Improving Wikipedia-based Place Name Disambiguation in Short Texts Using Structured Data from DBpedia

Yingjie Hu [1], Krzysztof Janowicz [1], and Sathya Prasad [2]
presented by Grant Mckenzie [1]

[1] STKO Lab, UC Santa Barbara
[2] Applications Prototype Lab, Esri Inc.

Outline

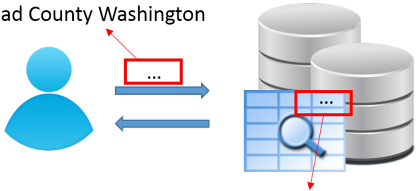
- Introduction
- Wikipedia-based Place Name Disambiguation
- Proposed Method
- Experiments
- Conclusions and Future Work

Introduction

- Place name disambiguation is important for geographic information retrieval (GIR)

- Two aspects of GIR that can benefit from disambiguation:
  - □ Understanding user's input query
  - □ Indexing records in database

User input: Hempstead County Washington



Description: Washington is home to the Historical Washington State Park

Introduction

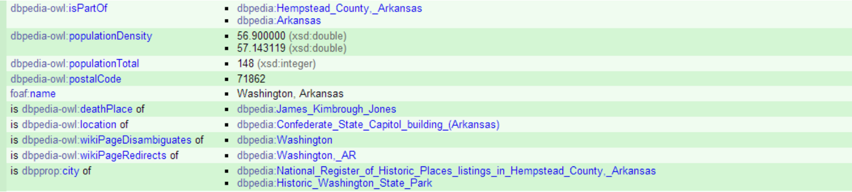
- Challenges of place name disambiguation– ambiguity of toponyms
- Synonymy: different places can have the same name
  - □ e.g., 25 Washingtons in the U.S.
  - □ Using context surrounding the place name for disambiguation
- Polysemy: the same place have multiple different names
  - □ e.g., California is also called Golden state
  - □ Add alias for places in a database

Introduction

- Many applications require disambiguation in **short texts**
  - □ Users input queries are often short
  - □ Data descriptions (snippets) are often short
  - □ ...
- Short texts contain very limited context information
  - □ Recognizing entities is important for increasing the disambiguation accuracy in short texts
  - □ e.g., UCSB is located in Santa Barbara

Introduction

- This work integrates the structured data from DBpedia to enhance Wikipedia-based place name disambiguation
- What is DBpedia?
  - DBpedia is the Semantic Web version of Wikipedia
  - The content of DBpedia is based on Wikipedia, but use structured data to represent entities (e.g., places, persons, organizations) and their relations



dbpedia-owl:isPartOf	▪ dbpedia:Hempstead_County,_Arkansas ▪ dbpedia:Arkansas
dbpedia-owl:populationDensity	▪ 56.900000 (xsd:double) ▪ 57.143119 (xsd:double)
dbpedia-owl:populationTotal	▪ 148 (xsd:integer)
dbpedia-owl:postalCode	▪ 71862
foaf:name	▪ Washington, Arkansas
is dbpedia-owl:deathPlace of	▪ dbpedia:James_Kimbrough_Jones
is dbpedia-owl:location of	▪ dbpedia:Confederate_State_Capitol_building_(Arkansas)
is dbpedia-owl:wikiPageDisambiguates of	▪ dbpedia:Washington
is dbpedia-owl:wikiPageRedirects of	▪ dbpedia:Washington,_AR
is dbpprop:city of	▪ dbpedia:National_Register_of_Historic_Places_listings_in_Hempstead_County,_Arkansas ▪ dbpedia:Historic_Washington_State_Park

Wikipedia-based Place Name Disambiguation

A two-stage process

- Stage 1: spotting
  - □ Goal: from the descriptions, identify the terms (called *surface forms*) that can be used to represent place names
  - □ E.g., recognizing "Washington" can be used for place name without disambiguating which "Washington" it refers to

- Three Wikipedia sources for spotting:
  - □ Article titles: provide the formal name of a place, e.g., *Washington, D.C.*
  - □ Redirect pages: provide the common alias of a place, e.g., *United States Capital*
  - □ Disambiguation page: provide the place names which people often use and which may refer to multiple places, e.g., *Washington*

