A Coherent Unsupervised Model for Toponym Resolution

Ehsan Kamalloo and Davood Rafiei

University of Alberta

At a Glance

- *Goal*: Map location mentions in a document to a geographical reference
- *Challenges*: Different places with same name are abundant
  - Paris, France
  - Paris, Ontario, Canada
  - Paris, Texas, U.S.
- Related Works
- Unsupervised Approaches
- Evaluations

Problem: GeoTagging

- Given a document D

… The jobless rate for wider Northeast Georgia, which includes Barrow and Jackson counties, inched closer to double-digit figures in February, ...

- The objective is to annotate location mentions in D using geographical references
- Performed in two phases

Phase I: Recognition

- Given a document D

… The jobless rate for wider Northeast Georgia, which includes Barrow and Jackson counties, inched closer to double-digit figures in February, ...

- *Goal:* Detect location mentions (a.k.a **toponyms**)
- *Output:* A sequence of toponyms $T = t_1, \cdots, t_K$
- Typically done using Named Entity Recognizers (NER)

Phase II: Resolution

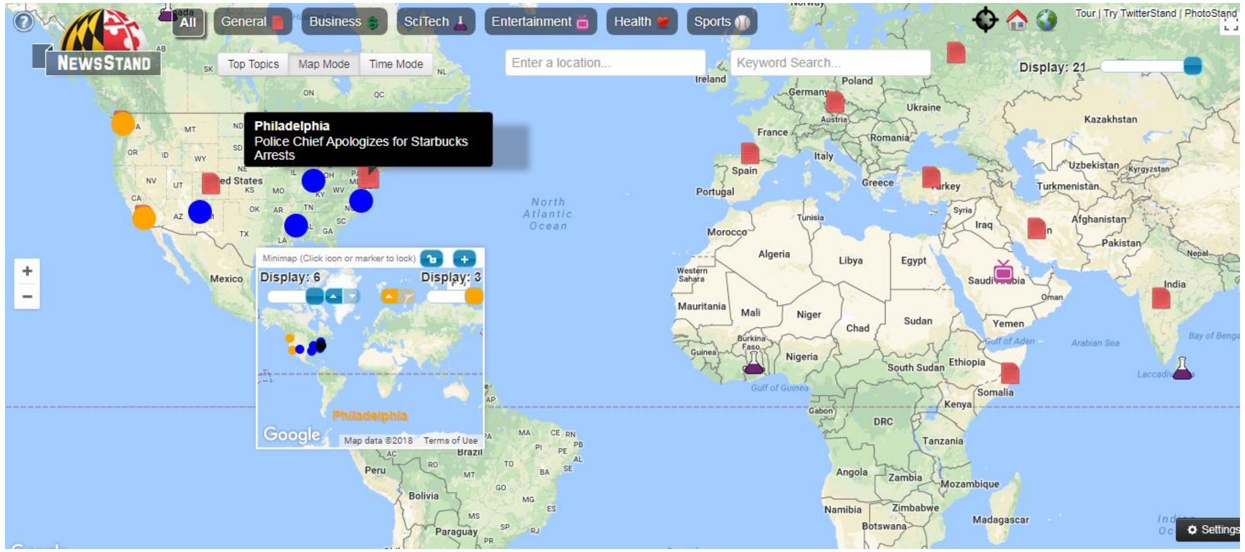
- Given a document D

… The jobless rate for wider Northeast Georgia, which includes Barrow and Jackson counties, inched closer to double-digit figures in February, ...

