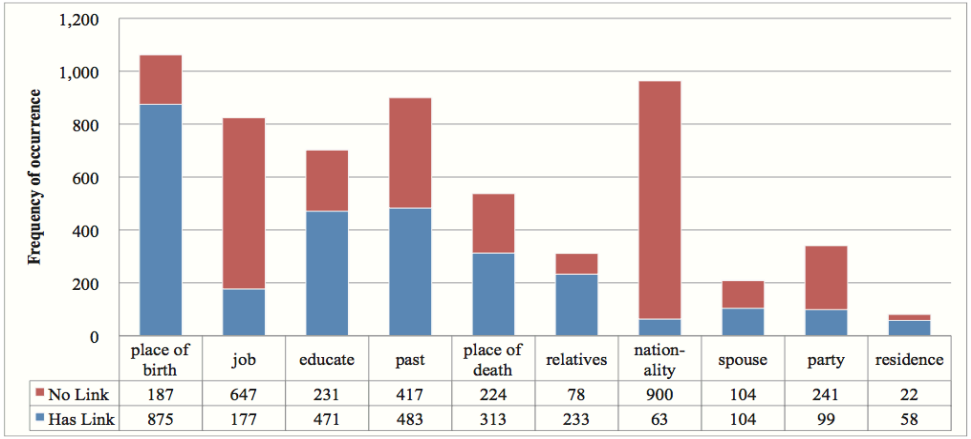
Entity links in infoboxes are important for defining semantic relations between entities. However, there are a lot of entity links are missing in infoboxes.

Motivation



	birth_ place	death_ place	occupation	birth_ name	spouse	nation-ality	awards	children	known_ for	years_ active
No Link	7,287	5,144	39,508	22,266	18,370	11,739	3,253	9,960	4,796	14,520
Has Link	92,253	33,998	29,992	68	8,252	10,781	15,267	2,338	7,455	323

Links in person infoboxes in English Wikipedia



	place of birth	job	educate	past	place of death	relatives	nation-ality	spouse	party	residence
No Link	187	647	231	417	224	78	900	104	241	22
Has Link	875	177	471	483	313	233	63	104	99	58

Links in person infoboxes in Chinese Wikipedia

How to automatically add entity links in the attribute values of infoboxes to enrich the relations between entities?

Related Work

- Entity Linking
  - Identifying entities in documents and linking them to a knowledge base, such as Wikipedia and DBpedia.
    - Wikify! [Mihalcea 2007]
    - M&W [Milne 2008a]
    - DBpedia Spotlight [Mendes 2011]
    - …
- Instance Matching
  - Adding "*sameAs*" links among RDF datasets by finding equivalent entities.
    - Silk [Volz 2009]
    - idMesh [Cudre-Mauroux 2009]
    - KnoFuss [Nikolov 2008]
    - …

Proposed Approach

Step 1

Mention Identification

Identify candidate name mentions in infoboxes that might refer to entities in Wikipedia

Step 2

Entity Link Prediction

Identify the correct corresponding entity for each mention

Mention Identification

- (1) Extract Mention-Entity vocabulary

  Mention: the anchor text of a hyperlink

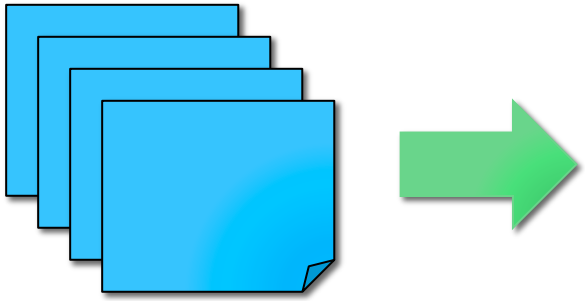
Beijing is the second largest Chinese city by *urban population* after *Shanghai* and is the nation's *political*, *cultural*, and *educational center*.

Beijing is the second largest Chinese city by [[List of cities in the People's Republic of China by urban population|urban population]] after [[Shanghai]] and is the nation's [[Politics of the People's Republic of China|political]], [[Chinese culture|cultural]], and [[List of universities and colleges in Beijing|educational center]].

