YaLi: a Crowdsourcing Plug-In for NERD

Yafang Wang, Lili Jiang, Johannes Hoffart, Gerhard Weikum
Max-Planck-Institut für Informatik, Saarbrücken, Germany
{ywang,ljiang,jhoffart,weikum}@mpi-inf.mpg.de

ABSTRACT

We demonstrate the YaLi browser plug-in which discovers named entities in Web pages and provides background knowledge about them. The plug-in is implemented with two purposes. From a user perspective, it enriches the browsing experience with entities, helping users with their information needs. From the research perspective, we aim to improve the methods that are used for named entity recognition and disambiguation (NERD) by leveraging the plug-in as an implicit crowdsourcing platform. YaLi tracks the system's errors and the users' corrections, and also gathers implicit training data for improving NERD accuracy.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Crowdsourcing, Browser Plug-in, Named Entity Disambiguation, Named Entity Recognition

1. INTRODUCTION

When users browse news or social media on the Web, often names of entities (e.g., people, places, companies, movies, songs, etc.) catch their attention, and users would then like to obtain additional information on these entities. Some news sites satisfy this need to some extent, by providing links to Wikipedia, but this features is limited to very prominent entities or relies on manually curated links. The work presented here generates such links on the fly and fully automatically, whenever a user highlights a noun phrase in a Web page. The challenge that we face here is the ambiguity of names and phrases. For example, "Washington" most prominently refers to the capital of the USA, but it can have other meanings depending on context: the first American president, a film name, or the name of a state. To address this problem, we propose a browser plug-in that calls a server for named entity recognition and disambiguation (NERD) in a real-time manner and transparently for the user. In this way we achieve two goals:

1) From a user perspective, we enrich the browsing experience by providing background knowledge about entities (from Wikipedia or knowledge bases) on demand in a pop-up frame.
2) From a research perspective, we collect user feedbacks, as a form of implicit, user-transparent crowdsourcing, thus gathering statistics on name-entity pairs and implicit training data for improving NERD accuracy.

Crowdsourcing has become popular for many tasks such as image tagging, language translation, or recommendations. Platforms such as Amazon Mechanical Turk (`www.mturk.com`) and CrowdFlower (`crowdflower.com`) allow applications to submit micro-tasks to a worker community and perform the matchmaking between tasks and workers based on payment and profiles. Research projects such as [2, 5, 6] have leveraged crowdsourcing for tasks like entity co-reference resolution, image search, etc. However, users contribute only because they are paid and the topics of the input tasks are given to them by the application. In contrast, our approach adopts the rationale that users should voluntarily and transparently engage in NERD tasks, by embedding these tasks into the users' daily browsing activities. We achieve this by the YaLi plug-in. No extra burden is imposed on users, and YaLi works on arbitrary Web pages so that users can freely choose their own topics of interest.

Contributions. We present the YaLi system which transparently harnesses NERD methods to help users with richer interpretation of Web pages. YaLi gathers user feedback to obtain implicitly labeled data for training and further improvement of NERD performance. Unlike explicit crowdsourcing, YaLi does not impose any extra burdens on users and comes with the incentive of enhancing users' browsing experience and their personal knowledge.

