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

Elizabeth L. Murnane
Cornell University
Information Science
elm236@cornell.edu

Bernhard Haslhofer
Cornell University
Information Science
bh392@cornell.edu

Carl Lagoze
University of Michigan
School of Information
clagoze@umich.edu

ABSTRACT

We address the Named Entity Disambiguation (NED) 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.

Categories and Subject Descriptors

H.3 [**Information storage and retrieval**]

Keywords

Entity Resolution; Social Web; Semantic Knowledge Graph; User Interest Modeling; Personalized IR

1. INTRODUCTION

The Named Entity Disambiguation (NED) problem arises when mentioned entities (people, places, concepts, etc.) have multiple candidate meanings. User utterances on social media like Twitter, YouTube, and Flickr are a particularly challenging scenario because texts are short, linguistic features are unreliable, and local lexical context is sparse.

Such content causes severe degradation in unique identification of entity concepts for traditional NED approaches, which are trained on long and formal texts. For instance, P@1 values for DBPedia Spotlight and Wikipedia Miner drop to below 40% when used for entity disambiguation on tweets; and the F1 score of the Stanford NER, which is trained on the CoNLL-03[1] news article dataset, drops from over 90% to just under 46% when applied to Twitter data.

Others' attempted solutions either rely on crowd-sourcing or focus on entity extraction without addressing the disambiguation problem. Also, nearly all have only been trained to handle tweets, making it unclear whether results are generalizable to personal utterances outside of Twitter.

[1] `http://www.cnts.ua.ac.be/conll2003`

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

We present a novel system called RESLVE (**R**esolving **E**ntity **S**ense by **L**e**V**eraging **E**dits) that augments traditional techniques with a personalized approach based on the idea that a user is more likely to mention an entity about a topic drawn from a domain of personal interest than from a domain of non-interest. Cognitive psychology affirms that a user possesses core interests in a consistent set of topics for which she tends to seek and produce content online, both in social media and in knowledge production communities.

Modeling interests from a user's other posts on the same social Web platform is not an effective approach because users do not generally produce a large enough volume of such posts and because the ambiguity in them is typically equally as high and offers no informative context [3]. Therefore we instead bridge a user's accounts between a social media platform and an external knowledge base in order to resolve the intended meaning of ambiguous entities encountered in the former by leveraging the power of structured information relevant to personal interests available in the latter. Here we use Wikipedia because it is well-established and domain-spanning, but our general model can be instantiated on any knowledge context that categorizes topics; DBPedia and Freebase are two additional high-coverage examples.

Our key contributions are: (i) a model to represent user interest with respect to a knowledge base, (ii) a ranking method that incorporates this interest model, and (iii) the results of an empirical evaluation of our system and an annotated dataset of disambiguated entities from Twitter, YouTube, and Flickr. The dataset and system implementation are available at `https://github.com/emurnane/RESLVE`.

2. APPROACH: THE RESLVE SYSTEM

RESLVE addresses the disambiguation problem by (i) connecting a user's social Web identity and Wikipedia editor identity, (ii) modeling that user's personal interests based on Wikipedia article edits, and (iii) ranking entity candidates by measuring how similar each candidate's associated topic is to the most salient topics in the user's interest model.

Given a short text a user posted on Twitter, YouTube, or Flickr, RESLVE first attempts to find a Wikipedia account belonging to the same person. Our current approach simply does string matching on usernames since prior research demonstrates feasibility [1]. RESLVE then models the user's topics of interest using the bridged Wikipedia account's editing-history metadata. The number of times a user edits an article is used as a signal of interest in a topic, and trivial edits like vandalism reversion or typo correction are ignored to avoid noise in the interest model [2].

During disambiguation, RESLVE performs pre-processing on the short texts (e.g., filtering URLs, normalizing or removing Twitter-specific tag characters, and bypassing Flickr machine tags) and then uses Wikipedia Miner and DBPedia Spotlight to extract an entity's candidate topic meanings, supplied in the form of Wikipedia URIs.