[[Entity | Mention]] [[Entity or Mention]]

Mention Identification

- (1) Extract Mention-Entity vocabulary



Wikipedia articles

Mention	Entities
m_1	$e_{11}, e_{12}, \ldots, e_{1p}$
m_2	$e_{21}, e_{22}, \ldots, e_{2q}$
m_3	$e_{31}, e_{32}, \ldots, e_{3h}$
…	…
m_n	$e_{n1}, e_{n2}, \ldots, e_{nk}$

Mention Identification

- (2) Identify mentions in the attribute values of infoboxes

Mention	Entities
m_1	$e_{11}, e_{12}, \ldots, e_{1p}$
m_2	$e_{21}, e_{22}, \ldots, e_{2q}$
m_3	$e_{31}, e_{32}, \ldots, e_{3h}$
…	…
m_n	$e_{n1}, e_{n2}, \ldots, e_{nk}$

n-gram matching

Infobox

Attribute	Value
a_1	v_1
a_2	v_2
…	…

$\{<m_{i1}, E_{i1}>, \ldots\}$

$\{<m_{i2}, E_{i2}>, \ldots\}$

Mentions and their candidate entities

Link Prediction

- Given a mention *m* and a set of its candidate entities E_m
  - 7 features are computed for each mention-entity pair
  - An aggregated score is computed for each mention-entity pair

$$s(m,e) = \omega_1 \times f_1(m,e) + \ldots + \omega_6 \times f_6(m,e) + \omega_7 \times f_7(m,e)$$

  - Entity e^* that maximizes $s(m,e^*)$ is predicted as the destination entity of *m*.

- Two important problems
  - What kind of features are useful for link prediction?
  - How to decide the weights of different features?

Features for Link Prediction

- Feature 1: Entity Occurrence

$$f_1(e,m) = \begin{cases} 1 \text{ if } e \in C_{article}(m) \\ 0 \text{ otherwise} \end{cases}$$

$C_{article}(m)$ is the set of entities appearing in the main text of the current article.

- Feature 2: Link Probability

$$f_2(e,m) = \frac{count(m,e)}{count(m)}$$

count(m,e) is the number of times that *m* links to *e*;
count(m) is the number of times *that m* appears.

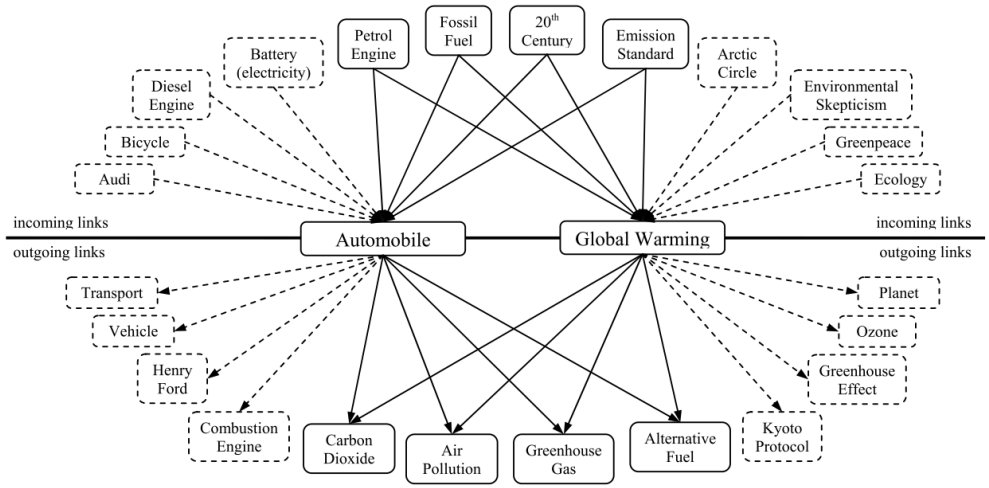
Features for Link Prediction

- Semantic Relatedness
  - Relatedness between two entities [Milne 2008b]