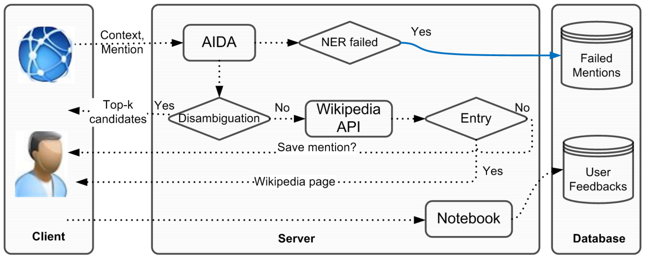
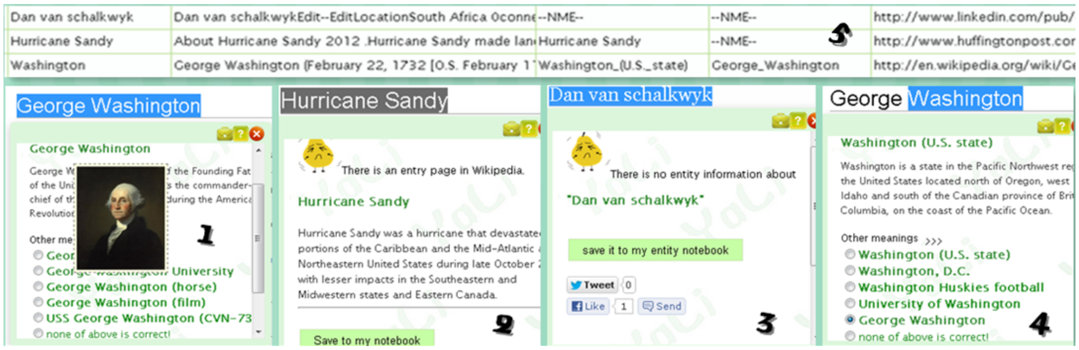


Figure 1: The System Architecture of YaLi

2. SYSTEM ARCHITECTURE

Figure 1 depicts the architecture of YaLi. YaLi integrates itself into the Google Chrome browser, listening for click or highlight events. When a user highlights a span of text, YaLi transfers both the highlighted text (mention) and its context to a NERD service, which returns the most likely entities for the highlighted text. The best entity candidates are shown in a pop-up frame next to the highlighted text, including short descriptions of the entities. Users can also save interesting entities in a *notebook* for later reference. To do this, they have to select the correct entity in the candidate list, thus implicitly providing training data.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.



Dan van schalkwyk	Dan van schalkwykEdit--EditLocationSouth Africa Oconne	--NME--	--NME--	http://www.linkedin.com/pub/
Hurricane Sandy	About Hurricane Sandy 2012 .Hurricane Sandy made lan	Hurricane Sandy	--NME--	http://www.huffingtonpost.cor
Washington	George Washington (February 22, 1732 [O.S. February 1	Washington_(U.S._state)	George_Washington	http://en.wikipedia.org/wiki/Ge

George Washington

George Washington

George W f the Founding Fat of the Uni s the commander- chief of th during the America Revolutio

Other me

○ Geor
○ George Washington University
○ George Washington (horse)
○ George Washington (film)
○ USS George Washington (CVN-73
○ none of above is correct!

Hurricane Sandy

There is an entry page in Wikipedia.

Hurricane Sandy

Hurricane Sandy was a hurricane that devastate portions of the Caribbean and the Mid-Atlantic Northeastern United States during late October with lesser impacts in the Southeastern and Midwestern states and Eastern Canada.

Save to my notebook

Dan van schalkwyk

There is no entity information about "Dan van schalkwyk"

save it to my entity notebook

George Washington

Washington (U.S. state)

Washington is a state in the Pacific Northwest re the United States located north of Oregon, west Idaho and south of the Canadian province of Bri Columbia, on the coast of the Pacific Ocean.

Other meanings >>>

○ Washington (U.S. state)
○ Washington, D.C.
○ Washington Huskies football
○ University of Washington
● George Washington
○ none of above is correct!

Figure 2: User Interface Screenshots of YaLi

YaLi uses AIDA [3] for entity disambiguation, which uses the Stanford NER Tagger [1] for mention recognition. AIDA uses the YAGO (`yago-knowledge.org`) knowledge base as a catalog of entities and their surface names, and returns Wikipedia identifiers for entities. When the highlighted phrase cannot be disambiguated by AIDA, YaLi tries to map it to Wikipedia directly by string matching. This is useful for new entities that are not registered in AIDA but already have a Wikipedia article. If there is no match in Wikipedia either, the background server for YaLi remembers the highlighted phrase as a potentially new entity or as a new name for an existing entity. As evidence is collected from several users, this information is used to extend the back-end knowledge base.

3. USE CASES

The user feedback and implicit training data that YaLi gathers can be used for strengthening NERD methods in a number of ways.

