Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior

Keith Maki, Michael Miller Yoder, Yohan Jo and **Carolyn Penstein Rosé**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
{kmaki,yoder,yohanj,cprose}@cs.cmu.edu

Abstract

In this work we investigate how role-based behavior profiles of a Wikipedia editor, considered against the backdrop of roles taken up by other editors in discussions, predict the success of the editor at achieving an impact on the associated article. We first contribute a new public dataset including a task predicting the success of Wikipedia editors involved in discussion, measured by an operationalization of the lasting impact of their edits in the article. We then propose a probabilistic graphical model that advances earlier work inducing latent discussion roles using the light supervision of success in the negotiation task. We evaluate the performance of the model and interpret findings of roles and group configurations that lead to certain outcomes on Wikipedia.

1 Introduction

In this paper we explore the discussion strategies and configurations of conversational roles that allow Wikipedia editors to influence the content of articles. In so doing, we contribute both a new public dataset and proposed model that advance work towards induction of latent conversational roles using light supervision.

Online production communities like Wikipedia, an online encyclopedia which anyone can edit, have the potential to bring disparate perspectives together in producing a valuable public resource. Individual Wikipedia editors unavoidably carry their own perspectives; these voices can explicitly or subtly influence the jointly produced article content even when editors strive for neutrality[1].

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

This work explores the interaction between individual editors and the collaborative process that supervises the development of a Wikipedia article.

Wikipedia editors discuss article improvements, coordinate work and resolve disagreements on talk pages associated with each article (Ferschke, 2014). Pairing talk page discussions with simultaneous edits in shared content, we introduce a task predicting the success of a particular editor's article edits based on the corresponding discussion.

We propose a lightly supervised probabilistic graphical model of discussion roles and behaviors that offers advances over the prior discussion role modeling work of Yang et al. (2015), which employs a more restricted conceptualization of role taking. While the earlier model only allowed each role to be played by one editor, our extended model learns a distribution over roles for each editor. Furthermore, it can assign roles to an arbitrary number of editors rather than being restricted to a specific number.

This model allows the interpretation of configurations of roles that are conducive or detrimental to the success of individual editors. We find that the greatest success is achieved by detail-oriented editors working in cooperation with editors who play more abstract organizational roles.

2 Related Work

This work investigates influence in discussion as part of the collaborative editing process of Wikipedia, but achieving influence through discussion has also been studied in online environments other than Wikipedia. For example, other work in language technologies has studied the effectiveness of argumentative speech in changing others' minds (Tan et al., 2016) and revealing subgroups of users with similar attitudes (Hassan et al., 2012).

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 1026–1035,
Taipei, Taiwan, November 27 – December 1, 2017

Our work fits with research on editor behavior on Wikipedia, which is relatively well-studied on article pages and somewhat less studied on talk pages. Wikipedia has been a popular source of data for modeling social interaction and other issues of language behavior from multiple perspectives including collaboration (Ferschke et al., 2012), authority (Bender et al., 2011), influence (Bracewell et al., 2012; Swayamdipta and Rambow, 2012), and collegiality and adversity (Bracewell et al., 2012).

Much work analyzing behavior in Wikipedia has focused on types of edit behavior. Yang et al. (2016) use an LDA-based model to derive editor roles from edit behaviors. They then find correlations between certain editor roles and article quality improvements. Their approach differs from ours in that our model is supervised with an outcome measure and that we define editor roles based on talk page behavior.

Behavior in discussion can be characterized at multiple levels of granularity. Viégas et al. (2007) categorize talk page contributions into 11 classes, and find that the most common function of talk page behavior is to discuss edits to the corresponding article, but that requests for information, references to Wikipedia policies, and off-topic remarks are also commonly found. Bender et al. (2011) annotate authority claims and agreement in Wikipedia talk pages.

Above the level of individual contributions to discussion, the notion of a conversational role is relevant both for characterizing the rights and responsibilities an individual has within an interaction as well as the configuration of conversational behaviors the person is likely to engage in. Therefore, it is not surprising that prior work has revealed that the process of becoming a Wikipedia moderator is associated both with changes in language use and in the roles editors play on the talk pages (Danescu-Niculescu-Mizil et al., 2012).

