Leveraging User Interest to Improve Entity Disambiguation on Short Text

I. Connecting a user’s social Web identity and Wikipedia editor identity

Underlying assumptions: a user possesses a core set of interests, these interests are encoded in various forms of media, and a user’s interest in a topic is reflected in their engagement with that topic. We address the Named Entity Disambiguation problem for short, online texts by modeling a user’s engagement with a topic as a time series of clicks. By analyzing this time series, we are able to infer the user’s topic of interest and disambiguate the entity that best matches the user’s intended meaning from multiple candidate meanings.

EVALUATION

Results

• Baselines: randomly sorted candidates (RC), prior-freq (PF), RESOLVE given a random Wikipedia user’s interest model (RU), Wikipedia Miner (WM) and DBpedia Miner (DM) (for the shortest tweet).

• Best performance on YouTube texts (longest) due to content-based sim

• Outperforms on tweets (more personal) due to fewer noisy candidates

• Less effective on impersonal text (e.g., geo-location) by high prior freq evoked by standard methods suffice

Figure 2

Future Work

• Identity disambiguation & Modeling interests

• Computational

• Pruning strategy by b/millions of articles & hundreds of thousands of categories => require new vector items to meet “relevance” threshold to current items + standard strategies to remove very high or low freq items

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1144153 and is also supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (PIFOR:GA-2009-232506).

Table 1

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Coverage</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>0.8</td>
<td>0.72</td>
<td>0.76</td>
<td>0.71</td>
</tr>
<tr>
<td>Location</td>
<td>0.9</td>
<td>0.87</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>Organization</td>
<td>0.7</td>
<td>0.65</td>
<td>0.69</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Coming soon...

We address the Named Entity Disambiguation problem for short, user-generated texts on the social Web. In such settings, the lack of linguistic features and sparse lexical context result in high degrees of ambiguity and sharp performance drops of nearly 50% in the accuracy of conventional NER systems. We handle these challenges by developing a general model of user-interest with respect to a personal knowledge context and instantiating it using Wikipedia. We conduct extensive evaluations using individuals’ posts from Twitter, YouTube, and Flickr and demonstrate that our novel technique is able to perform well beyond state-of-the-art NER methods.

DISAMBIGUATION

Examples

Disambiguation hard in social Web setting (even for humans!)

Twitter (R of tweets/day) = 10 million page book: aahh one more day until final!!! Scanscan

YouTube video title: Flicker photo tag (7 billion images, 2x/month, user/mnt of just 4 yr ago)

Flickr "office" holiday: Beetle

Wikipedia Disambiguation pages show dozens of candidate meanings

Short Length & Informal

High Ambiguity

• Fundamentally different than traditional training text

• Longest still shorter than even shortest texts from NER task corpora like Reuters or Brown

• Many potential candidates (2 to 163 with avg. of 5.6-5.8 median 4)

• Tweets severely degree even text/metrics pairs:

- Stanford NER (trained on news articles): F1 drops 90% to 46%
- DBpedia Spotlight & Wikipedia Miner: Precision under 40%
- Recent attempts of NER & NER on Twitter also have limitations:

- Some use crowd-sourcing, but this causes dependency on reliable human workers (Demartini, 2012; Finn, 10)
- Automated attempts do extraction not disambiguation & generalization beyond Twitter itself (Davis, 2012; Li, 2011, Ritter, 2011)

Our Solution

Personalized NER: compensate for sparse lexical context with personal context, building a user interest model from external structured data and use it to select the sense most similar to personal interests.

RESOLVE

RESOLVE (Resolution Entity Sense by LeVoraging Edits) addresses NED.

Connecting a user’s social Web identity and Wikipedia editor identity

Modeling that user’s interests based on her Wikipedia article edits

Ranking entity candidate by measuring how similar each candidate’s topic is to salient topics in user’s interest model

Phase I: Account-binding

Given social Web post, RESOLVE takes username and try to get Wikipedia account belonging to same person. Simply string-matching usernames (prior work shows feasibility Itofu, 2011; Perito, 2011)

RESOLVE models user’s interests using bridged Wikipedia account on social Web site.