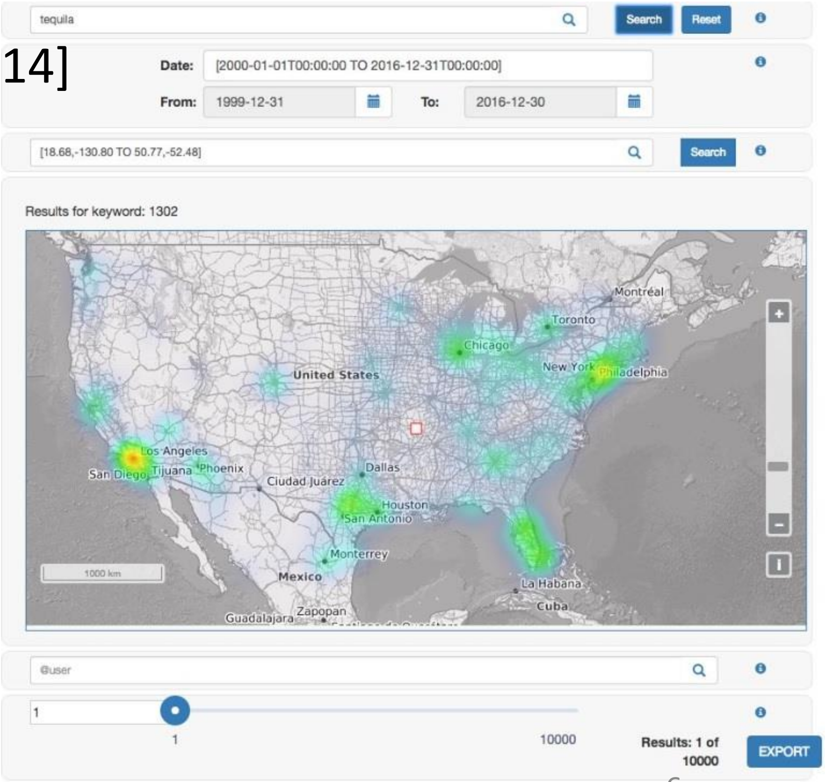
- And a sequence of toponyms $T = t_1 , \cdots , t_K$
- *Goal:* ground each toponym t_i to a geographic footprint (latitude/longitude)
- Coordinates are derived from a location database (a.k.a **Gazetteer**)
- **GeoNames** is adopted as gazetteer

Applications

- NewsStand [Teitler et al. 2008]
- TwitterStand [Sankaranarayananet al. 2009]
- VisCAT: Event detection on Twitter [Ghanem et al. 2014]
- Spatio-Temporal Search Plaform [Lewis et al. 2016]



http://newsstand.umiacs.umd.edu/web/



Challenges: Name Ambiguities

- Many place names have multiple interpretations

The November 2015 Paris attacks were the deadliest in the country since World War II.

Paris was voted ‘the Prettiest Little Town in Canada’ by Harrowsmith Magazine.

- GeoNames lists **97** candidates for *Paris*

Challenges: Immense Search Space

- Consider an article about U.S. states

… Washington (113) … California (225) … Florida (228) … Colorado (230) … Arizona (63) … Texas (53) …

The number of interpretations in GeoNames

- Leads to more than 4 billion cases
- In our datasets, news articles include 8 toponyms on average
- Heuristics such as picking largest population can help
  - Works poorly in dealing with localized context

Minimality Properties [Leidner 2007]

- Based on Cooperative Principle
  - Documents are encapsulated by extra-linguistic context where the audience is believed to understand the intention of an ambiguous term.

1. **One-sense-per-referent**
2. **Spatial-minimality**

- Adopted by virtually all toponym resolvers

Today Georgia skates in Red Deer, Innisfail and Edmonton for additional training and practises with coaches.

Related Works

- Unsupervised and rule-based
- Knowledge-based
  - TopoCluster [DeLozier et al. 2015]
- Supervised
  - Adaptive [Lieberman and Samet 2012]
- Entity-linking

Unsupervised Approach

- Why Unsupervised methods?
  - Lack of large enough annotated data
  - Data collected for a specific region
- *Goal:* Design an off-the-shelf resolver wherein no additional information other than gazetteer is required
- How?
  - Using contextual features of text as clues
  - Interactions between toponym interpretations

Context-Bound Hypotheses (CBH)

- Inspired by a named entity geotagging method [Yu and Rafiei 2016]
  - Given a named entity and a set of documents, capture the geographic focus of the named entity
- A probabilistic model grounded on two hypotheses
  1. Geo-centre Inheritance
  2. Near-location