Improving NER. The text spans highlighted by users are good indicators of entity mentions and also reflect the evolving interests of user communities. The spans can be compared to the mentions found by the Stanford NER tagger [1] employed in AIDA. Recording the cases when the NER tagger failed is valuable data for future re-training and improvement of the NER tagger. It is also valuable input for extending the names dictionary that AIDA maintains. For example, phrases about entity roles such as "the first American president" (for George Washington) or "the tenor sax player" (in the context of a jazz band) can be easily picked up, and possibly added to the corresponding entities.

Improving statistics on name-entity pairs. Most NED methods make uses, among other assets, of a prominence-based prior probability that a certain noun phrase denotes a specific entity [4]. YaLi can improve this prior by observing users' click frequencies. An example is that "Washington" more frequently denotes Washington DC, and less frequently the US state of Washington. Moreover, as the entity associations in the user community change over time, it is crucial to estimate the prior in a dynamic manner rather than solely relying on static information such as link anchor texts in a Wikipedia snapshot. For example, "Sandy" most likely referred to the singer "Sandy Denny" until last winter, when "Hurricane Sandy" occurred. YaLi can track such trends and continuously improve the statistics for the NED prior.

Improving NED. YaLi increasingly collects labeled mention-entity pairs from the users' browsing activities. This data can be used for re-training the parameters of NED methods. A concrete example are the various hyper-parameters (coefficients) that AIDA uses for combining a prominence-based prior, context similarity measures, and the coherence (semantic relatedness) among entities.

Providing ground-truth for NERD evaluation. YaLi continuously gathers ground-truth information on mention-entity pairs from the users' clicks. This can serve as a growing basis for systematically evaluating and comparing different NERD methods and systems, going way beyond the currently used benchmarks (e.g., the CoNLL'03 corpus used in [3]).

4. DEMO FEATURES

Figure 2 shows several screenshots to illustrate the demo.

Entity information. Highlighting a span of text triggers a small icon to be displayed next to it. To avoid the situation that users highlight the text casually, the icon will disappear if users continue browsing without clicking it. Once the user clicks the icon, a pop-up frame is shown and three situations are handled as shown in the No. 1 - 3 screenshots respectively:

1) If the highlighted mention is disambiguated by AIDA, a list of entity candidates is returned. The best entity candidate is shown at the top, together with a short description and an image. For all candidates, the corresponding Wikipedia link is displayed and the first sentence from the Wikipedia page is provided as hovering text. If the user wants to save an entity for later reference, she can choose the candidate she believes is correct. She can also click on "none of the above is correct".
2) If the disambiguation of AIDA fails but there is a matched entity in Wikipedia, that information is shown and the user can save the entry if she deems it correct.
3) When both of the above cases fail, a new entity is recorded into the notebook by the user.

Entity notebook. As shown in the No. 5 screenshot, a personal notebook is made available by YaLi, to edit or search the recorded entities. The first row shows the feedback that the disambiguation failed; the second row shows the case where AIDA failed but a match was found in Wikipedia. The third row records the user feedback about the mention "Washington" (see the No. 4 screenshot). The generated best entity by AIDA is U.S. state but the user corrects it by choosing George Washington.

Acknowledgments *This work is supported by the 7th Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105.*

5. REFERENCES

[1] J. R. Finkel, T. Grenager, C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. *ACL*, 2005.

[2] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin. CrowdDB: Answering queries with crowdsourcing. *SIGMOD*, 2011.

[3] J. Hoffart, et al.. Robust disambiguation of named entities in text. *EMNLP*, 2011.

[4] V. I. Spitkovsky, A. X. Chang. A cross-lingual dictionary for English Wikipedia concepts. *LREC*, 2012.

[5] J. Wang, T. Kraska, M. J. Franklin, J. Feng. CrowdER: Crowdsourcing entity resolution. *VLDB*, 2012.

[6] T. Yan, V. Kumar, D. Ganesan. CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones. In *MobiSys*, 2010.