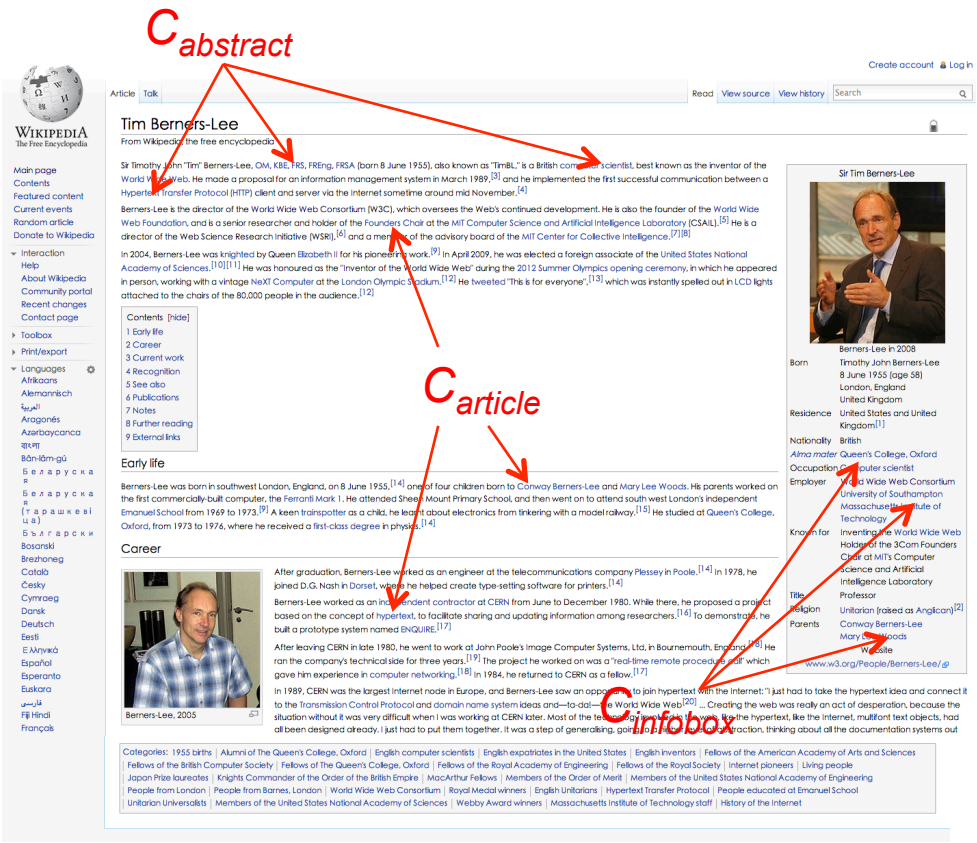
$$r(a,b) = 1 - \frac{log(max(|I_a|,|I_b|)) - log(|I_a \cap I_b|)}{log(|W|) - log(min(|I_a|,|I_b|))}$$

  - Relatedness between one entity and a set of entities

$$SR(a,B) = \frac{1}{|B|}\sum_{b \in B} r(a,b)$$



Features for Link Prediction



Feature 3: Infobox Context Relatedness

$$f_3(e,m) = SR(e, C_{infobox}(m))$$

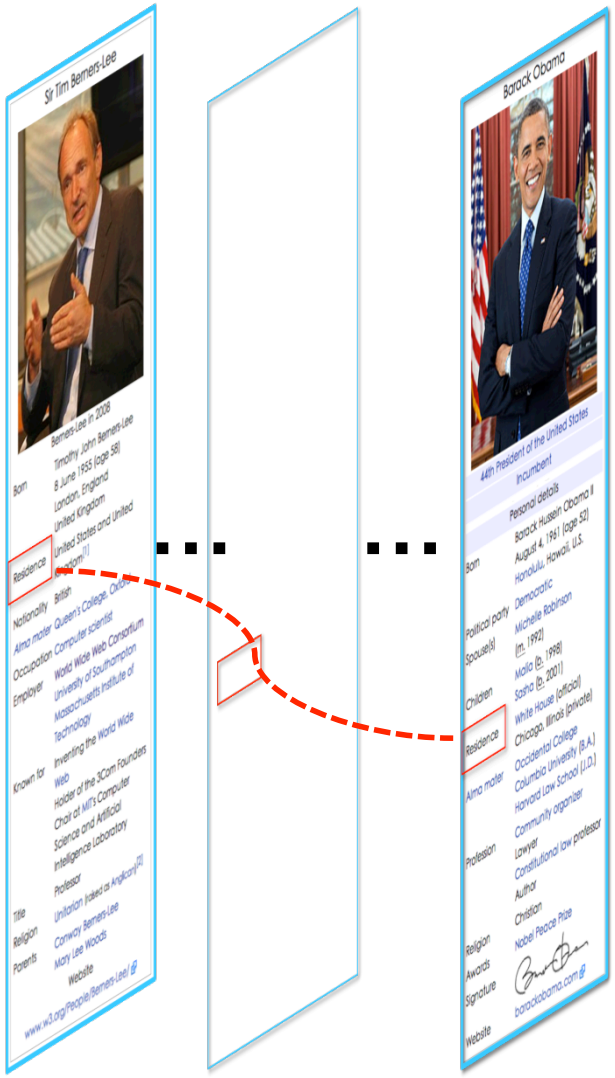
Feature 4: Article Context Relatedness

$$f_4(e,m) = SR(e, C_{article}(m))$$

Feature 5: Abstract Context Relatedness

$$f_5(e,m) = SR(e, C_{abstract}(m))$$

Features for Link Prediction



C_{att_rang}: set of entities that appear in the value of the concerned attribute