Attempts have been made to understand roles Wikipedia editors play. Arazy et al. (2017) find self-organizing roles based on the edit behavior of thousands of editors. Editors frequently move in and out of those roles, but on the aggregate the proportions of these roles are relatively stable.

Our work is similar to that of Ferschke et al. (2015), who apply the role identification model of Yang et al. (2015) to Wikipedia talk page contributions. This model learns a predefined number of user roles, each of which is represented as weights on a set of user behaviors, and assigns the roles to the participants in each discussion. The roles are induced by rewarding latent role representations with high utility in selecting users whose behavior was highly predictive of the task outcome of article quality. We extend this work by predicting an outcome that is specific to one discussion participant, i.e. the editing success of a particular editor within an interaction. We also relax the constraint that every role must be assigned to a single participant and that each participant can take at most one role. Our model is thus more flexible in capturing more nuanced configurations of roles.

3 Data and Task

One of the contributions of this work is the creation of a new public dataset[2] and task for predicting the influence of editors in Wikipedia discussions. The dataset comprises 53,175 instances in which an editor interacts with one or more other editors in a talk page discussion and achieves a measured influence on the associated article page. In this section we detail the motivation for the conceptualization of the task as an influence prediction task, and the details for the construction of the dataset.

3.1 Task Conceptualization

Wikipedia talk pages are not stand-alone discussion forums. They are explicitly designed to support coordination in editing of their associated article pages. In order to extract task instances, we pair discussions with the record of concurrent edits to the associated article page.

Once a discussion has been paired with a sequence of edits, an assessment can be made for each editor who participated both in the discussion and in article edits of how successful that editor was in making changes to the article page. It is this assessment that forms the class value of our predictive task. In this study we explore negotiation strategies and role configurations that affect article editing; each data point in our task provides both discussion and an article edit success value for each editor involved.

[2]This dataset is available at `http://github.com/michaelmilleryoder/wikipedia-talk-scores`

3.2 Data Acquisition

To form our dataset, we extracted all versions (*revisions*) of English Wikipedia articles from 2004 to 2014 and removed much of the Mediawiki markup using the Java Wikipedia Library (JWPL) (Ferschke et al., 2011). The most recent revisions of talk pages corresponding to the articles were split into turns using paragraph boundaries and edit history. We grouped discussion posts under the same section headings as *discussion threads*.

We sampled 100,000 articles with talk page discussions and filtered to include discussion threads with 2 or more participants who made edits to the article page from 24 hours before the discussion began to 24 hours after the discussion ended. Discussion thread beginnings and endings are defined as the time of the first post and last post, respectively. Statistics on our discussion dataset can be seen in Table 1.

number of articles	7,211
number of discussion threads	21,108
number of editor-discussion pairs	53,175
average #editors/discussion	2.52

Table 1: Dataset statistics

3.3 Editor Success Scores

Editors frequently enter into talk page discussions to modify the article page in a particular way or challenge others' edits. We wish to quantify the success of editors on the article page as related to these goals on the talk page. In prior work, editor success has been measured with respect to the longevity of edits made to a page (Priedhorsky et al., 2007), and we take a similar approach. We define a success score y for each editor in a specific discussion. Intuitively, this measure is computed as the change in word frequency distribution associated with an editor's edits between the article revision prior to discussion and the article revision when the discussion ends. In particular, this score is the proportion of an editor's edits–words deleted and words added–that remain 1 day after the discussion ends. Note that this score only reflects changes in word frequencies and does not take word re-ordering into account.

