A Corpus of Wikipedia Discussions: Over the Years, with Topic, Power and Gender Labels

Vinodkumar Prabhakaran*, **Owen Rambow**‡
* Department of Computer Science, Stanford University, California, CA 94305, USA.
‡ Center for Computational Learning Systems, Columbia University, New York, NY 10027, USA.
vinod@cs.stanford.edu, rambow@ccls.columbia.edu

Abstract

In order to gain a deep understanding of how social context manifests in interactions, we need data that represents interactions from a large community of people over a long period of time, capturing different aspects of social context. In this paper, we present a large corpus of Wikipedia Talk page discussions that are collected from a broad range of topics, containing discussions that happened over a period of 15 years. The dataset contains 166,322 discussion threads, across 1236 articles/topics that span 15 different topic categories or domains. The dataset also captures whether the post is made by an registered user or not, and whether he/she was an administrator at the time of making the post. It also captures the Wikipedia age of editors in terms of number of months spent as an editor, as well as their gender. This corpus will be a valuable resource to investigate a variety of computational sociolinguistics research questions regarding online social interactions.

Keywords: computational sociolinguistics, Wikipedia, online interactions, power, gender, dialog

1. Introduction

Computational analysis of online social interactions has become an active field of research in recent years. Researchers have studied the linguistic and dialogic patterns of these interactions, the network structures they form, as well as how these patterns and structures relate to the social relations that exist between the interactants. These studies cover a wide range of genres such as social networking websites, email interactions, and online discussion forums. Within the genre of online discussion forums, the discussions happening in Wikipedia Talk pages (forums where Wikipedia editors discuss and debate the edits to the Wikipedia articles) have garnered special attention due to the fact that Wikipedia Talk is one of the very few online sources for task-oriented interactions.

In this paper, we present a large corpus of Wikipedia Talk page discussions that are collected from a broad range of topics, containing discussions that happened over a period of 15 years. The dataset contains 166,322 discussion threads, across 1236 articles/topics that span 15 different topic categories or domains. The dataset also captures whether the post is made by an registered user or not, and whether he/she was an administrator at the time of making the post. It also capture the Wikipedia age of editors in terms of number of months spent as an editor, as well as their gender. This corpus will be a valuable resource to investigate a variety of computational sociolinguistics research questions regarding online social interactions.

2. Related Work

There is a wide array of computational studies analyzing the dynamics of the collaborative editing process of building Wikipedia. One line of work focuses mainly on meta information such as history of edits, deletes, reverts, and dispute tags (e.g., (Vuong et al., 2008; Rad and Barbosa, 2012; Jurgens and Lu, 2012)), whereas others analyze the interaction dynamics exhibited by the editors in the Wikipedia Talk pages. At the level of modeling the language and structure of these interactions, researchers have attempted to assign dialog acts (Ferschke et al., 2012), to assign social acts (Bender et al., 2011), and to identify agreements, disagreements and disputes (Wang and Cardie, 2014b; Wang and Cardie, 2014a) as well as biases (Recasens et al., 2013) in these interactions. There is also work connecting the linguistic patterns to the social context of these interactions, such as power (Danescu-Niculescu-Mizil et al., 2012), influence and pursuit of power (Biran et al., 2012; Swayamdipta and Rambow, 2012; Strzalkowski et al., 2012; Nguyen et al., 2013), and social roles (Ferschke et al., 2015). Most of these studies use data collected specifically for the research questions they investigate, whereas we present a large general purpose corpus that captures broader aspects of interactions and their participants.

3. Corpus

In this section, we present our WikiTalk corpus, describe its construction process, and discuss the format in which the discussions are represented in it.

3.1. Data source: Discussion Threads

Our starting point is the list of controversial issues in Wikipedia that is collaboratively compiled by Wikipedia editors.[1] This list comprises of articles that are often re-edited in a circular manner, or are the focus of many editing disputes. Because of the controversial nature of these articles, they also tend to have relatively more and longer discussions in the corresponding talk pages, many of which have hundreds of archived pages of discussions. Another list of controversial topics we considered was from (Yasseri et al., 2014), in which they find the top ten

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

controversial articles in Wikipedia across ten different languages. In our preliminary effort, we used their top 10 list to extract the talk pages, resulting in a rather small number of threads (around 10,000), which prompted us to extract discussions from the Wikipedia-curated list of controversial articles, resulting in a much broader corpus around two orders of magnitude larger. In future, we plan to extend the corpus to other languages.