Wikipedia-based Place Name Disambiguation

A two-stage process

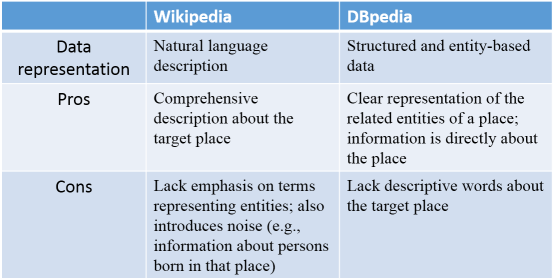
- Stage 2: disambiguation
  - Goal: find the actual place entity that a place name refers to
  - Existing methods include: entity prominence, context similarity, and a combined approach
- Entity prominence:
  - Importance of place entities:
    - E.g., Washington D.C. is generally more important than other Washingtons
  - How to quantify this importance:
    - Page in links, i.e., how many other pages linking to this page
    - Geographic features: Population, total area, ...
    - These information are available from DBpedia

Wikipedia-based Place Name Disambiguation

- Stage 2: disambiguation
- Context similarity
  - □ How similar is the context information of a place name compared with the Wikipedia descriptions.
  - □ E.g., Cosine similarity
- Combining Entity prominence with Context Similarity
  - □ A place name generally refers to the most popular place, unless there is strong context evidence that suggests otherwise
  - □ E.g., Bayesian theorem

Proposed Method

- Integrating DBpedia into place name disambiguation
- Pros and cons of Wikipedia and DBpedia



	Wikipedia	DBpedia
Data representation	Natural language description	Structured and entity-based data
Pros	Comprehensive description about the target place	Clear representation of the related entities of a place; information is directly about the place
Cons	Lack emphasis on terms representing entities; also introduces noise (e.g., information about persons born in that place)	Lack descriptive words about the target place

Example: "Greenville is one of the newest and smallest towns in Hillsborough County." will be converted into two triples:

:Greenville a :Town.
:Greenville :isPartOf :Hillsborough County

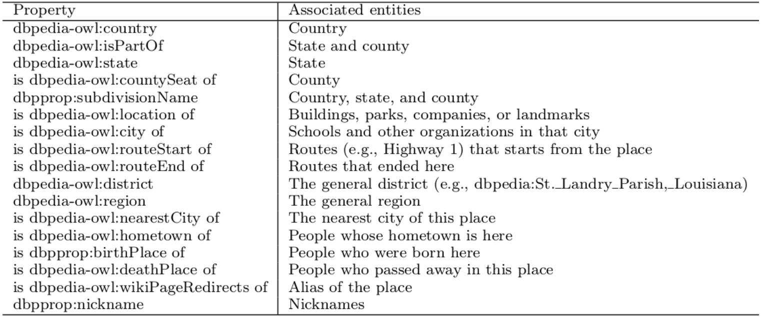
Proposed Method

- Three steps for integrating DBpedia and Wikipedia
- Step 1: employ the vector space model in existing works, but using the content from both of the two knowledge bases
  - □ Merits: puts more emphasis on the entity terms, while keep the descriptive words
  - □ Limits: breaks the structured nature of DBpedia data

Example: "Greenville is one of the newest and smallest **towns** in **Hillsborough County**."

Proposed Method

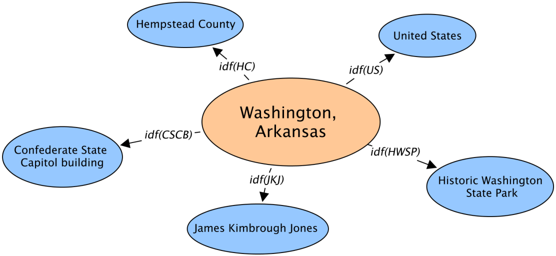
- Step 2: identify place-related entities from DBpedia to improve place name disambiguation
  - □ We use a list of DBpedia properties which connect the target place to related entities:



Property	Associated entities
dbpedia-owl:country	Country
dbpedia-owl:isPartOf	State and county
dbpedia-owl:state	State
is dbpedia-owl:countySeat of	County
dbpprop:subdivisionName	Country, state, and county
is dbpedia-owl:location of	Buildings, parks, companies, or landmarks
is dbpedia-owl:city of	Schools and other organizations in that city
is dbpedia-owl:routeStart of	Routes (e.g., Highway 1) that starts from the place
is dbpedia-owl:routeEnd of	Routes that ended here
dbpedia-owl:district	The general district (e.g., dbpedia:St._Landry_Parish,_Louisiana)
dbpedia-owl:region	The general region
is dbpedia-owl:nearestCity of	The nearest city of this place
is dbpedia-owl:hometown of	People whose hometown is here
is dbpprop:birthPlace of	People who were born here
is dbpedia-owl:deathPlace of	People who passed away in this place
is dbpedia-owl:wikiPageRedirects of	Alias of the place
dbpprop:nickname	Nicknames