Formally, we consider each edit $\mathbf{e}$ as a vector of word frequency changes, both positive (additions) and negative (deletions) for each word type, stopwords removed. For an example in English, an edit that changed one instance of *suggested* to *insinuated*, as well as adding *old* might be represented as {'suggested': -1, 'insinuated': +1, 'old': +1'}. For each edit $\mathbf{e}_i$, let vector $\mathbf{c}_i$ be the changes in word frequencies from that edit to the final revision after the discussion. This change vector represents how many tokens that an editor deleted were put back and how many tokens the editor added were afterward deleted. Let $|\mathbf{e}|$ be the number of tokens changed in that edit and $|\mathbf{c}|$ be the total word frequency changes (deletions if tokens of the word were added in the edit, or vice versa) in those specific word types from the edit to the final revision. The score y of a particular Wikipedia editor u in thread t across edits $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ made by u in t is:

$$y(u,t) = 1 - \frac{\sum_{i=1}^{n} |\mathbf{c}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|}$$

Each editor's score is the proportion of tokens they changed that remain changed, so $s \in [0, 1]$.

The goal of this editor score is to capture the "ground truth" of an editor's influence on the article page. To validate this editor success measure, we sampled 20 conversations, read through the corresponding article edits by those editors, and made sure our automated editor success scores were reasonable compared with the success that editors seemed to achieve.

In our experiments, we aim to predict this editor success measure calculated from article revisions with behaviors and interactions simultaneously occurring on the talk page. This assumes that talk page discussions in our data are related to the simultaneous article edits that those same editors are doing. To validate that editors who were editing the article while having a discussion on the talk page simultaneously were talking about those simultaneous article edits, and not something else, we manually went through 20 conversations and simultaneous edits. Nineteen out of the 20 conversations directly related to simultaneous edits, and the only one not specifically about simultaneous edits related to similar content on the article page.

4 Model

We present a model which attempts to learn both discussion behaviors of the target editor (editor we are predicting the success of) and roles of other discussion participants that influence the success of a particular editor.

4.1 Role Modeling Task

The task of role modeling as described is to identify latent patterns of behavior in discourse which explain some conversational outcome measure. The learned roles can then be intuitively interpreted to better understand the nature of the discourse and the interactions between the participants with respect to the chosen outcome measure.

4.2 Prior Approach: Role Identification Model (RIM)

A similar task was explored by (Ferschke et al., 2015) and (Yang et al., 2015), who represented role modeling as a bipartite matching problem between participants and roles. More specifically, RIM learns conversational roles from discussion behaviors, supervised by discussion outcome. A role is defined as a weight vector over discussion behaviors, where the weights represent the positive or negative contribution of the behaviors toward outcome measures.

However, this approach suffers from several simplifying assumptions which reduce its applicability to realistic conversation settings:

1. All roles are present in every conversation.
2. Each role is played by exactly one editor.
3. Each editor plays exactly zero or one roles.
4. All behaviors from editors with a role contribute to the outcome metric under that role.
5. No behaviors from editors without a role contribute to the outcome metric.

The proposed approach addresses these limitations by using a probabilistic graphical model that encodes a more appropriate hierarchical structure for the task.

4.3 Probabilistic Role Profiling Model (PRPM)

For modeling roles in discourse, we propose a generative model shown in Figure 1, whose generative story is shown in Figure 2.

4.3.1 Inference

Appropriate values for the parameters η, β, and τ may be inferred from data, and represent the settings with which the data is best explained (i.e. has the highest likelihood) under the generative story.

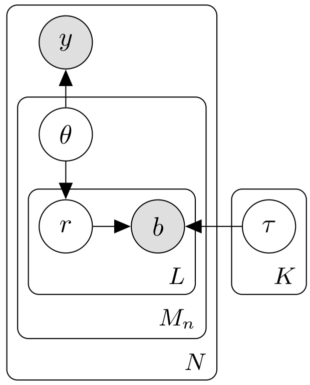

Figure 1: PRPM plate diagram relating for each conversation N the outcome measure y and each user M's L behaviors b.

