Master Thesis

_named Entity Recognition and Disambiguation with Wikidata on Arbitrary English Text_

Yi-Chun Lin  2021/09/22
Problem
The Tasks

Named Entity Recognition (NER): given a text, tell which words belongs to a named entity

Named Entity Disambiguation (NED): given the text span of a named entity, link it to its corresponding entry in a knowledge base.
NER + NED Example

Amazon was founded by Jeff Bezos

Q3884 (wikidata)  Q3783  Q312556
Problem Definition

input: plain text

Amazon was founded by Jeff Bezos.

output: text span of named entities + corresponding entries in Wikidata

“Amazon”: Q3884,
“Jeff Bezos”: Q312556
Solution
Base Model

Entity Index

Candidate List

Query Text

NER
Examine all possible text spans in query, speed up with POS-tag filter

NED
Consider popularity of each candidate and its similarity to query context

Output
Entity Index

key:  entity name and synonyms

value: candidate lists
      (all entities that have the name or synonym)

<table>
<thead>
<tr>
<th>QID</th>
<th>Name</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3884</td>
<td>Amazon</td>
<td>Amazon.com</td>
</tr>
<tr>
<td>Q3783</td>
<td>Amazon</td>
<td>Amazon River</td>
</tr>
</tbody>
</table>

“Amazon”: [Q3884, Q3783]
“Amazon.com”: [Q3884]
“Amazon River”: [Q3783]
**NER - 1/2**

**Task:** locate the text span of named entities in the query text.

**Basic approach:** part-of-speech (POS) tagging. Determine the grammatical category of each word.

✔️ Amazon was founded by Jeff Bezos  ✗ United States of America

<table>
<thead>
<tr>
<th>NNP</th>
<th>VBD</th>
<th>VBN</th>
<th>IN</th>
<th>NNP</th>
<th>NNP</th>
<th>NNP</th>
<th>NNP</th>
<th>IN</th>
<th>NNP</th>
</tr>
</thead>
</table>

9 📚 Named Entity Recognition and Disambiguation with Wikidata on Arbitrary English Text  Yi-Chun Lin
## NER - 2/2

Our approach: examine text spans starting with `NNP` or `NN`. Choose the longest match.

<table>
<thead>
<tr>
<th>Amazon was founded by Jeff Bezos</th>
<th>United States of America</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP VBD VBN IN NNP NNP</td>
<td>NNP NNP IN NNP</td>
</tr>
<tr>
<td>“Amazon”, “Amazon was”, ...</td>
<td>“United”, “United States”,</td>
</tr>
<tr>
<td>“Amazon was founded by Jeff Bezos”</td>
<td>“United States of”,</td>
</tr>
<tr>
<td>“Jeff”, “Jeff Bezos”</td>
<td>“United States of America”</td>
</tr>
</tbody>
</table>
NED - 1/3

Task: After NER, for each text span, choose the most suitable item among the candidates.

Basic approach: choose the most popular candidate.

✓ Obama was the president of the US.
✗ Obama is a city in Japan.
Query context plays an important role.

How to measure the similarity between context and each candidate:

- “context”: *NN* and *NNP* in the query
- “candidate”: words in its name, synonym and description
- “similarity”: overlaps between the two
Our approach: choose the candidate with the highest score, where score =

\[\text{score} = \text{popularity}_\text{score} + \text{similarity}_\text{score}\]

\[\text{popularity}_\text{score} = \text{sitelinks} \ (0 \sim P_{\text{max}})\]

\[\text{similarity}_\text{score} = \text{overlaps} \times \begin{cases} \frac{P_{\text{max}}}{3} & \text{when longer contexts (> 10 words)} \\ \frac{P_{\text{max}}}{2} & \end{cases}\]
Configurable Features

On top of the base model, each feature can be turned on/off to see its effectiveness.

- **Synonym Expansion**
  - Family Name
  - Demonym

- **KB Enrichment**
  - Large DB
  - Wiki Abstract

- **FP Reduction**
  - NNP Reduction
Family Name

Problem: Amazon was founded by Bezos.

“Jeff Bezos”: { Q312556 }
“Bezos”: {Q4900382 }

Solution: if an entity is of type “person” and has the property “family name”, add its family name to its synonym. “Bezos”: {Q4900382, Q312556 }
Demonym

Problem: Amazon is an American (adj.) company.

“Demonym” denotes the natives or inhabitants of a particular country, state, city, etc.

Solution: if the entity is a country and has the property “demonym”, add its demonym to its synonym. “USA”: {Q30}, “American”: {Q30}
Large DB

Problem: the condensed version of Wikidata excludes less popular entities, leads to recognition limits.

Solution: try the full version of Wikidata
Problem: Wikidata description too short, the algorithm falls back to depend only on popularity.

Solution: represent each entity with the words in corresponding Wikipedia abstract. Adjust weight s.t.

similarity score $\propto \frac{1}{\log(\# \text{ words})}$
NNP Reduction

Problem: “Bank Duta” recognized as “Bank” “Duta”
“Bank Duta”: an Indonesia bank (not in Wikidata)
“Bank”: a film by Charlie Chaplin (in Wikidata)
“Duta”: a family name (in Wikidata)

Solution: remove consequent single-word named entities. Exception: any of them’s score > $P_{max}$. 

Quick Recap

- **NER**
  - Examine all possible text spans in query, speed up with **POS-tag filter**

- **NED**
  - Consider **popularity** of each candidate and its **similarity to query context**

**Synonym Expansion**
- Family Name
- Demonym

**KB Enrichment**
- Large DB
- Wiki Abstract

**FP Reduction**
- NNP Reduction
Evaluation
Datasets

(1) AIDA CoNLL-YAGO news, manual annotated, 217 words/doc, 20 entities/doc

(2) ClueWeb12 FACC1 mixed, automatic annotated, 26 words/doc, 1.6 entities/doc, subset of 50000 docs.