C_{att_dom} : set of entities that described by the concerned attribute

Feature 6: Attribute Range Context Relatedness

$$f_6(e, m) = SR(e, C_{att\_rang}(m))$$

Feature 7: Attribute Domain Context Relatedness

$$f_7(e, m) = SR(e, C_{att\_dom}(m))$$

Learn the weights of features

- Given a mention *m* and the set of its candidate entities E_m

$$s(m,e) = \omega_1 \times f_1(m,e) + \ldots + \omega_6 \times f_6(m,e) + \omega_7 \times f_7(m,e)$$

- Entity e^* in E_m that maximized $s(m,e^*)$ is predicted as the destination entity of *m*.

$$\boldsymbol{\omega} \cdot (\boldsymbol{f}(m,e^*) - \boldsymbol{f}(m,e)) > 0, (e \in E_m, e \neq e^*)$$

$$P(e_1 \succ e_2) = \frac{1}{1 + e^{-\boldsymbol{\omega} \cdot (\boldsymbol{f}(m,e_1) - \boldsymbol{f}(m,e_2))}}$$

$$\begin{cases} s(m,e_1) > s(m,e_2),\ P(e_1 \succ e_2) > 0.5 \\ s(m,e_1) < s(m,e_2),\ P(e_1 \succ e_2) < 0.5 \end{cases}$$

Learn the weights of features

Input: a set of known entity links $L = \{< m_i, e_i >\}_{i=1}^{k}$

Output: A dataset $D = \{(\boldsymbol{x_j}, y_j)\}_{j=1}^{m}$ is generated

For each mention m_i

For each entity $e' \in (E_{m_i} - \{e_i\})$

Generate a positive example $(\boldsymbol{f}(m_i, e_i) - \boldsymbol{f}(m_i, e'), positive)$

or generate a negative example $(\boldsymbol{f}(m_i, e') - \boldsymbol{f}(m_i, e_i), negative)$

End For

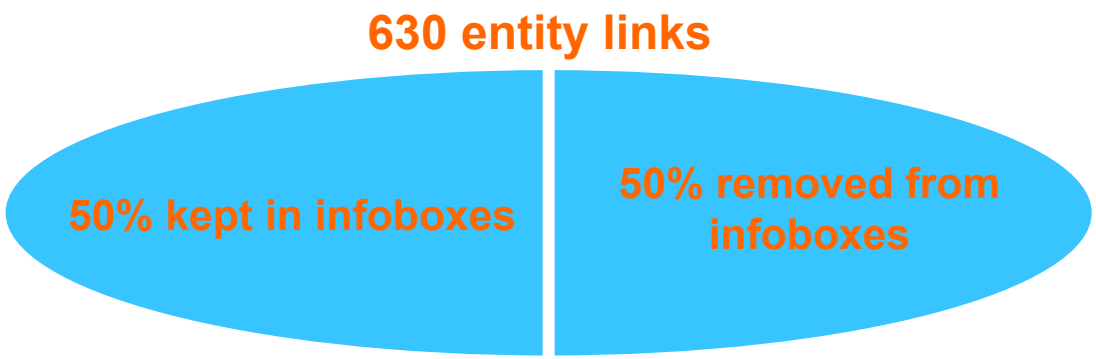
End For

Train a Logistic Regression model using dataset $D = \{(\boldsymbol{x_j}, y_j)\}_{j=1}^{m}$

Optimal weights $\boldsymbol{\omega} =< \omega_1, ..., \omega_7 >$

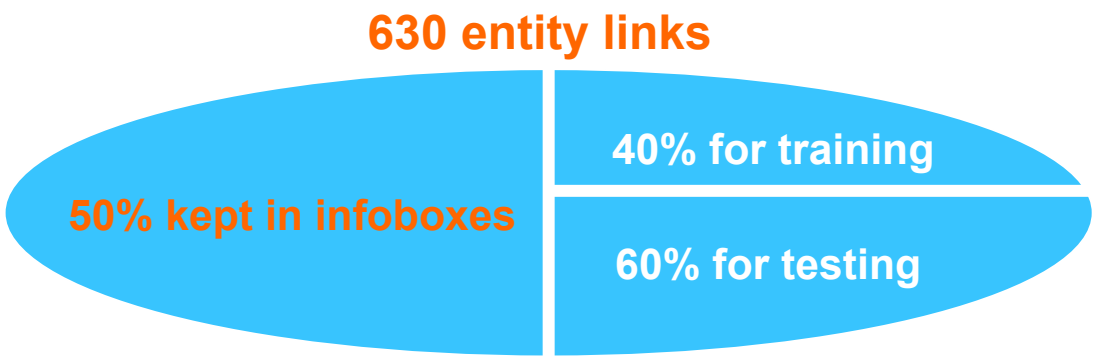
Experiments

- Dataset
  - English Wikipedia XML dump (August 2012)
  - 100 randomly chosen infoboxes



