JVN-TDT Entity Linking Systems at TAC-KBP2012

Hien T. Nguyen[a] Huy H. Minh[a,b] Tru H. Cao[b,c] Trong T. Nguyen[b,c]

[a]Ton Duc Thang University
[b]John von Neumann Institute
[c]Ho Chi Minh City University of Technlogy

Outline

- JVN_TDT1 System
  - Features
  - Coreference-based Entity Linking
  - Experiments
- JVN_TDT2 System
  - Heuristics
  - VSM for Entity Linking
  - Experiments
- Conclusion

JVN_TDT1 System

- Improving the system of (Milne and Witten, 2008)
  - Two features: Prior probability and Semantic relatedness
  - Training a classifier using Bagged C4.5 with two these features on 500 articles randomly chosen from Wiki
  - Tuning parameters using other 100 articles randomly chosen from Wiki
- Exploiting coreference relations among mentions

Prior probability Medelyan et al, 2008

$$P(e \mid m) = \frac{count_m(e)}{\sum_{e_i \in CE_m} count_m(e_i)}$$

- For instance: assuming that in Wiki, a mention *m* occurs 10 times and refers to three different entities *a*, *b*, *c*, in which 7 times *m* refers to *a*, 2 times *m* refers to *b* respectively; then $P(a|m) = 7/10 = 0.7$, $P(b|m) = 2/10 = 0.2$, $P(c|m) = 1/10 = 0.1$; therefore, *a* is considered as more popular than *b* and *c*.

Semantic relatedness Milne and Witten, 2008

- Semantic relatedness between two entities
- Semantic relatedness between a candidate of a mention and contextual entities
  - a contextual entity is an entity that has identified

Semantic relatedness Milne and Witten, 2008

- A_1 be the set of all Wiki articles that link to e_1
- A_2 be the set of all Wiki articles that link to e_2
- W is the set of all articles in Wikipedia

$$Sem(e_1, e_2) = 1 - \frac{\log(\max(|A_1|, |A_2|)) - \log(|A_1 \cap A_2)|}{\log(|W|) - \log(\min(|A_1|, |A_2|))}$$