- For each role $k \in \{1, \ldots, K\}$,
  - Draw behavior distribution $\tau_k \sim \text{Dir}(\alpha)$.
- For each conversation $n \in \{1, \ldots, N\}$,
  - For each user $m \in \{1, \ldots, M\}$,
    - Observe user participation z_{nm}.
  - For each user $m \in M_n$,
    where $M_n = \{m | z_{nm} = 1\}$,
    - Draw role distribution $\theta_{nm} \sim \text{Dir}(\gamma)$.
    - For each behavior $l \in \{1, \ldots, L\}$,
      - Draw role $r_{nml} \sim \text{Multi}(\theta_{nm})$.
      - Draw behavior $b_{nml} \sim \text{Multi}(\tau_{r_{nml}})$.
  - Draw outcome $y_n \sim \mathcal{N}(\mu_n, \sigma)$,
    where $\mu_n = \sum_m z_{nm}\theta_{nm} \cdot \beta$.

Figure 2: PRPM generative story

Computationally efficient methods for exact inference will not work for the proposed model due to the model structure, so approximate inference is used to estimate the parameter values.

We implement the model sampler using the JAGS framework (Plummer, 2003), which uses Gibbs sampling to generate dependent samples from the posterior distribution. These samples are used to obtain posterior mean estimates of the model parameters.

5 Features

5.1 Dialogue Act Features

We are interested in linguistic moves that characterize editors in conversation, and so we extract features that represent conversational acts. In particular, we extract dialogue act features from the model of Jo et al. (2017), an HMM-based unsupervised dialogue act identification method that has been found to usefully separate between content-related words that are relatively static across conversations and words more related to dialogue acts, which change over the course of discussion. These features were found to yield better performance with our model than unigrams with tf-idf selection.

The model of Jo et al. (2017) learns separate language models for dialogue acts (DA LMs) and topical content (content LMs), where each word can be generated from either type of language model. This structure helps the model identify content words that are consistent throughout a conversation and separate them out from language models for dialogue acts.

To identify dialogue acts on talk pages that may be related to conversational roles of interest, we ignore content-specific words by providing pre-trained content LMs trained using LDA over the content pages. Each conversation is provided with the topic distribution of the content page of the same article, and in the modified model, each word may come from a different content LM independently chosen from the provided distribution over content LMs.

5.2 Behavior Features

To be used in combination with roles, we extract general discussion features motivated by relevance in other work.

Along with a simple bag of words of each editor's talk contributions and the contributions of all other editors, we consider the following discussion features.

5.2.1 Position of the editor in a discussion.

- Number of editor turns
- Number of other editors' turns
- Whether the editor takes the first turn
- Whether the editor takes the last turn

5.2.2 Style characteristics.

Drawn from (Tan et al., 2016), these may reflect the style and state of editors.

- Number of definite/indefinite articles
- Number of singular/plural personal pronouns
- Examples: number of occurrences of "for example", "for instance", and "e.g."
- URLs: number of URLs that end with ".com", ".net", ".org", or ".edu"
- Questions: number of question marks that follow an alphabetic character

5.2.3 Authority claims.

Bender et al. (2011) define these authority claim categories annotate them in Wikipedia talk pages. For each word type in their annotated data, we calculated the pointwise mutual information for each category. In our data, we scored each sentence with the log sum of the word scores for each category.

The categories used are:

- Credentials: education or occupation
- Experiential: personal involvement
- Forum: policy or community norms
- External: outside authority, such as a book
- Social expectations: expected behavior of groups

5.2.4 Emotion expressed by editors.

For a simple measure of emotion, we use LIWC (Tausczik and Pennebaker, 2010).

- Counts of positive/negative emotion words

6 Experiments

We frame our task as a regression problem, predicting editor scores based on discussion behaviors of the target editor and the other editors. Our outcome measure is the editor success score of a single editor. Since there are multiple editors in a discussion, we have multiple instances per discussion.

We use root mean squared error (RMSE) between the true scores and the predicted scores as an evaluation metric. We hypothesize that in specifying our model with latent roles as mediators between the raw discussion data and the predictive task we can achieve a lower RMSE than from a baseline that takes only the behaviors into account, especially for conversations with a greater number of participants, for which there can be more interaction.

Furthermore, to the extent to which the proposed graphical model better captures a valid conceptualization of roles, we hypothesize that we can achieve a lower RMSE than the model of Yang et al. (2015). In this section we first specify the baselines used for comparison in our experiments, and then explain the testing process with our own model and experimental design.