The Wikipedia-curated list of controversial articles assigns each article into one or more of 15 topic categories, which roughly corresponds to the following domains: *Politics and economics*, *History*, *Religion*, *Science, biology, and health*, *Sexuality*, *Entertainment*, *Environment*, *Law and order*, *Linguistics*, *Philosophy*, *Psychiatry*, *Technology*, *Media and culture*, *People*, and *Sports*. We preserve these category labels in the WikiTalk corpus, so that one could study if there are differences in the collaboration dynamics across different topic categories, and if so, why.

3.2. Data source: Editors' Gender and Wiki-age

We use the MediaWiki API to obtain information present in each editor's Wikipedia user account.[2] In particular, we extract the gender, registration date, and aggregate edit count of each editor. Only 12.3% of the registered editors in our corpus have revealed their gender in their user accounts. Nonetheless, given the size of our corpus, it still gives us a sizable collection of gender-labeled posts. The registration date helps us compute the "Wiki Age" of each editor, and the edit count helps measure how active they were.

3.3. Data source: Labeling Posts by Admins

In Wikipedia, some editors are promoted to the administrator status through an election process. Although the adminship is a user attribute, we assign the label at post level in order to distinguish between posts made by the editor before and after becoming an administrator. We use the Wikipedia page that keeps track of all the successful requests for adminships for editors over the years,[3] to determine when the editor was promoted to be an administrator. We obtained 2065 successful adminship requests and their corresponding dates. We verified each of the associated usernames to ensure their User pages still exist in Wikipedia. Out of these admins, 69 were since removed from Wikipedia due to various violations (e.g., maintaining multiple user accounts), but we kept them in the database so that we can capture their behavior as admins while they were still active in Wikipedia editing. Another four of the editors had since renamed their usernames (e.g., *Reedy_Boy* to *Reedy*). We kept both versions of their usernames in our records. Figure 1 shows the number of successful adminship request each month.

3.4. Discussion Thread Format

We use the Apache UIMA (Ferrucci and Lally, 2004) framework to design and build our dataset. The dataset will be released in both a simple XML format as well as the UIMA specific XMI format. The schema of the dataset is given in Table 1. The data types shown in the table are specific to the UIMA representation. In the simple XML representation, each data type is translated to appropriate XML tags.

Thread: uima.tcas.Annotation
— Uri: uima.cas.String
— SourceName: uima.cas.String
— ForumName: uima.cas.String
— GatherDate: uima.cas.String
— Posters: uima.cas.FSArray (Person)
— Posts: uima.cas.FSArray (Post)
— ThreadName: uima.cas.String
Person: uima.cas.TOP
— Name: uima.cas.String
— RegisteredUser: uima.cas.Boolean
— RegisteredDate: uima.cas.String
— Gender: uima.cas.String
— EditCount: uima.cas.Integer
— AdminDate: uima.cas.String
— AllPosts: uima.cas.FSArray (Post)
Post: uima.tcas.Annotation
— Author: Person
— Date: uima.cas.String
— UID: uima.cas.Integer
— ReferencePost: Post
— LinksMentioned: uima.cas.FSArray (WikiLink)
— isAuthorAdmin: uima.cas.Boolean
WikiLink: uima.tcas.Annotation
— Url: uima.cas.String

Table 1: UIMA type system (data structure) for WikiTalk discussion threads
Note: uima.cas.FSArray stands for an array of feature structures

The dataset is a collection of *Thread* objects, which are of the type Annotation (i.e., tied to a span of input text). The fields associated with each *Thread* object are: *Uri*, the unique resource identifier; *Sourcename*, the web source, in our case, `en.wikipedia.org`; *ForumName*, the specific article talk page; *GatherDate*, the date on which the thread was downloaded; *Posters*, an array of objects of the type *Person* (see below); *Posts*, an array of objects of the type *Post* (see below); and *ThreadName*, the title of the discussion. Each *Poster* has the following fields: *Name*, the name of the editor, which could also be their IP address if they were not logged in; *RegisteredUser*, a boolean feature denoting whether the editor is a registered editor; *Regis-*