Semantic relatedness Milne and Witten, 2008

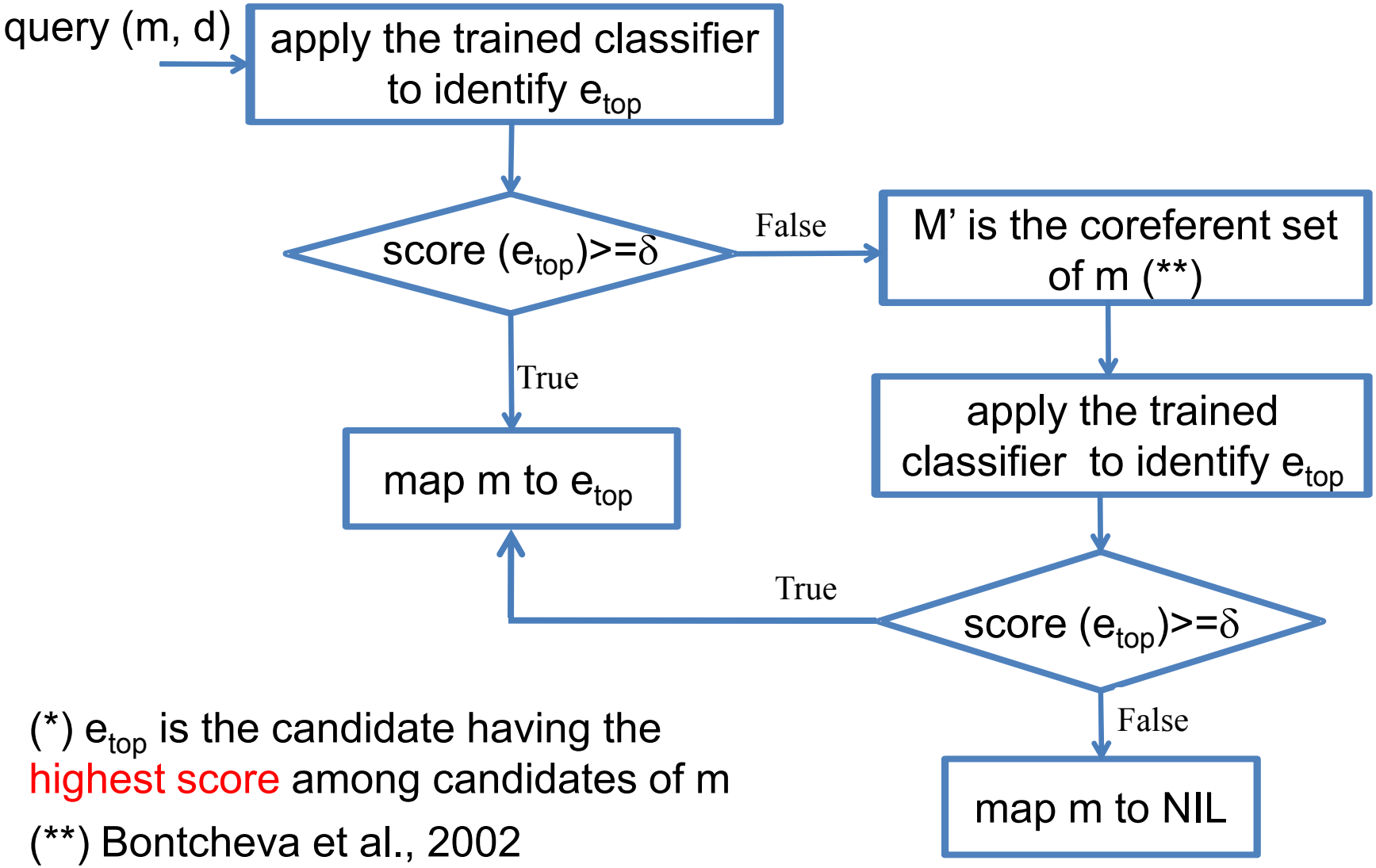
- Let E be the set of contextual entities
- Let m be a query mention and e be an its candidate

$$SR(e) = \frac{\sum_{e' \in E} Sem(e, e')}{|E|}$$

JVN_TDT1 System

- Improving the system of (Milne and Witten, 2008)
  - Two features: Prior probability and Semantic relatedness
  - Training a classifier using Bagged C4.5 with two these features on 500 articles randomly chosen from Wiki
  - Tuning parameters using other 100 articles randomly chosen from Wiki
- Exploiting coreference relations among mentions

Linking Algorithm



(*) e_{top} is the candidate having the highest score among candidates of m

(**) Bontcheva et al., 2002

Experiments

Feature	All	NIL	Non-NIL
P	75.3%	87.8%	62.8%
P+SR	**82.5%**	**95.0%**	69.9%

TABLE 1 - The MAA overall results TAC-KBP2011 dataset using coreference

Feature	All	NIL	Non-NIL
P	72.7%	85.0%	61.5%
P+SR	**79.5%**	**91.3%**	68.4%

TABLE 2 - The B-Cubed+ F1 overall results TAC-KBP2011 dataset using coreference

Experiments

Feature	All	NIL	Non-NIL
P	68.3%	90.6%	46.0%
P+SR	72.7%	**96.6%**	48.7%

TABLE 3 - The MAA overall results TAC-KBP2011 dataset not using coreference

Feature	All	NIL	Non-NIL
P	65.5%	87.6%	44.9%
P+SR	69.6%	93.0%	47.3%

TABLE 4 - The B-Cubed+ F1 overall results TAC-KBP2011 dataset not using coreference

Experiments

Query	MAA	B-Cubed+ F1
All (2,226)	67.7%	58.6%
NIL (1,049)	84.9%	71.0%
Non-NIL (1,177)	52.4%	49.8%

TABLE 6 - The MAA and B-Cubed+ F1 overall results TAC-KBP2012 dataset using coreference

JVN_TDT2 System

- Heuristics
- VSM for entity linking
- Experiments

Heuristics

- Heuristic 1:

  Among candidate entities of mention m, the ones whose title-hints occur around m in a context window are chosen.