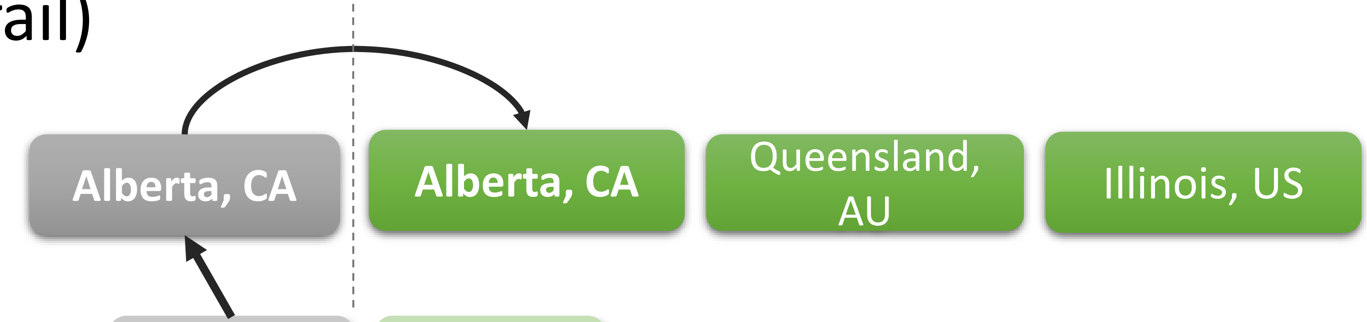
1. Geo-Centre Inheritance

- The geographic scope of document can disambiguate toponyms
- Given the scope of the following document is Canada:

Today Georgia skates in Red Deer, Innisfail and Edmonton for additional training and practises with coaches.

2. Near-Location

- Nearby Toponyms are more likely to be linked to one another
  - Comma-groups [Lieberman et al. 2010]
  - Object/containers [Lieberman et al. 2010]
- A known mapping (Red Deer) is exploited to resolve a neighboring toponym (Innisfail)



Today Georgia skates in Red Deer, Innisfail and Edmonton for additional training and practises with coaches.

Preliminary Resolution

- CBH is preceded by a preliminary disambiguation phase
  - Estimate the graphic scope of document
  - Find an initial setting for near-location hypothesis

Preliminary Resolution Example

Minimum distance = 10 words

Minimum distance = 2 words

He doesn't have any connections to London, Ont., ... But the police officer from Hampshire, England took off the day from work Wednesday so he … to the largest military cemetery in the UK to honour Sanborn's life... To make sure Canada recognizes that sacrifice, he submitted Sanborn's name to London's commemorative street name program.*

Minimum distance = 20 words

Minimum distance = 30 words

London candidates	Score
England, UK	$\frac{1}{10} + \frac{1}{20} = 0.15$
Ontario, CA	$\frac{1}{2} + \frac{1}{30} = \mathbf{0.53}$
Kentucky, US	0

Ontario, CA

*An excerpt from cbc.ca news

Problem in Pre. Resolution

- Tie breaker: **highest population heuristic**
- Works poorly when no mentions of location in spatial hierarchy found
  - Ties occur frequently
  - Resolution would stick to the most populous candidate

King's Highway 401, commonly referred to as Highway 401 … is a controlled-access 400-series highway … Toronto … London … Kingston …*

Canada U.K. Jamaica

*From “Ontario Highway 401” Wikipedia article

CBH: Probabilistic model

- Resolution proceeds to compute hypotheses probabilities
- Resolution method
  - Starts with the lowest non-leaf spatial division (i.e., “county”)
  - Picks a toponym to compute the probabilities
  - Confidence: the linear combination of the estimated probabilities
  - Resolution rectified only if the candidate with highest confidence altered
  - Otherwise, continues to the parent division
- The procedure repeats until no modification can be performed or the number of iterations exceeds a limit

1. Geo-center inheritance

- Maximum likelihood of term frequency for an ancestor at division d in all toponyms
- At division d=‘*country*’, estimating geo-center hypothesis for *London*

King's Highway 401, commonly referred to as Highway 401 … is a controlled-access 400-series highway … Toronto … London … Kingston …

Canada

Jamaica

- *London* interpretations = {Canada, U.K., U.S.}

d=Country	**tf**	$P_{\text{inh}}^{(\text{d})}$
Canada	2	$^{2}/_{4}$
U.K.	1	$^{1}/_{4}$
U.S.	1	$^{1}/_{4}$