6.1 Baselines

These two hypotheses suggests different baseline models. Our first hypothesis is that introducing a model with latent roles improves over simply using discussion features, and the second is that PRPM better captures interaction than the prior RIM model.

6.1.1 Linear Regression

The simplest baseline model allows us to evaluate the first hypothesis. This model assumes that the whole is not greater than its parts. In other words, it assumes that the sum total of positive impact the features can achieve on performance is just through their inclusion as separate features. For this baseline, we use a simple linear regression model. We bound the linear regression predictions to be between 0 and 1, the range of the editor scores. The full set of features in this model are included twice, once from the target editor in the discussion, and once from an aggregation across all non-target editors in the discussion.

6.1.2 RIM

We evaluate our model against RIM, introduced by Yang et al. (2015). RIM was originally applied to Wikipedia talk page discussions in Ferschke et al. (2015), who assigned a single success score to each page. In our work, for each discussion, we evaluate the success of each editor in each discussion thread separately. Since there is differential success between editors in the same interaction, the same interaction is associated with multiple different success measures. We handle this by slightly tweaking the original RIM model such that the first role is reserved exclusively for target editors, i.e., editors whose success measure is being evaluated. The other roles represent the roles of other editors in terms of their influence on the success of the target editor. Additionally, for conversations having fewer editors than the number of roles, we leave some of the roles unassigned by adding dummy editors whose behavior values are zero.

To predict the success measure of an editor for a test instance, RIM first assigns the learned roles to the editors. This process is identical to the training process, except that there is only the role assignment step without the weight adjustment step. Specifically, the first role is assigned to the target editor as in training, and the other roles are assigned according to the original model. Once the roles are assigned, the predicted score is simply the sum over roles of the inner product of a role's weight vector and the behavior vector of the editor who is assigned the role.

6.2 PRPM

To infer role distributions for each editor in a test instance conversation, we first fix the model parameters to the estimates learned during the training phase. Gibbs sampling is then used to infer the non-target users' role distributions θ_m and the conversation outcome measure y over the unseen data. The role distributions for each non-target editor are then averaged together and concatenated with the target editor role distribution. Finally, a linear regressor is used analogously to the above baseline to evaluate the predictive power of the PRPM roles in aggregating the information from editor behavior features.

6.3 Experimental Design

In order to evaluate our approach and model, we split our data into a training set of 60%, a development set of 20% to train regression weights on the roles learned from the training set, and a test set of 20%.

For the original and proposed role identification models, we manipulated the number of latent roles the learned models were allowed to include.

7 Results and Discussion

Results from baselines and PRPM are presented in Table 2. We do not include scores with unigram tf-idf counts as features, as this decreases the performance of all models. The pattern of results is consistent with the hypotheses, i.e., role information and our model's configuration improves performance over both baselines.

First, the relatively high RMSE values indicate the challenging nature of this task. Talk page discussion is only one factor in editor success, and undoubtedly much interaction between editors comes from edit behavior, past interactions between editors, and even the short edit comments that editors leave about their edits. We were not able to find a comprehensive study of the effect of Wikipedia talk pages on article pages, but links from discussion features to outcomes in collaborative editing are often tenuous (Wen et al., 2016).

Our model performs slightly better than the linear regression baseline, though it performs

Model	Setting	2	3	4	5+	All
LinReg	tgt editor	**0.286**	**0.302**	**0.287**	0.302	0.292
LinReg	all	0.287	**0.302**	0.289	0.301	0.292
RIM	K=2	0.316	0.317	0.308	0.342	0.318
RIM	K=3	0.307	0.320	0.310	0.337	0.314
RIM	K=4	0.307	0.314	0.311	0.327	0.311
RIM	K=5	0.309	0.315	0.308	0.321	0.312
PRPM	K=2	**0.286**	**0.302**	0.288	0.297	0.292
PRPM	K=3	**0.286**	**0.302**	0.288	**0.295**	**0.291**
PRPM	K=4	**0.286**	**0.302**	0.289	**0.295**	**0.291**
PRPM	K=5	**0.286**	**0.302**	0.288	**0.295**	**0.291**

Table 2: RMSE for baselines and models. Rows are model settings. Scores are reported for different numbers of participants, which are the columns headings. (LinReg: editor uses only the target editor's features, and all uses all participants' features. RIM and PRPM: K is the number of roles.)

substantially better than the previously proposed RIM model. One advantage of our role-based model above the linear regression baseline is clear when looking at conversations with more editors (columns in Table 2 denote the number of discussion participants in analyzed conversations). This points to the utility of using role information with larger groups, when roles are likely more relevant.