- Title-hints of an entity are extracted from its title and redirecting titles

Title-hint instances

Atlanta (GA)

From Wikipedia, the free encyclopedia

Redirect page

↳ Atlanta

Atlanta, Georgia

From Wikipedia, the free encyclopedia

Redirect page

↳ Atlanta

title-hints

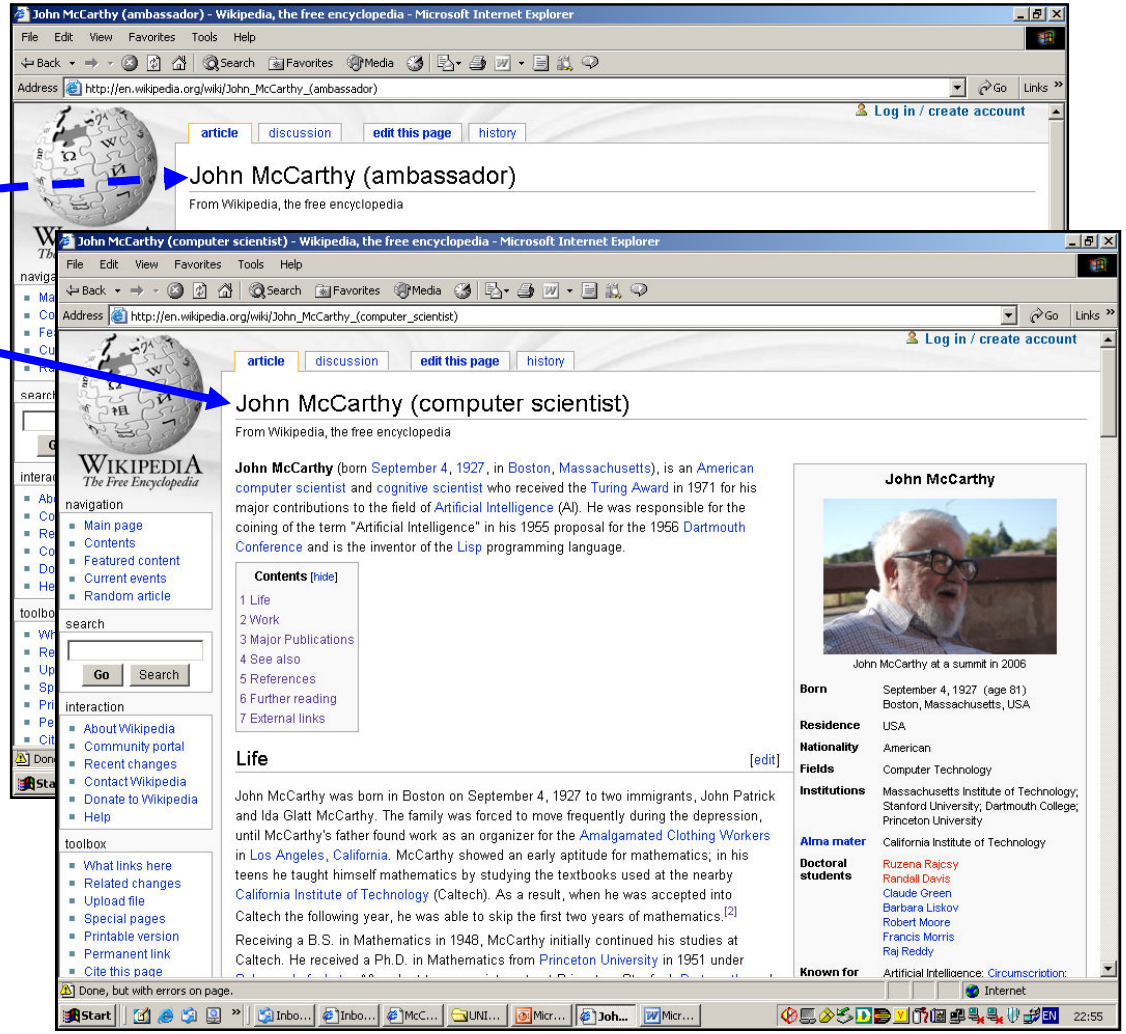
Example 1

title-hint

A state of emergency has been declared in the **US state** of Georgia after two people died in storms, a day after a tornado hit the city of Atlanta.