Identify candidate by measuring how similar each entity’s topic is to a user’s interest model

Phase II: Representing Topics

Wikipedia article as topic for merging Wikipedia pages

Phase III: Ranking

Wikipedia article as topic for merging Wikipedia pages

<table>
<thead>
<tr>
<th>Phase I: Account Binding</th>
<th>Phase II: Representing Topics</th>
<th>Phase III: Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity disambiguation</td>
<td>Association of user interests</td>
<td>User interests used</td>
</tr>
<tr>
<td>with a Wikipedia entity</td>
<td>with Wikipedia entities</td>
<td>to distinguish users</td>
</tr>
<tr>
<td>based on user editorial</td>
<td>based on Wikipedia articles</td>
<td>from different users</td>
</tr>
<tr>
<td>activity</td>
<td>and user editor activity</td>
<td>in the same context</td>
</tr>
</tbody>
</table>

| Table 2

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.72</td>
<td>0.71</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Implementation: The RESOLVE System

User’s contributions fairly homogeneous across race

Same topics: Ambig. YouTube post: office, december 3rd; same user’s logged Wikipedia edit: <item used=xander>xx user=xwxx pages=31841130 titles=The Office (U.S. season 9x)>

Same categories: On average, 52.4% of entities a user mentions in social Web ("Java") have at least 1 candidate meaning in same parent category of Wikipedia article she edited ("Programming language") & 60% if go just 4 parents up category hierarchy

Knowledge graph of categories, topics, and descriptions

User-topic graph

Instantiation on any knowledge base that categorizes topics is possible (e.g., Wikipedia, DBpedia, Freebase). Using Wikipedia because:

• Articles & categories effectively represent topics (Syed, 2008)

• Broadly covers entity concepts and rare word senses (Zesch, 2007) [key given topic diversity on social Web]

• Proven effective to model interests from editing behavior (Cosley, 2007)

• Can treat articles as in NER, article content as topic’s unique description, and “Category” as resources in NER

Theory

• Contribution: Users produce online content about key set of personally interesting topics because it is fulfilling and seen as better cost benefit (Harper, 2007; Ling, 2006, Collaku, 2008, NER on Tweets & Social Web)

• Modeling: Effective to model topics interesting to lexical features of the text-based contributions (Chen, 2010; Cosley, 2007; Pennacchiotti, 2011)

Motivations

Qualitative Analysis

QUICK DESIGN GUIDE

• Wikipedia Disambiguation pages show dozens of candidate meanings

• YouTube, and Flickr and demonstrate that our novel technique is able to perform well beyond state-of-the-art NER methods.

• Recent attempts of NER & NER on Twitter also have limitations:

- Some use crowd-sourcing, but this causes dependency on reliable human workers (Demartini, 2012; Finn, 10)
- Automated attempts do extraction not disambiguation & generalization beyond Twitter itself (Davis, 2012; Li, 2011, Ritter, 2011)

• Tweets severely degree even text/metrics pairs:

- Stanford NER (trained on news articles): F1 drops 90% to 46%
- DBpedia Spotlight & Wikipedia Miner: Precision under 40%

• Identifying interests is a complex problem:

- Content similarity, simUWERTS IC-BOW with TF-IDF terms of topics of articles user edited, candidate’s article title, words from those articles’ categories & title category

- Knowledge context based similarity, simUWERTS IC-BOW: Vectors of articles’ category IDs weighted by distance b/t topics & categories in know. graph

- Cosine sim between user & vector categories

Table 2

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.72</td>
<td>0.71</td>
<td>0.77</td>
</tr>
</tbody>
</table>

• Underlying assumptions: a user possesses a core set of interests, these interests are encoded in various forms of media, and a user’s interest in a topic is reflected in their engagement with that topic. We address the Named Entity Disambiguation problem for short, user-generated texts on the social Web. In such settings, the lack of linguistic features and sparse lexical context result in high degrees of ambiguity and sharp performance drops of nearly 50% in the accuracy of conventional NER systems. We handle these challenges by developing a general model of user-interest with respect to a personal knowledge context and instantiating it using Wikipedia. We conduct extensive evaluations using individuals’ posts from Twitter, YouTube, and Flickr and demonstrate that our novel technique is able to perform well beyond state-of-the-art NER methods.