Another advantage of PRPM over the linear regression baseline is that it allows interpretation of both target editor strategies and group dynamics that characterize the success or failure of a target editor. Where linear regression allows only the characterization of behaviors that make individual editors successful, PRPM captures roles in interaction with other roles in group conversation. In this way, PRPM allows a more full interpretation of group interaction.

7.1 PRPM Role Analysis

Our best-performing model classified editors into 5 different roles. We identified the combinations of roles that are predictive of editor success (or failure).

To assess roles, we examined the text and discussion features of editors who scored highly, as well as considered the weights assigned to each feature for each role. The relative frequencies of each behavior for each role are shown in Figure 3. A characteristic example discussion post for each role is given in Table 3. Each role is named and described qualitatively below.

Moderator. This role primarily helps discussion flow without getting too involved, performing and summarizing the results of administrative tasks. High probability dialogue act features for this role include asking questions of other editors and discussing itemized content. The moderator role is less likely than other roles to have success as a target editor and has the lowest target editor success when paired with other editors playing the moderator role.

Architect: This role is predominantly focused on page hierarchy, with the bulk of its probability focused on the page_format dialogue act, which is relevant to discussions of adding new page sections, merging, archiving, and creating new pages. The architect role is moderately likely to have success as a target editor.

Policy Wonk: This role is an knowledgeable Wikipedia user, frequently mentioning source accountability, fair use or copyright policy for images. Dialogue act features which have high probability for the policy wonk include appealing to Wikipedia policy and discussing engagement with other users on user talk pages. The policy wonk role is moderately unlikely to have success as a target editor.

Wordsmith: This role is predominantly concerned with the naming, creation, and wording of pages. Dialogue act features which have high probability for the wordsmith include discussing the spelling, pronunciation, or translation of words and phrases, as well as discussing the (re-)naming of new or existing pages or sections. The wordsmith role is strongly correlated with target editor success, especially when combined with the moderator or architect.

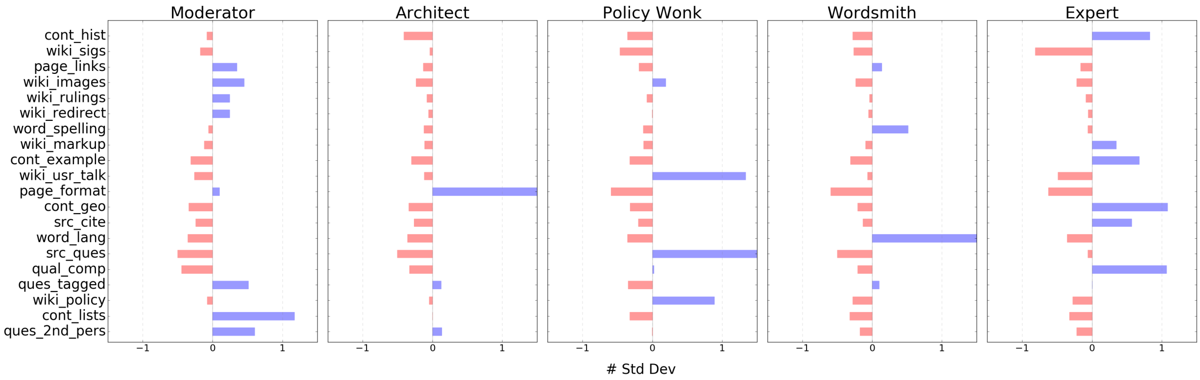

Figure 3: Behavior distributions for each role, expressed for each behavior as the number of standard deviations above the mean.