Example 2

In 1955 the **computer scientist** John McCarthy, who has died aged 84, coined the term artificial intelligence, or AI.



Heuristics

- Heuristic 2

  if *m* is a title-hint of an already identified entity around it, the chosen candidates are the ones that have outlinks to the identified entity or this identified entity has outlinks to these candidates.

Example 3

ATLANTA — The political movement that spread nationally in opposition to corporate bailouts and President Barack Obama's health care overhaul cannot seem to find a unified voice on Georgia's proposed constitutional amendment on charter schools.

Atlanta

From Wikipedia, the free encyclopedia

This article is about the city in the U.S. state (disambiguation).

Atlanta (/ətˈlæntə/, stressed /ætˈlæntə/, locally /ætˈlænə/) is the capital of and the most populous city in the U.S. state of Georgia, with a 2010 population of 420,003.[9] Atlanta is the cultural and economic center of

Georgia (U.S. state)

From Wikipedia, the free encyclopedia

"State of Georgia" redirects here. For TV series, see State of Georgia (TV (country). For other uses, see Georgia (disambiguation).

Georgia (ˈ/ˈdʒɔrdʒə/ *JOR-juh*) is a state located in the southeastern United States. It was established in 1732, the last of the original Thirteen Colonies.[4] Named after King George II of Great Britain,[5] Georgia was the fourth state to ratify the United States Constitution, on January 2, 1788.[6] It declared its secession from the Union on January 21, 1861, and was one of the original seven Confederate states.[6] It was the last state to be restored to the Union, on July 15, 1870.[6] Georgia is the 24th most extensive and the 9th most populous of the 50 United States. From 2007 to 2008, 14 of Georgia's counties ranked among the nation's 100 fastest-growing, second only to Texas.[7] Georgia is known as the *Peach State* and the *Empire State of the South*.[6] Atlanta is the state's capital and its most populous city.

Example 3

ATLANTA — The political movement that spread nationally in opposition to corporate bailouts and President Barack Obama's health care overhaul cannot seem to find a unified voice on Georgia's proposed constitutional amendment on charter schools.

Atlanta

From Wikipedia, the free encyclopedia

This article is about the city in the U.S. state (disambiguation).

Atlanta (/ətˈlæntə/, stressed /ætˈlæntə/, locally /ætˈlænə/) is the capital of and the most populous city in the U.S. state of Georgia, with a 2010 population of 420,003.[9] Atlanta is the cultural and economic center of

Georgia (U.S. state)

From Wikipedia, the free encyclopedia

"State of Georgia" redirects here. For TV series, see State of Georgia (TV (country). For other uses, see Georgia (disambiguation).

Georgia (ˈ/ˈdʒɔrdʒə/ *JOR-juh*) is a state located in the southeastern United States. It was established in 1732, the last of the original Thirteen Colonies.[4] Named after King George II of Great Britain,[5] Georgia was the fourth state to ratify the United States Constitution, on January 2, 1788.[6] It declared its secession from the Union on January 21, 1861, and was one of the original seven Confederate states.[6] It was the last state to be restored to the Union, on July 15, 1870.[6] Georgia is the 24th most extensive and the 9th most populous of the 50 United States. From 2007 to 2008, 14 of Georgia's counties ranked among the nation's 100 fastest-growing, second only to Texas.[7] Georgia is known as the *Peach State* and the *Empire State of the South*.[6] Atlanta is the state's capital and its most populous city.