SOCCER - FRANCE BEAT MEXICO 2-0 IN FRIENDLY. PARIS 1996-08-31 France beat Mexico 2-0 (halftime 0-0) in a friendly soccer international on Saturday. Scorers: Nicolas Ouedec (49th minute), Youri Djorkaeff (53rd) Attendance: 18,000

English colonists brought asparagus to North America, but asparagus did not become a commercial crop in the United States until the 19th century.
Metrics

We report **Micro F1** and **Macro F1** scores.

**F1 Score**: $2 \times P \times R / (P + R)$

**Precision**: ratio of correctly reported NEs among algorithm output

**Recall**: ratio of correctly reported NEs among the ground truth

**Micro**: aggregates data from all documents to compute one score

**Macro**: one score per document and takes average over all documents
### Results of Base Model + Single Feature

<table>
<thead>
<tr>
<th>configuration</th>
<th>Clueweb Micro F1</th>
<th>Clueweb Macro F1</th>
<th>AIDA Micro F1</th>
<th>AIDA Macro F1</th>
<th>memory (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>39.49</td>
<td>39.98</td>
<td>50.9</td>
<td>50.32</td>
<td>3.80</td>
</tr>
<tr>
<td>base + family name</td>
<td>39.62</td>
<td>40.48</td>
<td>53.47</td>
<td>52.29</td>
<td>3.89</td>
</tr>
<tr>
<td>base + demonym</td>
<td>41.78</td>
<td>42.74</td>
<td>55.65</td>
<td>56.34</td>
<td>3.80</td>
</tr>
<tr>
<td>base + large database</td>
<td>39.86</td>
<td>40.71</td>
<td>51.08</td>
<td>50.68</td>
<td>5.19</td>
</tr>
<tr>
<td>base + Wikipedia abstract</td>
<td>38.92</td>
<td>39.29</td>
<td>51.03</td>
<td>50.13</td>
<td>5.86</td>
</tr>
<tr>
<td>base + NNP reduction</td>
<td>47.06</td>
<td>42.95</td>
<td>54.26</td>
<td>53.22</td>
<td>3.80</td>
</tr>
</tbody>
</table>
### Effectiveness of Each Feature

<table>
<thead>
<tr>
<th>configuration</th>
<th>false positive</th>
<th></th>
<th>false negative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>counts</td>
<td>% change</td>
<td>counts</td>
<td>% change</td>
</tr>
<tr>
<td>base</td>
<td>89,305</td>
<td>-</td>
<td>53,194</td>
<td>-</td>
</tr>
<tr>
<td>base + family name</td>
<td>90,711</td>
<td>1.57%</td>
<td>51,745</td>
<td>-2.72%</td>
</tr>
<tr>
<td>base + demonym</td>
<td>89,066</td>
<td>-0.27%</td>
<td>48,696</td>
<td>-8.46%</td>
</tr>
<tr>
<td>base + large database</td>
<td>91,659</td>
<td>2.64%</td>
<td>52,056</td>
<td>-2.14%</td>
</tr>
<tr>
<td>base + Wikipedia abstract</td>
<td>89,943</td>
<td>0.71%</td>
<td>53,659</td>
<td>0.87%</td>
</tr>
<tr>
<td>base + NNP reduction</td>
<td>52,771</td>
<td>-40.91%</td>
<td>53,972</td>
<td>1.46%</td>
</tr>
</tbody>
</table>
## Comparison to AmbiverseNLU

<table>
<thead>
<tr>
<th>configuration</th>
<th>Clueweb Micro F1</th>
<th>Clueweb Macro F1</th>
<th>AIDA Micro F1</th>
<th>AIDA Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>enhanced</td>
<td>49.82</td>
<td>46.21</td>
<td>61.47</td>
<td>60.98</td>
</tr>
<tr>
<td>enhanced + large database</td>
<td>49.61</td>
<td>46.59</td>
<td>61.39</td>
<td>61.26</td>
</tr>
<tr>
<td>enhanced + Wikipedia abstract</td>
<td>48.78</td>
<td>45.18</td>
<td>60.53</td>
<td>60.17</td>
</tr>
<tr>
<td>full</td>
<td>48.55</td>
<td>45.56</td>
<td>62.31</td>
<td>61.50</td>
</tr>
<tr>
<td>AmbiverseNLU</td>
<td>44.75</td>
<td>33.58</td>
<td>68.57</td>
<td>67.78</td>
</tr>
</tbody>
</table>

*enhanced = base + family name + demonym + NNP reduction*
## Comparison to AmbiverseNLU

<table>
<thead>
<tr>
<th>model</th>
<th>Clueweb</th>
<th></th>
<th></th>
<th>AIDA</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tp</td>
<td>fp</td>
<td>fn</td>
<td>tp</td>
<td>fp</td>
<td>fn</td>
<td></td>
</tr>
<tr>
<td>enhanced</td>
<td>40,176</td>
<td>43,721</td>
<td>37,213</td>
<td>16,340</td>
<td>9,316</td>
<td>11,167</td>
<td></td>
</tr>
<tr>
<td>full</td>
<td>40,146</td>
<td>47,833</td>
<td>37,243</td>
<td>17,208</td>
<td>10,522</td>
<td>10,299</td>
<td></td>
</tr>
<tr>
<td>AmbiverseNLU</td>
<td>26,083</td>
<td>16,119</td>
<td>48,299</td>
<td>17,136</td>
<td>5,319</td>
<td>10,393</td>
<td></td>
</tr>
</tbody>
</table>