[2] `https://www.mediawiki.org/wiki/API:Users`

[3] `https://en.wikipedia.org/wiki/Wikipedia:Successful_requests_for_adminship`

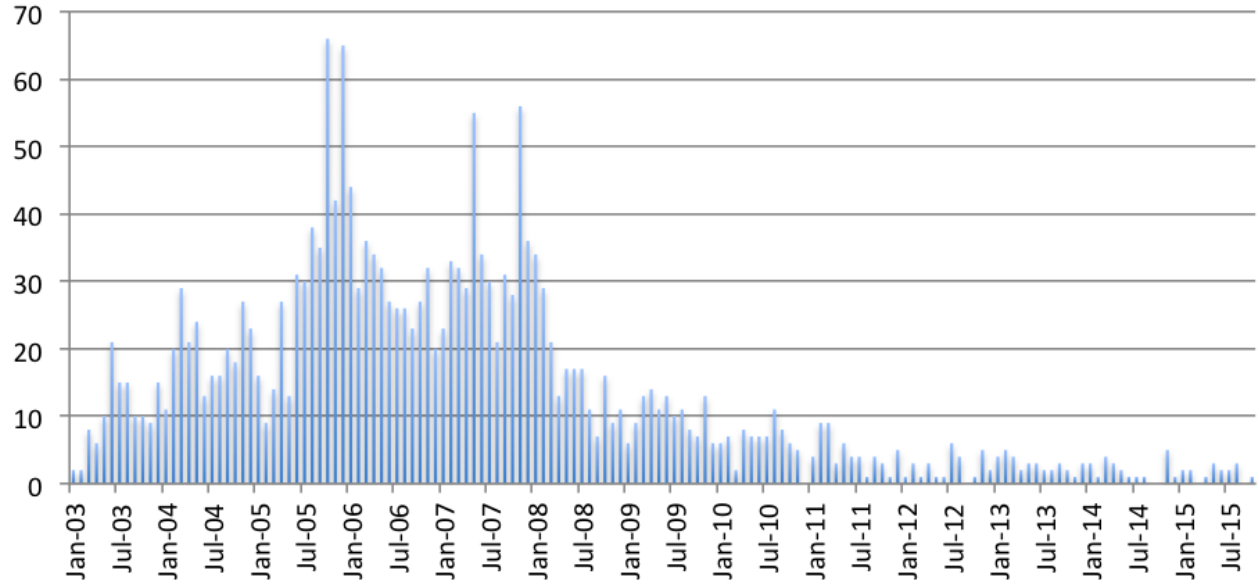

Figure 1: Number of successful adminship requests each month (Jan 2003 - Oct 2015)

teredDate, the date on which the editor registered; *Gender*, whether the editor's gender is male, female, or unknown; *EditCount*, aggregated number of edits made by the editor; *AdminDate*, the date on which the poster became an administer, if he is one; and *AllPosts*, an array of posts authored by the poster in the current thread. Each *Post* is of the type Annotation, and contains the following fields: *Author*, the object of type *Person* who authored the post; *Date*, date on which the post was authored; *UID*, a unique identifier for the post; *ReferencePost*, the post that this post is in reply to; *LinksMentioned*, an array of *WikiLink*s mentioned in the post; *isAuthorAdmin*, a boolean variable indicating whether the author of this post was an administer at the time of this post. Finally, *WikiLink* is an Annotation object that also stores the destination url.

4. Statistics

In this section, we present the various preliminary statistics we obtained on the corpus. Table 2 presents the aggregate counts of threads, posts and posters in the corpus. Table 3 presents the number of topics (i.e., Wikipedia articles) in each topic category, and the total number of discussion threads in each. It also shows the number of threads per topic in each topic category. For example, controversial articles in *Law and order* and *Entertainment* have relatively smaller number of discussions threads, whereas those in *Politics and economics* and *History* has around four times as many discussions per article, on average.