Experiments

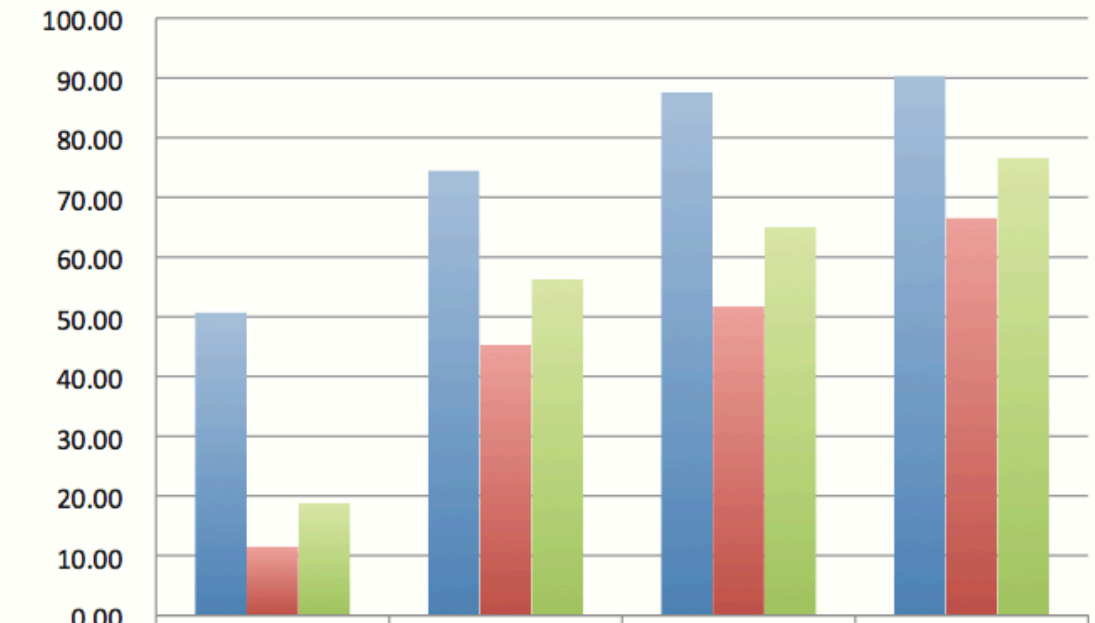
- Dataset
  - English Wikipedia XML dump (August 2012)
  - 100 randomly chosen infoboxes



- Evaluation Metrics
  - Precision, Recall, F1-measure
- Comparison Methods
  - Wikify! [Mihalcea 2007] ; M&W [Milne 2008a] ;
  - SVM