Example 3

ATLANTA — The political movement that spread nationally in opposition to corporate bailouts and President Barack Obama's health care overhaul cannot seem to find a unified voice on Georgia's proposed constitutional amendment on charter schools.

Atlanta

From Wikipedia, the free encyclopedia

This article is about the city in the U.S. state (disambiguation).

Atlanta (/ətˈlæntə/, stressed /ætˈlæntə/, locally /ætˈlænə/) is the capital of and the most populous city in the U.S. state of Georgia, with a 2010 population of 420,003.[9] Atlanta is the cultural and economic center of

Georgia (U.S. state)

From Wikipedia, the free encyclopedia

"State of Georgia" redirects here. For TV series, see State of Georgia (TV (country). For other uses, see Georgia (disambiguation).

Georgia (/ˈdʒɔrdʒə/ *JOR-juh*) is a state located in the southeastern United States. It was established in 1732, the last of the original Thirteen Colonies.[4] Named after King George II of Great Britain,[5] Georgia was the fourth state to ratify the United States Constitution, on January 2, 1788.[6] It declared its secession from the Union on January 21, 1861, and was one of the original seven Confederate states.[6] It was the last state to be restored to the Union, on July 15, 1870.[6] Georgia is the 24th most extensive and the 9th most populous of the 50 United States. From 2007 to 2008, 14 of Georgia's counties ranked among the nation's 100 fastest-growing, second only to Texas.[7] Georgia is known as the *Peach State* and the *Empire State of the South*.[6] Atlanta is the state's capital and its most populous city.

Example 4

Under Pienciak's leadership at the Daily News, the investigative team has won numerous awards for its work, most notably for its exhaustive series "9/11 Money Trough," which examined how the $21.4 billion the federal government gave New York to recover from the Sept. 11 attacks was misspent and mismanaged.

Daily News (New York)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2011)*

The ***Daily News*** of New York City is the fourth most widely circulated daily newspaper in the United States.[2]

Daily News

New York City

From Wikipedia, the free encyclopedia

Coordinates: 40°39'51"N 73°56'19"W

"NYC" and "New York, New York" redirect here. For other uses, see NYC (disambiguation) and New York, New York (disambiguation).

This article is about the city. For other uses, see New York City (disambiguation).

New York is the most populous city in the United States[10] and the center of the New York Metropolitan Area, one of the most populous metropolitan areas in the world.[11][12][13] The city is referred to as **New York City** or **The City of New York**[14] to distinguish it from the State of New York, of which it is a part.[15] A global power city,[16] New York exerts a significant impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment. The home of the United Nations Headquarters,[17] New York is an important center for international diplomacy[18] and has been described as the cultural capital of the world.[19]

Located on one of the world's largest natural harbors,[20] New York City consists of five boroughs, each of which is a state county.[21] The five boroughs—The Bronx, Brooklyn, Manhattan, Queens, and Staten Island—were consolidated into a single city in 1898.[22][23] With a Census-estimated 2011 population of 8,244,910[24] distributed over a land area of just 305 square miles (790 km^2),[25][26][27] New York is the most densely populated major city in the United States.[28] As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world.[29] The New York City Metropolitan Area's population is the United States' largest, with 18.9 million people distributed over 6,720 square miles (17,400 km^2),[30][31] and is also part

New York

— City —

The City of New York

Example 4

Under Pienciak's leadership at the Daily News, the investigative team has won numerous awards for its work, most notably for its exhaustive series "9/11 Money Trough," which examined how the $21.4 billion the federal government gave New York to recover from the Sept. 11 attacks was misspent and mismanaged.

Daily News (New York)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2011)*

The ***Daily News*** of New York City is the fourth most widely circulated daily newspaper in the United States.[2]

Daily News

New York City

From Wikipedia, the free encyclopedia

Coordinates: 40°39'51"N 73°56'19"W

"NYC" and "New York, New York" redirect here. For other uses, see NYC (disambiguation) and New York, New York (disambiguation).