Of all the 906,671 posts in our corpus, 42,767 did not have a date assigned to it. This is probably from the period when Wikipedia had not enforced the format of editors' signatures when they were making posts. Figure 2 shows the number of posts made over the years. The period between 2005 and 2008 saw the peak of editor collaborations in our corpus. This is also reflected in the number of editors who became administrators (Figure 1).

Number of threads	166,322
Average number of posts per thread	5.45
Average number of participants per thread	2.84
Total number of posts	906,671
Number of posts by a registered user	834,067
Number of posts by an administrator	82,437
Totoal number of unique editors	104,982
Number of registered editors	59,451

Table 2: Aggregate statistics of WikiTalk corpus

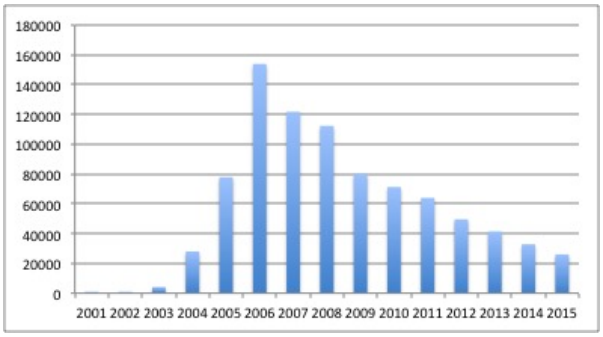

Figure 2: Number of posts per year

Figure 3 plots the percentage of editors with most posts against the percentage of posts they collectively authored. It shows that about 80% of the posts in our corpus is authored by 20% of the editors. In other words, our corpus represents a considerably large number of interactions between the same set of people. Out of all the registered

Topic category	No. of Threads	No. of Topics	No. of threads per topics
Entertainment	883	19	46.47
Environment	3154	50	63.08
History	19877	100	198.77
Law and order	600	14	42.86
Linguistics	2964	42	70.57
Media and culture	7745	68	113.90
People	60108	387	155.32
Philosophy	1114	6	185.67
Politics and economics	36374	177	205.50
Psychiatry	516	4	129.00
Religion	17897	99	180.78
Science, biology, and health	22719	122	186.22
Sexuality	5796	47	123.32
Sports	477	6	79.50
Technology	3686	30	122.87

Table 3: Discussion threads across topics and topic categories

posters, only 7,286 (i.e., 12.3%) have updated the gender field in their Wikipedia user accounts. Figure 4 presents the gender split within the set of posters whose gender was extracted — only 8.2% of them were female.

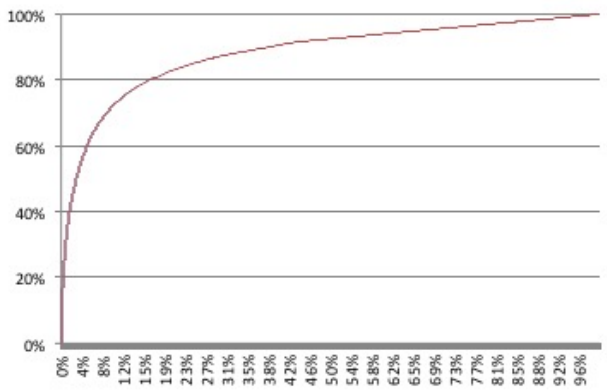

Figure 3: Percentage of posts (y-axis) by percentage of top posters (x-axis)

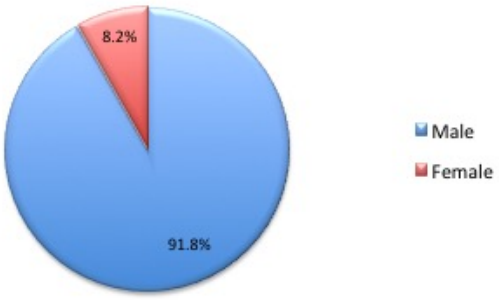

Figure 4: Number of Female vs. Male Wikipedia editors (based on self-declared gender information)