Experiments

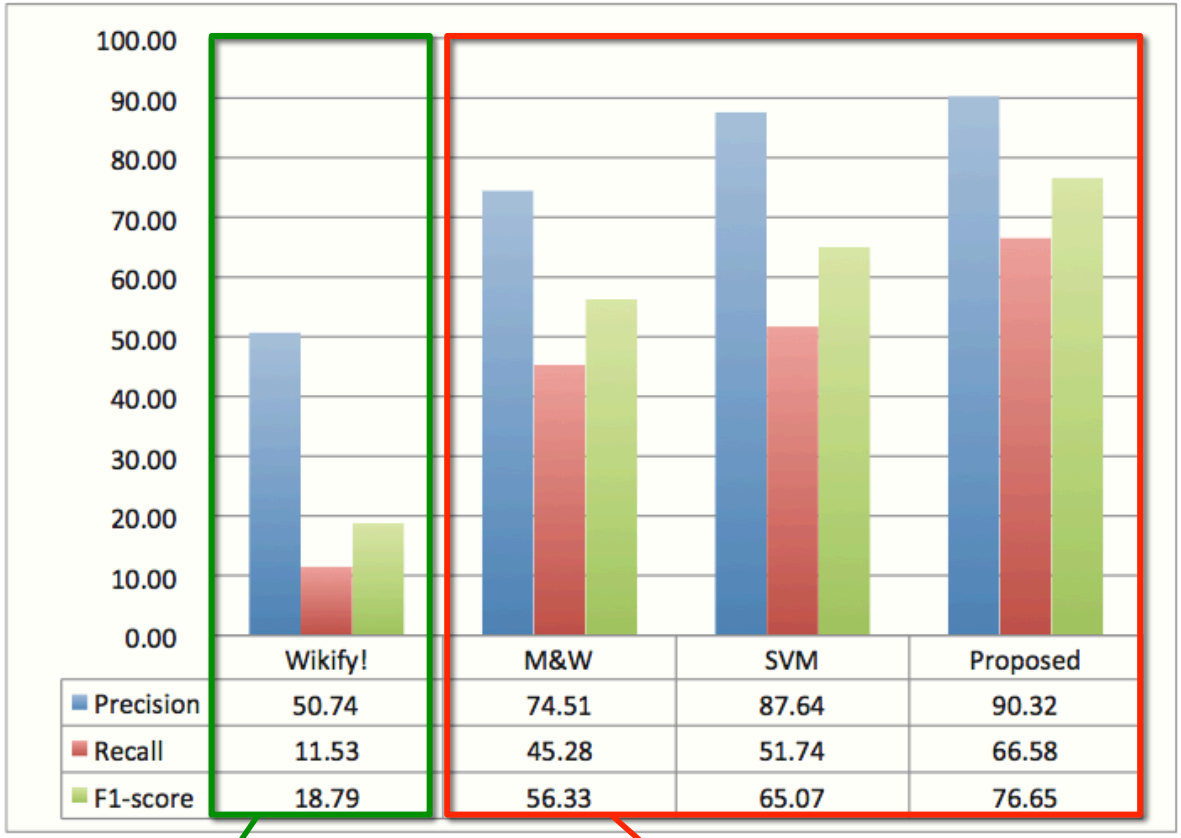
- Results



	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

Experiments

- Results



	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

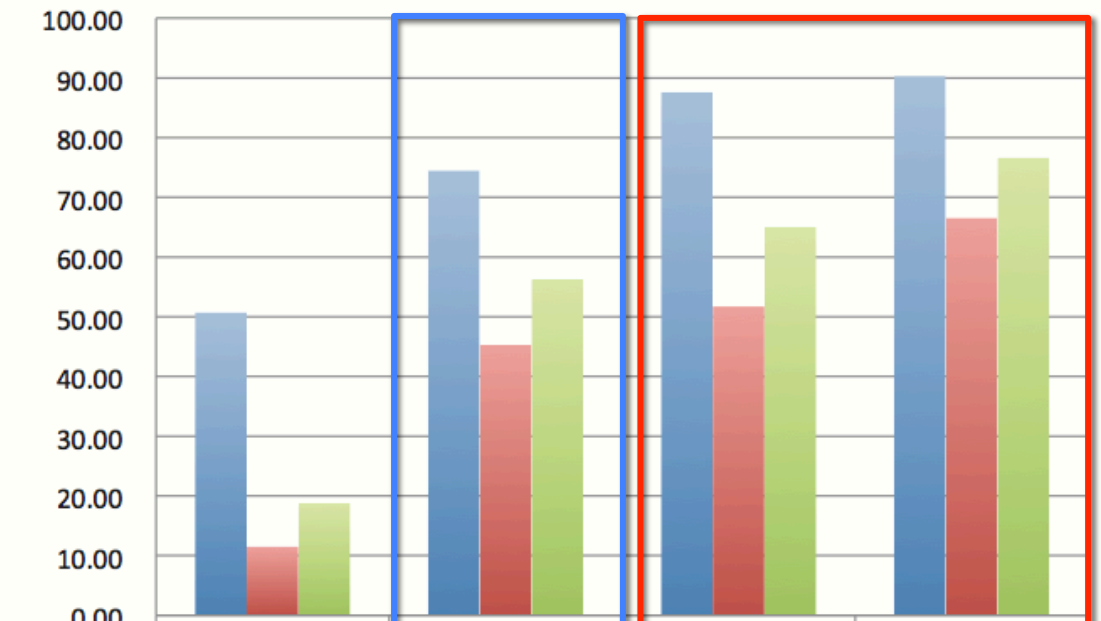
Trained with existing entity links

No training process

Trained with training data

Experiments

- Results



	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

3 features

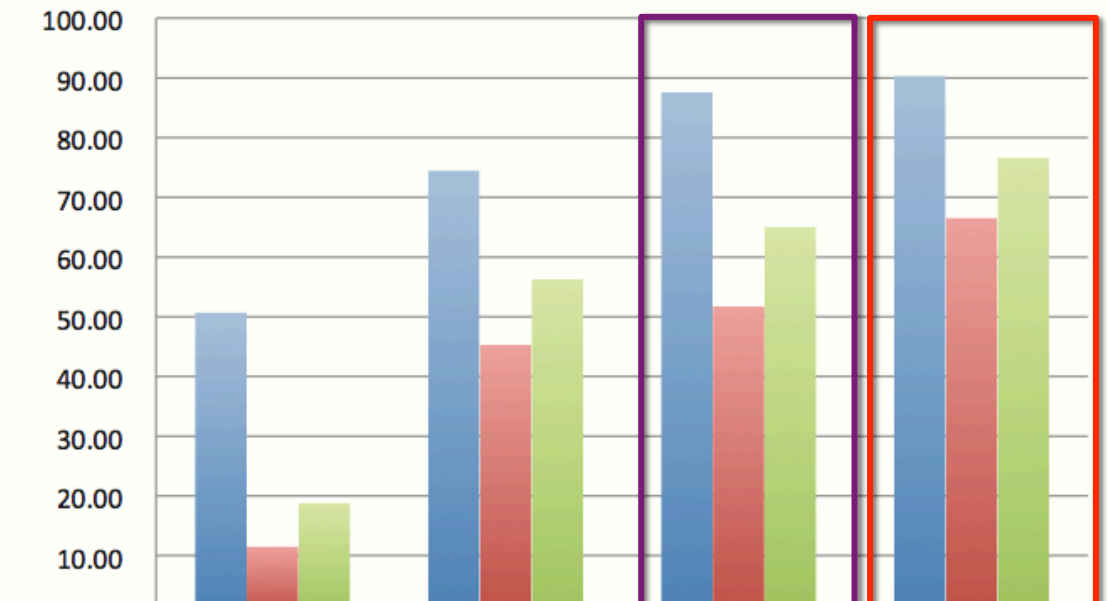
7 features

Trained with existing entity links

Specially designed features

Experiments

- Results



	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

Trained with existing entity links

Specially designed features

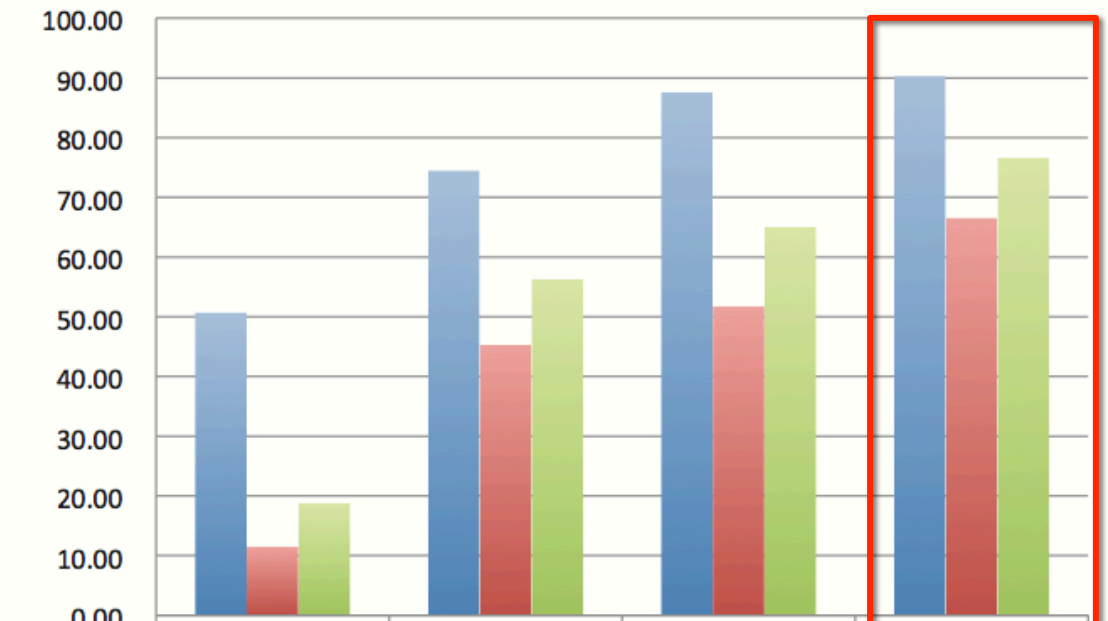
More effective learning algorithm

Classification-based method

Weighting and Ranking

Experiments

- Results



	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

Trained with existing entity links

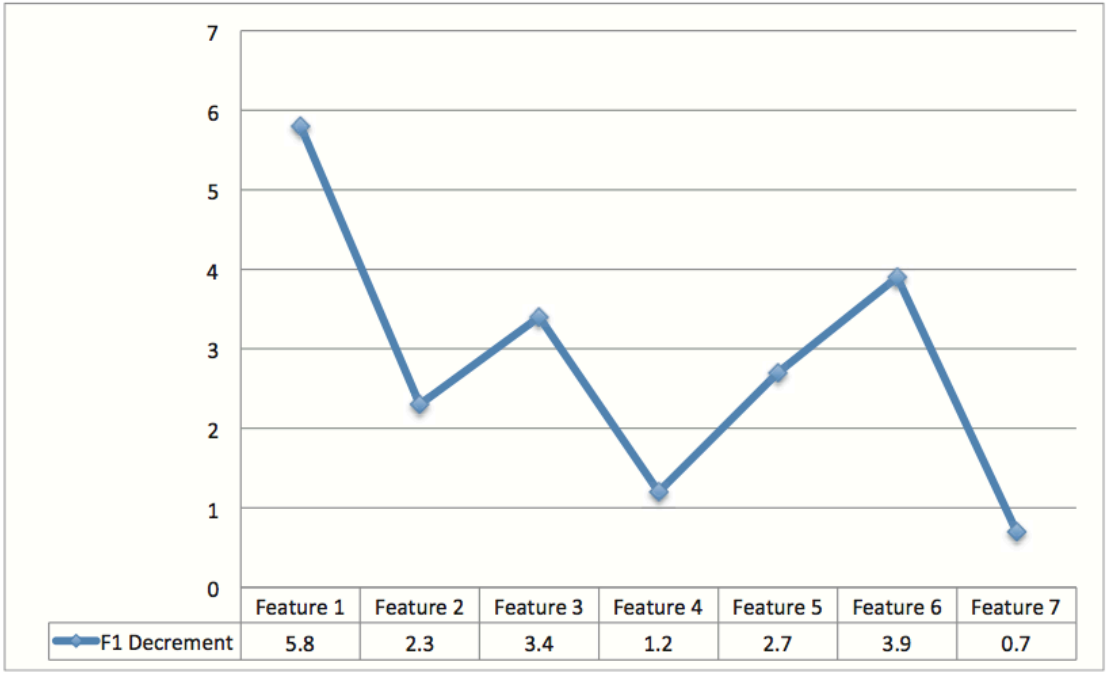
Specially designed features

More effective learning algorithm

Important factors for better results.

Experiments

- Feature Contribution Analysis
  - Our approach runs 7 times on the evaluation data, each time one feature is removed from the feature vectors of mention-entity pairs.