2. Near-Location

- Maximum likelihood of similarity between an ancestor at division d and all toponyms
- Similarity function: Inverse of minimum term distance between two mentions (as in Preliminary Resolution)
- At division d='*country*', near-location probability for *London*

King's Highway 401, commonly referred to as Highway 401 … is a controlled-access 400-series highway … Toronto … London … Kingston …

Minimum distance = 10 words

Canada

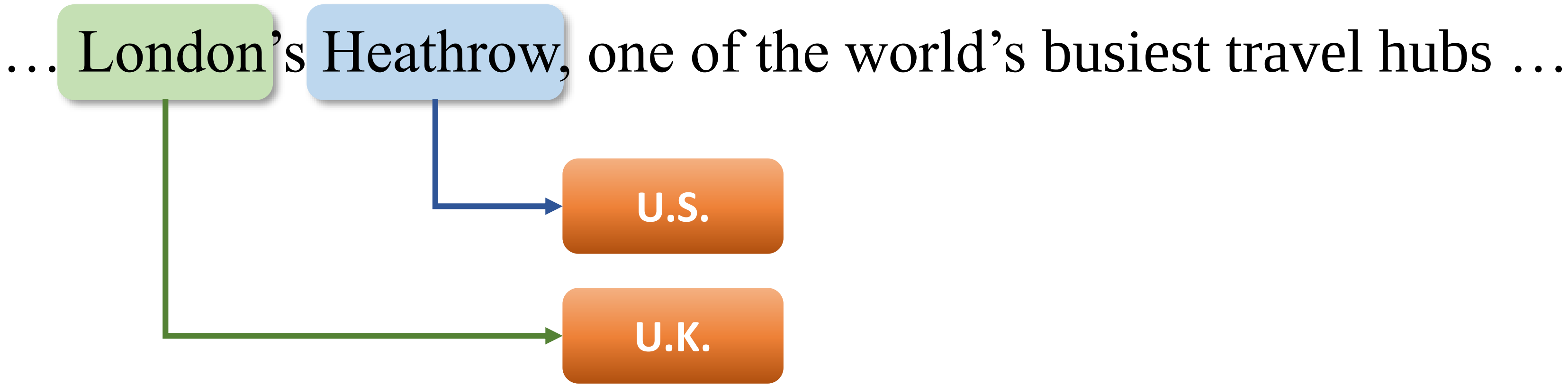
Jamaica

Minimum distance = 5 words

d=Country	sim	$P^{(d)}_{near}$
Canada	0.1	**1**
U.K.	0	0
U.S.	0	0

CBH: Infinite Loop Trap

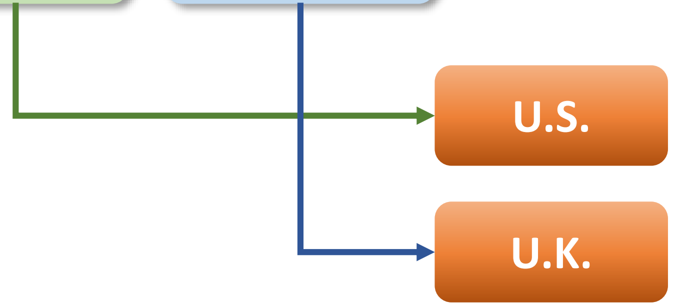
1. Preliminary Resolution
   - Highest population selected because no mentions of parents found



CBH: Infinite Loop Trap (cntd.)

2. First iteration: the probabilistic model
   - For London, Heathrow $\mapsto$ U.S. increases the probability of U.S.
   - For Heathrow, London $\mapsto$ U.K. increases the probability of U.K.

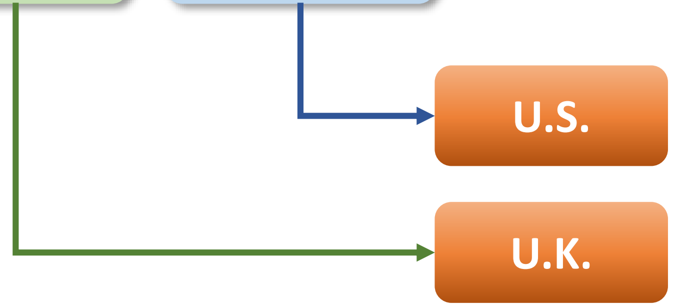
… London's Heathrow, one of the world's busiest travel hubs …



CBH: Infinite Loop Trap (cntd.)