The representation of topics and categories is formalized as a directed bipartite graph consisting of topic nodes and category nodes with edges weighted based on the distance of concepts in the Wikipedia category graph. In order to rank entity candidates, RESLVE combines content-based ($sim_{content}$) and category-based ($sim_{category}$) measures of similarity to get an overall measure of relatedness between the interests of a user u and the candidate meaning m of an ambiguous entity. Specifically, we define $sim(u,m) = \alpha * sim_{content}(u,m) + (1-\alpha) * sim_{category}(u,m)$ where α is a weighting parameter determined experimentally.

To compute $sim_{content}$, a TF-IDF weighted term vector is built from the titles of the articles a user has edited, the title of the candidate's article, the tokenized words from the articles' pages, and the titles of the articles' category pages. To compute $sim_{category}$, vectors are built using articles' category IDs and the distance between a category and article topic in the knowledge graph. Cosine similarity between the user and candidate vectors is used for both similarities.

3. EXPERIMENTS

From three separate sites that host content exhibiting qualitative differences, we sampled several thousand utterances that can be characterized as short texts: Twitter (T) tweets, YouTube (Y) video titles and descriptions, and Flickr (F) photo tags, titles, and descriptions. 91% of these short texts contained one or more ambiguous entities; nearly two-thirds of all detected entities were ambiguous; and very few of the short texts contained only non-ambiguous entities.

We took the social Web usernames of the posters, string-matched them to Wikipedia accounts (T:46%, Y:19%, F:21% of usernames existed on Wikipedia too), and used Mechanical Turk to confirm any matches belonged to the same person (T:47%, Y:48%, F:71% of matches were confirmed as the same person). To build the user interest models, we then collected the ID, title, page content, and categories for every Wikipedia article each user made non-trivial edits on.

We extracted entities from the short texts using Wikipedia Miner and DBPedia Spotlight with disambiguation threshold parameters set to retrieve all potential candidate meanings. RESLVE ranked these candidates given the corresponding user's interest model. Three different Mechanical Turkers also labeled the candidates, leaving a gold-standard of 918 labeled agreed-upon ($\kappa = 0.80$) entity meanings.

4. RESULTS

We compare RESLVE's performance to alternative ranking methods: randomly sorted candidates (RC), prior-frequency (PF), RESLVE given a random Wikipedia user's interest model (RU), Wikipedia Miner (WM) and DBPedia Spotlight (DS). Table 1 reports P@1, the proportion of entities for which the top ranked candidate is the correct meaning.

RESLVE performs best on YouTube, the longest texts in our dataset, mainly because of content-based similarity. It also outperforms existing NER services on Twitter texts, which are generally more personal than those on YouTube or Flickr, showing that considering user interest for NED can be effective in highly user-centric domains. Along the same lines, our approach with a random user's interest model as input (RU) often performs better than other baselines since the external data allows additional topic overlap with candidate entities; but it is not as accurate as the personalized approach, showing that the user-specific data does help.

Table 1: Precision (P@1) of ranking methods

	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

Conversely, RESLVE is less effective on more impersonal text. Misrankings result from automated posts, and we see lower performance on Flickr data, where many entities refer to non-subjective topics (e.g., geographic places), which have high prior frequencies and can be ably resolved with traditional approaches. Lastly, RESLVE is also less effective with users who make fewer contributions to Wikipedia.

5. CONCLUSIONS

We addressed the Named Entity Disambiguation problem for short, highly ambiguous texts using a personalized approach that leverages external semantic knowledge about a user's key topics of interest. Our method does not depend on language-specific or local information, which is often hard to process or missing entirely in user-generated content. We reveal the advantages of our strategy over a variety of baseline and state-of-the-art methods, achieving gains especially when text contains content of a highly personal nature.

The shortcomings and relatively high costs of finding confirmed cross-platform identities limits the sample size of our current evaluation. Improving on this with more sophisticated account matching or collaborative filtering techniques is our top priority for future research.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1144153, and this work is also supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (PIOF-GA-2009-252206).

7. REFERENCES

[1] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *UMUAI*, 23:169–209, 2013.

[2] D. Cosley, D. Frankowski, L. G. Terveen, and J. Riedl. SuggestBot: using intelligent task routing to help people find work in wikipedia. *IUI*, pages 32–41, 2007.

[3] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *AND '10: Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, Oct. 2010.