	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7
F1 Decrement	5.8	2.3	3.4	1.2	2.7	3.9	0.7

Decrease of F1-score for each feature

Entity Occurrence

Attribute Range

Infobox Context

Abstract Context

Link Probability

Article Context

Attribute Domain

More important

Less important

Conclusion and Future Work

- Conclusion
  - An approach for automatically discovering the missing typed relations between entities in Wikipedia's infoboxes is proposed.
  - 7 features for predicting the entity links in infoboxes are defined; the effectiveness and importance are analyzed.
  - A new method for learning the weights of features is proposed, which achieves better results than baseline methods.

- Future Work
  - Discovering incorrect entity links in infoboxes.
  - Discovering new RDF links (not just "sameAs" links) between RDF datasets.

References

- [Mendes 2011] P.N.Mendes,M.Jakob,A.Garc´ıa Silva,andC.Bizer. DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, I-Semantics ’11, pages 1–8, New York, NY, USA, 2011. ACM.
- [Milne 2008a] D. Milne and I. H. Witten. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM ’08, pages 509–518, New York, NY, USA, 2008. ACM.
- [Mihalcea 2007] R.Mihalcea, A.Csomai.Wikify!:linking documents to encyclopedic knowledge. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM ’07, pages 233–242, New York, NY, USA, 2007. ACM.
- [Cudre-Mauroux 2009] P. Cudre-Mauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. idmesh: graph-based disambiguation of linked data. In Proceedings of the 18th international conference on World wide web, WWW ’09, pages 591–600, New York, NY, USA, 2009. ACM.
- [Nikolov 2008] A. Nikolov, V. S. Uren, E. Motta, and A. N. D. Roeck. Handling instance coref-erencing in the knofuss architecture. In P. Bouquet, H. Halpin, H. Stoermer, and G. Tummarello, editors, IRSW, volume 422 of CEUR Workshop Proceedings. CEUR-WS.org, 2008.
- [Volz 2009] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In ISWC ’09: Proceedings of the 8th International Semantic Web Conference, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.
- [Milne 2008b] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.

Thank you！

Email: zcwang@bnu.edu.cn

Discovering Missing Semantic Relations Between Entities in Wikipedia

Mengling Xu[1], Zhichun Wang[1], Rongfang Bie[1], Juanzi Li[2], Chen Zheng[1], Wantian Ke[1], and Mingquan Zhou[1]

[1]Beijing Normal University
[2]Tsinghua University