3. Next Iteration: the probabilistic model
    - For London, Heathrow ↦ U.K. increases the probability of U.K.
    - For Heathrow, London ↦ U.S. increases the probability of U.S.

… London's Heathrow, one of the world's busiest travel hubs …



4. And so on…

maxIterations parameter introduced to avoid these cases

Spatial Hierarchy Sets

- *Goal*: Preserve minimality properties
- The whole universe (gazetteer) are partitioned into geographically related structures
  - Based on containment and sibling relationships
- Find a minimal set of partitions to cover all toponyms

SHS Resolution

… London's Heathrow, one of the world's busiest travel hubs …

1 set vs. 2 sets

✓

SHS Weaknesses

- Minimality happens in ancestors
  - Unable to detect: Montreal $\mapsto$ Quebec and Windsor $\mapsto$ Ontario
    - Because there is Windsor $\mapsto$ Quebec
- Insufficient clues
  - Georgia $\mapsto$ Texas and Turkey $\mapsto$ Texas
  - Georgia (country) and Turkey (country)

Context Hierarchy Fusion (CHF)

- Use benefits of both models
  - Context-Bound Hypotheses
  - Spatial Hierarchy Sets
- Resolves based on CBH only if confidence is higher than a threshold
- Otherwise, SHS selects an interpretation

Experiment Setup

- Datasets
  - CLUST [Lieberman and Samet 2011]: 1082 articles, 11.5K toponyms
  - LGL [Lieberman et al. 2010]: 588 articles, 4.5K toponyms (contains geographically localized content)
  - TR-News: 118 articles, 1.3K toponyms
  - Toponyms not found in GeoNames: 3%
  - Wikipedia-linked toponyms: 94%
- Experiment Types
  - *GeoTag*: Recognition (NER) + Resolution
  - *Resolution*: Perfect Recognition + Resolution

Resolution Accuracy

- State-of-the-art techniques
  - *Supervised*: Adaptive context features [Lieberman et al. 2012]
  - *Unsupervised*: TopoCluster [DeLozier et al. 2015]
- Commercial products
  - Yahoo! YQL Placemaker
  - Thomson Reuter’s OpenCalais
  - Google Natural Language API

Unsupervised Comparison

	LGL					TR-News				
	P	R	F_1	P_{Resol}	M_{Resol}	P	R	F_1	P_{Resol}	M_{Resol}
Unsupervised										
CBH	66.8	40.6	50.5	68.6	**760**	74.9	53.0	62.1	79.2	**869**
SHS	**69.7**	**43.3**	**53.4**	68.3	1372	73.8	53.6	62.1	69.9	2305
CHF	68.5	43.1	52.9	**68.9**	818	**79.3**	**58.2**	**67.1**	**80.5**	942
TopoCluster	-	-	-	59.7	1228	-	-	-	68.8	1422