This article is about the city. For other uses, see New York City (disambiguation).

New York is the most populous city in the United States[10] and the center of the New York Metropolitan Area, one of the most populous metropolitan areas in the world.[11][12][13] The city is referred to as **New York City** or **The City of New York**[14] to distinguish it from the State of New York, of which it is a part.[15] A global power city,[16] New York exerts a significant impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment. The home of the United Nations Headquarters,[17] New York is an important center for international diplomacy[18] and has been described as the cultural capital of the world.[19]

Located on one of the world's largest natural harbors,[20] New York City consists of five boroughs, each of which is a state county.[21] The five boroughs—The Bronx, Brooklyn, Manhattan, Queens, and Staten Island—were consolidated into a single city in 1898.[22][23] With a Census-estimated 2011 population of 8,244,910[24] distributed over a land area of just 305 square miles (790 km^2),[25][26][27] New York is the most densely populated major city in the United States.[28] As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world.[29] The New York City Metropolitan Area's population is the United States' largest, with 18.9 million people distributed over 6,720 square miles (17,400 km^2),[30][31] and is also part

New York

— City —

The City of New York

Example 4

Under Pienciak's leadership at the Daily News, the investigative team has won numerous awards for its work, most notably for its exhaustive series "9/11 Money Trough," which examined how the $21.4 billion the federal government gave New York to recover from the Sept. 11 attacks was misspent and mismanaged.

Daily News (New York)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2011)*

The ***Daily News*** of New York City is the fourth most widely circulated daily newspaper in the United States.[2]

Daily News

New York City

From Wikipedia, the free encyclopedia

Coordinates: 40°39'51"N 73°56'19"W

"NYC" and "New York, New York" redirect here. For other uses, see NYC (disambiguation) and New York, New York (disambiguation).

This article is about the city. For other uses, see New York City (disambiguation).

New York is the most populous city in the United States[10] and the center of the New York Metropolitan Area, one of the most populous metropolitan areas in the world.[11][12][13] The city is referred to as **New York City** or **The City of New York**[14] to distinguish it from the State of New York, of which it is a part.[15] A global power city,[16] New York exerts a significant impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment. The home of the United Nations Headquarters,[17] New York is an important center for international diplomacy[18] and has been described as the cultural capital of the world.[19]

Located on one of the world's largest natural harbors,[20] New York City consists of five boroughs, each of which is a state county.[21] The five boroughs—The Bronx, Brooklyn, Manhattan, Queens, and Staten Island—were consolidated into a single city in 1898.[22][23] With a Census-estimated 2011 population of 8,244,910[24] distributed over a land area of just 305 square miles (790 km^2),[25][26][27] New York is the most densely populated major city in the United States.[28] As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world.[29] The New York City Metropolitan Area's population is the United States' largest, with 18.9 million people distributed over 6,720 square miles (17,400 km^2),[30][31] and is also part

New York

— City —

The City of New York

Heuristics

- Heuristic 3

> m_2 is the query mention. m_1 and m_2 are coreferent. Assume m_1 was linked to e_1 and occurs before m_2. m_2 also is linked to e_1.

- Two criteria:
  - m_1 occurs before any other its coreferent mentions and is the longest, or
  - m_1 occurs before any other its coreferent mentions and is the main alias of e_1

Example 5

Stanford University has acquired historic recordings of spiritual leaders the Dalai Lama and Jiddu Krishnamurti, author Joseph Campbell and thousands of other intellectual figures, the university announced Monday.

Stanford bought the 6,000-hour collection from New Dimensions Broadcasting Media Network, which airs interviews on public and community radio stations. The collection also includes recordings of Buckminster Fuller, Timothy Leary, Deepak Chopra, Bill Moyers, Alice Walker, Maya Angelou and about 3,000 others.

Example 5

Stanford University has acquired historic recordings of spiritual leaders the Dalai Lama and Jiddu Krishnamurti, author Joseph Campbell and thousands of other intellectual figures, the university announced Monday.