5. Highlights and Limitations

The WikiTalk corpus spans almost a decade of interactions between a large community of people who have come together with the common goal of enriching Wikipedia. The highlights of the corpus from a computational perspective are listed below:

- We capture two-level categorization of discussions: topic categories (e.g., science vs. history) and topics (e.g., American Revolution vs. Irish Potato Famine), enabling researchers to study how interaction dynamics differ across different kinds of topics.
- The discussions in the WikiTalk span over a period of 15 years (2001-2015), enabling researchers to study the temporal changes in behavioral patterns exhibited by the editors (e.g., do editors change over time), as well as in Wikipedia as a whole (e.g., are their macro-level effects of Wikipedia maturing as a platform that are reflected in these discussions).
- The corpus captures the gender of a subset of editors, which can be used to study gender differences in these interactions, and potentially understand why there is a skewed gender bias in the Wikipedia editorship.
- The corpus also captures the "Wiki Age" of each editor, enabling us to study the differences of interaction patterns exhibited by new vs. established editors.
- We also capture the aggregated edit count of each editor as a way to measure active they are.

- The corpus also captures whether the posters were admins at the time of making a post, enabling us to study the manifestations of power in these interactions.

Like the rest of Wikipedia, the list of controversial articles that this dataset is based off of also suffers from inaccuracies and omissions. For example, the article about *Noam Chomsky* was not listed as one of the controversial articles in the *People* category, even though the Wikipedia talk page for *Noam Chomsky* goes on to 15 archives of discussions. Despite such omissions, our WikiTalk is representative of the breadth of Wikipedia, and is the largest of its kind, to our knowledge.

6. Conclusion

This paper introduces WikiTalk corpus: a new large social interaction dataset of online task-oriented interactions, collected from the discussions on Wikipedia about making edits to controversial articles. The corpus contains 166,322 discussions that happened over 15 years, organized across 15 topic categories and 1,236 topics, and capture a range of attributes of the participants such as gender and power. Bringing together these multiple social aspects of interactions makes this a valuable resource to further computational sociolinguistics research on manifestations of social contexts in online interactions.

Acknowledgments

This paper is based upon work partly supported by the DARPA DEFT Program and by the NSF via award IIS-1159679. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We thank Rhea Goel for the work she did in the initial phase of this project, especially in building scripts to obtain some of the data. We thank several anonymous reviewers for their constructive feedback.

7. Bibliographical References

Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B., and Ostendorf, M. (2011). Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics.

Biran, O., Rosenthal, S., Andreas, J., McKeown, K., and Rambow, O. (2012). Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, Montréal, Canada, June. Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. (2012). Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, New York, NY, USA. ACM.

Ferrucci, D. and Lally, A. (2004). Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012). Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics.

Ferschke, O., Yang, D., and Rosé, C. P. (2015). A lightly supervised approach to role identification in wikipedia talk page discussions. In *Ninth International AAAI Conference on Web and Social Media*.

Jurgens, D. and Lu, T.-C. (2012). Temporal motifs reveal the dynamics of editor interactions in wikipedia. In *ICWSM*.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., and Wang, Y. (2013). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, pages 1–41.

Rad, H. S. and Barbosa, D. (2012). Identifying controversial articles in wikipedia: A comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 7. ACM.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August. Association for Computational Linguistics.

Strzalkowski, T., Shaikh, S., Liu, T., Broadwell, G. A., Stromer-Galley, J., Taylor, S., Boz, U., Ravishankar, V., and Ren, X. (2012). Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING*, pages 2535–2552, Mumbai, India, December. The COLING 2012 Organizing Committee.

Swayamdipta, S. and Rambow, O. (2012). The pursuit of power and its manifestation in written dialog. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 22–29. IEEE.

Vuong, B.-Q., Lim, E.-P., Sun, A., Le, M.-T., Lauw, H. W., and Chang, K. (2008). On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 171–182. ACM.

Wang, L. and Cardie, C. (2014a). Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *ACL 2014*, page 97.

Wang, L. and Cardie, C. (2014b). A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 693–699.

Yasseri, T., Spoerri, A., Graham, M., and Kertész, J. (2014). The most controversial topics in wikipedia: A multilingual and geographical analysis.