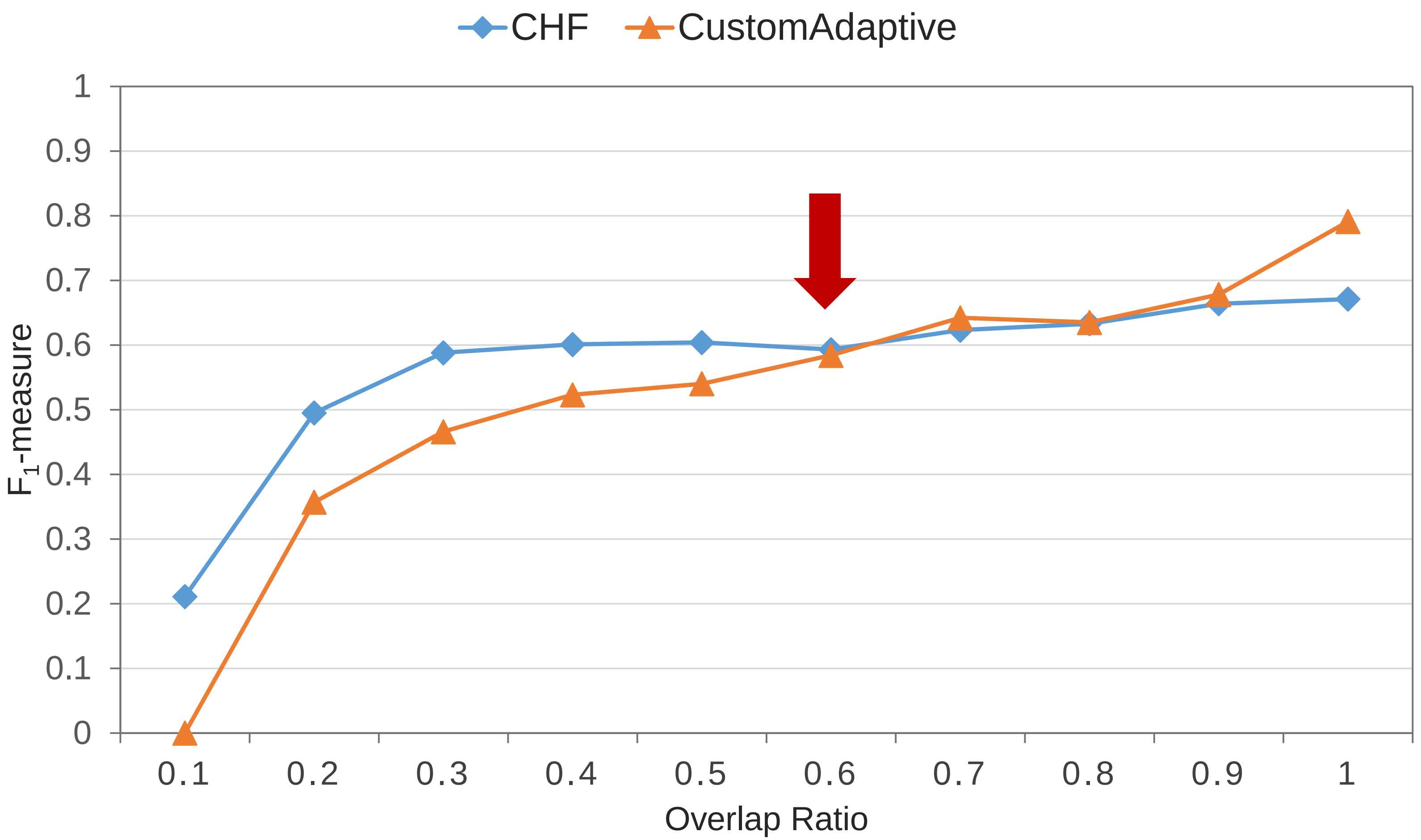
Spatial Hierarchies
- performs best in localized context
- yields high error distance

CHF performs best in more globalized context

Resolution Accuracy: comparison

	LGL					TR-News				
	P	R	F_1	P_{Resol}	M_{Resol}	P	R	F_1	P_{Resol}	M_{Resol}
Unsupervised										
CBH	66.8	40.6	50.5	68.6	**760**	74.9	53.0	62.1	79.2	**869**
SHS	**69.7**	**43.3**	**53.4**	68.3	1372	73.8	53.6	62.1	69.9	2305
CHF	68.5	43.1	52.9	**68.9**	818	**79.3**	**58.2**	**67.1**	**80.5**	942
TopoCluster	-	-	-	59.7	1228	-	-	-	68.8	1422
Supervised										
Adaptive*	**79.2**	**48.5**	**60.2**	**88.3**	**679**	**83.8**	**74.9**	**79.1**	**90.5**	**573**
Commercial										
Placemaker	73.5	**48.6**	**58.5**	-	-	80.8	**63.0**	**70.8**	-	-
OpenCalais	77.1	28.9	42.1	-	-	**81.3**	48.5	61.2	-	-
GoogleNL	**80.5**	34.0	47.8	-	-	80.2	38.4	51.9	-	-

Unseen Data Analysis



Overlap between toponyms in train data and test data channeled

Conclusions

- Introduced an unsupervised toponym resolver
- Future works
  - Investigate mixture models (supervised and unsupervised)
  - Study the correlation among the bounding-boxes of toponyms
- Code and data available at https://github.com/ehsk/CHF-TopoResolver