Stanford bought the 6,000-hour collection from New Dimensions Broadcasting Media Network, which airs interviews on public and community radio stations. The collection also includes recordings of Buckminster Fuller, Timothy Leary, Deepak Chopra, Bill Moyers, Alice Walker, Maya Angelou and about 3,000 others.

Linking Algorithm overview

- Pre-processing
  - Rule-based NE recognition
  - Rule-based coreference resolution
- Hybrid statistical and rule-based incremental algorithm
  - Step 1: Applying heuristics
  - Step 2: For remaining ambiguous names match their feature vector with those of their Wikipedia candidate entities

VSM for Entity Linking

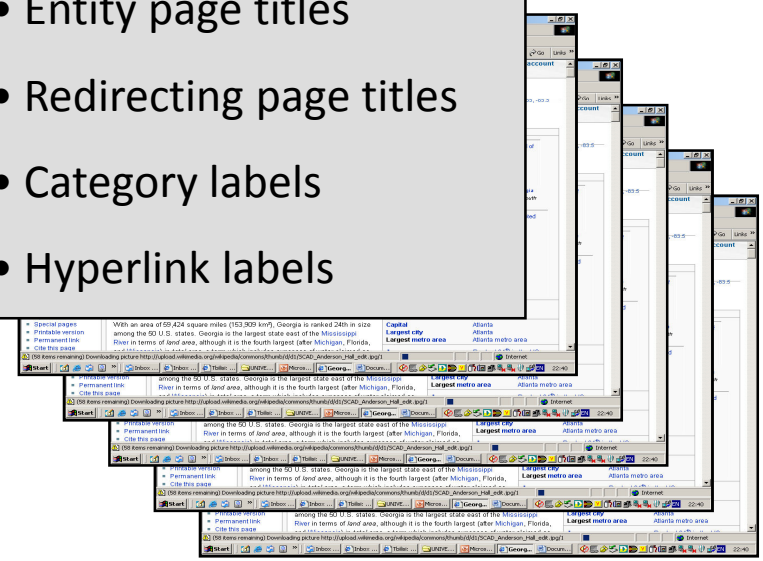
Text containing ambiguous mentions

- All mentions
- All words in the window text centred around the ambiguous mention and its coreferent ones
- Article titles of entities that have already been identified

TF-IDF vector similarity

Wikipedia article

- Entity page titles
- Redirecting page titles
- Category labels
- Hyperlink labels



Experiments

Query	MAA	B-Cubed+ F1
All (2,250)	75.8%	72.8%
NIL (1,126)	93.7%	90.4%
Non-NIL (1,124)	57.8%	56%

TABLE 7 - The MAA and B-Cubed+ F1 overall results of JVN_TDT2 on TAC-KBP2011 dataset

Query	MAA	B-Cubed+ F1
All (2,226)	57.1%	47%
NIL (1,049)	75.7%	60.9%
Non-NIL (1,177)	40.5%	37.5%

TABLE 8 - The MAA and B-Cubed+ F1 overall results of JVN_TDT2 on TAC-KBP2012 dataset

Conclusion

- We presented two methods.
  - The first one applied a learning model and exploited coreference relations among mentions to perform entity linking
  - The second one combined heuristics with a statistical model and performed entity linking in a incremental algorithms
- Experiment results showed that coreference relations among mentions significantly contribute to the performance of entity linking systems and
- And the proposed heuristics are potential for improving the performance of entity linking systems.

Thank you

Reference

- Milne, D. and Witten, I.H. (2008). Learning to Link with Wikipedia. In: *Proc. of the 17th ACM CIKM* (CIKM 2008), pp. 509-518.
- Medelyan, O., Witten, I. H., Milne, D. (2008). Topic indexing with Wikipedia. In *Proc. of Wikipedia and AI workshop at the AAAI-2008 Conference*.
- Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., and Cunningham, H. (2002). Shallow Methods for Named Entity Coreference Resolution. In *Proc. of TALN 2002 Workhop*.