Proposed Method

- Step 2: identify place-related entities from DBpedia to improve place name disambiguation
  - □ Entity importance is determined by the uniqueness of that entity to the target place
  - □ E.g., the entity **U.S.** has the lowest importance, since all the **Washington**s to be disambiguated are related to it
  - □ Inverse document frequency (IDF) for the importance of terms



Proposed Method

- Step 3: combine the previous two components using a smoothing parameter λ:

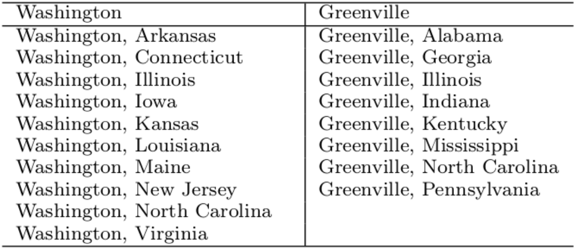
$$\begin{aligned} S(s \rightarrow e_i) = & \lambda Match(Context(s), Entities(e_i)) + \\ & (1-\lambda) Sim(Context(s), WD(e_i)) P(s \rightarrow e_i) \end{aligned}$$

Where $\lambda \in [0, 1]$, and it controls the relative importance of the two parts in the equation.

Experiments

- Experimental place names:
  - □ So far, two highly ambiguous place names to test our method
  - □ **Washington** (10 places) and **Greenville** (8 places)

- Experimental data:
  - □ Government description data (ground truth)
  - □ Wikipedia data
  - □ DBpedia data



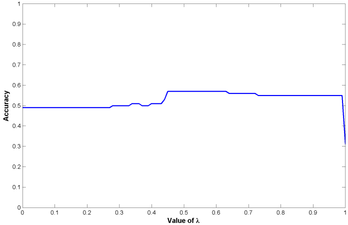
Washington	Greenville
Washington, Arkansas	Greenville, Alabama
Washington, Connecticut	Greenville, Georgia
Washington, Illinois	Greenville, Illinois
Washington, Iowa	Greenville, Indiana
Washington, Kansas	Greenville, Kentucky
Washington, Louisiana	Greenville, Mississippi
Washington, Maine	Greenville, North Carolina
Washington, New Jersey	Greenville, Pennsylvania
Washington, North Carolina	
Washington, Virginia	

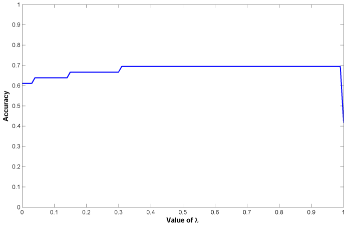
Experiments

- Experimental data preparation:
  - □ Government descriptions were separated into sentences using regular expression (to ensure short texts testing environment)
  - □ Numbers in DBpedia were kept in the model, while numbers in Wikipedia were removed

- Effects of λ:

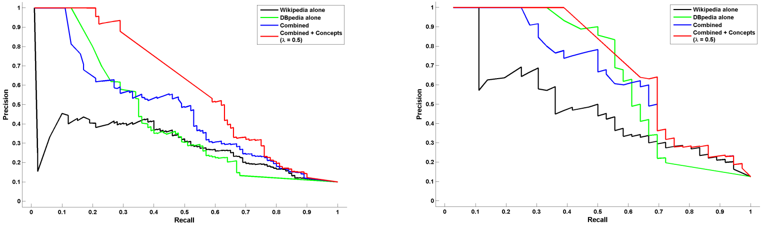
$$S(s \rightarrow e_i) = \lambda Match(Context(s), Entities(e_i)) + (1 - \lambda) Sim(Context(s), WD(e_i)) P(s \rightarrow e_i)$$





Experiments

- Comparing our method with three baselines
  - □ Only use Wikipedia
  - □ Only use DBpedia
  - □ Only use Wikipedia and DBpedia in vector space model
- Results:



Red: our approach; Black: Wikipedia; Green: DBpedia alone; Blue: vector space combining Wiki and DB

Conclusions

- A pure vector space model does not give enough emphasis to the terms representing entities
- DBpedia provides relatively comprehensive information about entities related to a place
- We propose a method which combines DBpedia and Wikipedia to improve place name disambiguation
- Data and source code is available on Github for further test: https://github.com/YingjieHu/Place-Disambiguation/

Feedback and questions are very welcome: yingjiehu@geog.ucsb.edu

Thank you for your attention!