RESLVE: Leveraging User Interest to Improve Entity Disambiguation on Short Text

Elizabeth L. Murnane elm236@cornell.edu
Bernhard Haslhofer bernhard.haslhofer@univie.ac.at
Carl Lagoze clagoze@umich.edu
Cornell University

INTRODUCTION

Abstract

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 a high degree of ambiguity and sharp performance drops of nearly 50% in the accuracy of conventional NED systems. We handle these challenges by developing a general model of user-interest with respect to a personal knowledge context and instantiate it using Wikipedia. We conduct systematic evaluations using individuals' posts from Twitter, YouTube, and Flickr and demonstrate that our novel technique is able to achieve performance gains beyond state-of-the-art NED methods.

Task Definitions

- **Named Entity Disambiguation (NED)**: Given *entity* detected during Named Entity Recognition (NER), NED is the resolution of its intended meaning from multiple *candidate meanings*
- ***Our focus***: disambiguating any entity (people, places, objects, concepts) detected in users' text-based utterances on the *social Web*

Examples

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

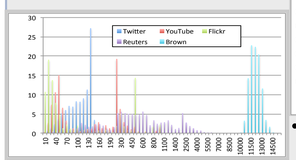
Twitter (# of tweets/day = 10 million page book):	YouTube video title (800 million users/month):	Flickr photo tag (7 billion images, 2x amount of just 4 yrs ago):
*aaahh one more day until **finn**!!! #cantwait*	*office holiday party*	*Beetle*

Wikipedia Disambiguation pages show dozens of candidate meanings

Challenges

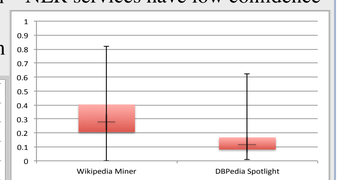
Short Length & Informal

- Fundamentally different than traditional training text
- Longest still shorter than even shortest texts from NER task corpora like Reuters or Brown



High Ambiguity

- 91% contain at least 1 ambig. entity
- 2/3 of detected entities ambig.
- NER services have low confidence



- Many potential candidates (2 to 163 with avg. of 5-6 & median 4)

Limitations of Extant Research

- Tweets severely degrade even state-of-the-art methods:
  - Stanford NER (trained on news articles): F1 drops **90% to 46%**
  - DBPedia Spotlight & Wikipedia Miner: **Precision under 40%**
- Recent attempts of NER & NED on Twitter also have limitations:
  - Some use crowd-sourcing, but this causes dependency on reliable human workers (Demartini, 2012; Finin, 2010)
  - Automated attempts do extraction not disambiguation & generalizability beyond Twitter unclear (Davis, 2012; Li, 2012; Ritter, 2011)

Our Solution

Personalized NED: 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.

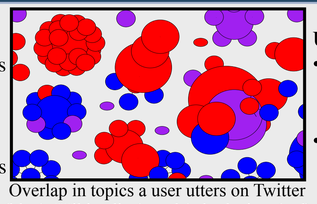
APPROACH

Underlying assumptions: a user possesses a core set of interests, these interests can be formally modeled with reliable and user-specific external semantic data, and a user is more likely to mention a topic from a domain of personal interest than non-interest, allowing candidate meanings to be ranked by personal relevance.

Motivations

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; Kollock, 1999; Maslow, 1970)
- **Modeling**: Effective to model these topic interests from lexical features of the text-based contributions (Chen, 2010; Cosley, 2007; Pennacchiotti, 2011)



Overlap in topics a user utters on Twitter (blue), Wikipedia (purple), both sites (red)

Qualitative Analysis

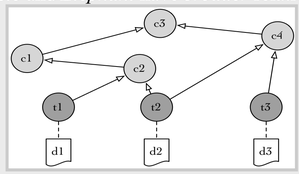
User's contributions fairly homogeneous across Web

- **Same topics**: Ambig. YouTube post: "*office*, december 3"; Same user's logged Wikipedia edit: <item userid=xx user=xx pageid=31841130 title=**The Office (U.S. season 8)**/>
- **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") & 100% if go just 4 parents up category hierarchy

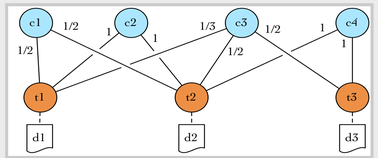
Modeling a Knowledge Context

A **knowledge base** is a semantic organization of entities, their types, and relations among them. We devise a graph representation $K=(N,E)$ with 2 node types: ***Category** nodes*. $N_{Category} \subset N$ with each $c \in N_{category}$ having a unique identifier $i \in I$ and a set of semantic relationships r with other nodes, so that $c = \{i,R\}$ (e.g., *Car* has *broader* relationship to *Vehicle*, and both *Vehicle* and *Elephant* have *broader* relationships to *Thing*).

***Topic** nodes*. $N_{Topic} \subset N$ with each $t \in N_{Topic}$ also having a unique identifier $i \in I$, belonging to one or more categories, and carrying a textual **description** d so that $t = \{I,d,C\}$ with $C \subseteq N_{Category}$ (e.g., *Volkswagen Beetle* is a topic of category *Car* and associated with a descriptive document – the article text in the case of Wikipedia.



Knowledge graph of categories, topics, and descriptions



User topic-interest graph

	c1	c2	c3	c4
t1	1/2	1	1/3	0
t2	1/2	1	1/2	1
t3	0	0	1/2	1

Edge-weight matrix of user interest 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 topics** in N_{topic}, article content as topic's unique **description**, and "Category:" resources as **categories** in $N_{category}$

Implementation: The RESLVE System

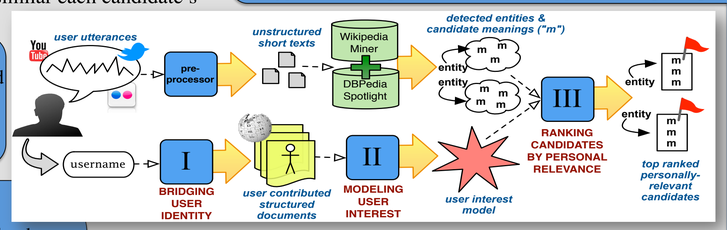
RESLVE (**R**esolving **E**ntity **S**ense by **LeV**eraging **E**dits) addresses NED by:

I. Connecting a user's social Web identity and Wikipedia editor identity
II. Modeling that user's interests based on her Wikipedia article edits
III. Ranking entity candidates by measuring how similar each candidate's topic is to salient topics in user's interest model

Pre-processing

Tweets: Normalize @name to MENTION, remove RT (retweet) tag, remove # but keep target concept. *YouTube, Flickr:* Bypass auto-gen names (IMG 336.jpg), remove file type suffix (.png) but leave filename, ignore auto-gen tags (hidden:filter=Boost).

All utterances: Remove URLs, non-English, single chars, parse errors, unambig entities (0 or 1 meaning), non Noun class (NN, NNS, NNP, NP) or IEER/ACE/CoNLL Named Entity class (loc, person, org)



Phase I: Account Bridging

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

Phase II: Representing Topics

RESLVE models user's topics of interest using bridged Wikipedia account's editing-history metadata (Table 1), ignoring trivial edits (Table 2) and compares the similarity of those topics to the topics associated with candidate meanings

- **Content similarity,** $sim_{content}(u.m)$: BOW with TF-IDF term vectors of titles of articles user edited, candidate's article title, words from those articles' pages & category titles
- **Knowledge-context based similarity,** $sim_{category}(u.m)$ **:** Vectors of articles' category IDs weighted by distance b/w topics & categories in know. graph
- Cosine sim between user & candidate vectors

Editing Behavior	How indicates user interest
Number of times user edits article	Repeated editing implies greater investment and interest
Type of edit	Trivial edits (Table 2) are weaker signal of interest
Article's global edit activity and number of editors	Generally popular articles less discriminative of individual interest and personal relevance
Editing time span	Long-term interests are stronger than fleeting ones
Edit quality w.r.t. Info. Qual. metrics	Substantiveness and quality indicate concern in topic

Table 1

Edits Ignored:	Patterns Cleaned:
• Trivial (typo fixes, vandalism reverts) • Articles w/ under 100 non-stop words	• Stopwords, punct. • Article maintenance Wiki Markup removed • Stem, tokenize, lowercase

Table 2

Phase III: Ranking

- $sim(u.m)=\alpha*sim_{content}(u.m)+(1-\alpha)*sim_{category}(u.m)$ (α is weighting param determined experimentally)
- Ouputs highest scoring candidate as intended meaning

EVALUATION

Experiments

Data Collection and Preparation

- Sampled from Twitter (tweets), YouTube (video titles, descriptions), Flickr (photo tags, titles, descriptions)
- MTurk Categorization Masters used to select correct entity meaning. Averaged observed agreement across all coders and items = 0.866 and average Fleiss Kappa = 0.803
- String-matched usernames of posters to Wikipedia accounts & used Mturk to confirm accounts were same person
- Removed posters of non-English content & non-active users (less than 100 social Web utterances or 100 Wiki edits)

	# Usernames	Exist on Wikipedia	Matches are same person
Twitter	479	46.1%	47%
YouTube	454	19.6%	48%
Flickr	226	21.7%	71%

Results

- **Baselines**: randomly sorted candidates (RC), prior-freq (PF), RESLVE given a random Wikipedia user's interest model (RU), Wikipedia Miner (WM) and DBPedia Spotlight (DS)
- Best performance on YouTube texts (longest) due to content-based sim
- Outperforms on tweets (more personal in nature)
- Less effective on impersonal text (e.g., photo geotags) b/c high prior freq so standard methods suffice

	Flickr	Twitter	YouTube
RESLVE	0.63	**0.76**	**0.84**
RC	0.21	0.32	0.31
PF	0.74	0.69	0.66
RU	0.51	0.71	0.78
WM	**0.78**	0.58	0.80
DS	0.53	0.67	0.63

Precision (P@1)

Future Work

Identity bridging & Modeling interests

a. True neg. (no identity in know. base)
b. False neg. (same person, diff usernames)
c. False pos. (string match but diff people)

- Consider wider array of profile attributes to match (Helps with b, potentially c)
- For opposite case (a), collab. filtering to approximate user's own interests with contributions of social connections
- Use other kb besides Wikipedia
- Model user interest from additional kinds of participation (page visits, favoriting)
- Interest drift over time -- does timing of utterances + edits affect topic overlap

Computability

- Pruning strategy b/c millions of articles & hundreds of thousands of categories --> require new vector items to meet "relatedness" 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 (PIOF-GA-2009-252206).