Role	Example post
Moderator	It was requested that this article be renamed but there was no consensus for it be moved.
Architect	I think a section in the article should be added about this.
Policy Wonk	The article needs more WP:RELIABLE sources.
Wordsmith	The name of the article should be ""Province of Toronto"" because that is the topic of the article.
Expert	There actually was no serious Entnazifizierung in East Germany.

Table 3: Examples of discussion posts from users in certain learned roles

Expert: This role is the most content-oriented role learned by our model. Dialogue act features which have high probability for the expert include making comparisons, discussing historical and geopolitical content, giving examples, and citing sources. The expert role is most strongly correlated with target editor success when combined with other users playing the expert role.

We find that the roles that lend themselves most strongly to target editor success (the Wordsmith and Expert) are more concrete edit-focused roles, while the roles associated with lower target editor success (the Moderator, Architect, and Policy Wonk) are more conceptual organizational roles. Note that it is not necessarily the case that editors that edit more frequently have higher scores. We find frequent editors across all roles.

Additionally, we find that configurations with multiple conceptual organizational roles lead to diminished outcomes for individual editors, suggesting that individual conceptual editors are unlikely to have their edits universally accepted. This could mean that talk page conversations that have multiple conceptual voices (which could be a measure of interesting discussion) are more likely to result in compromises or failure for a target editor. It is important to recognize that we are focusing on strategies and configurations of roles always in relation to the success of one editor; this editor score does not necessarily refer to a good, well-rounded discussion.

8 Conclusion and Future Work

The nature of collaboration on Wikipedia is still not fully understood, and we present a computational approach that models roles of talk page users with relation to success on article pages. We contribute both a new task with corresponding public dataset and a lightly-supervised graphical model for inducing role-based behavior profiles to predict the success of Wikipedia editors.

The proposed probabilistic graphical role model is unique in its structure of roles in relation to the outcome of one particular participant instead of group performance, and allows flexible mappings between roles and participants, assigning each participant a distribution over roles. The model we present retains one limitation of the RIM model, the assumption that editors in one conversation exist independently from those same editors in other conversations. Future work should address this.

Our model lends interpretability to combinations of talk page discussion roles. We find that detail-oriented roles are associated with success in combination with organizational roles, but that multiple participants taking organizational roles can lessen individual editing success.

We hope that this exploration into role-based discourse analysis will further enable systems to understand group interaction in text.

Acknowledgments

This work is supported in part by NSF GRFP Grant No. DGE1745016, NSF IIS STEM+C 1546393, NIH R01HL12263903, and by seedling funding from the Naval Research Lab.

References

Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. 2017. Turbulent Stability of Emergent Roles: The Dualistic Nature of Self-Organizing Knowledge Co-Production. *Information Systems Research*, page Forthcoming.

E.M. Bender, J.T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, (June):48–57.

David B Bracewell, Marc T Tomlinson, Mary Brunson, Jesse Plymale, Jiajun Bracewell, and Daniel Boerger. 2012. Annotation of Adversarial and Collegial Social Actions in Discourse. *6th Linguistic Annotation Workshop*, (July):184–192.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, pages 699–708, Lyon, France. ACM.

Oliver Ferschke. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Ph.D. thesis, Technische Universität, Darmstadt.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics EACL 2012*.

Oliver Ferschke, Diyi Yang, and Carolyn P. Rosé. 2015. A Lightly Supervised Approach to Role Identification in Wikipedia Talk Page Discussions. (2009):43–47.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 97–102, Portland, Oregon. Association for Computational Linguistics.

Ahmed Hassan, A Abu-Jbara, and D Radev. 2012. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, July, pages 59–70.

Yohan Jo, Michael Miller Yoder, Hyeju Jang, and Carolyn P Rosé. 2017. Modeling Dialogue Acts with Content Word Filtering and Speaker Preferences. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, September, pages 2169–2179.

Martyn Plummer. 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07*, page 259.

Swabha Swayamdipta and Owen Rambow. 2012. The pursuit of power and its manifestation in written dialog. *Proceedings - IEEE 6th International Conference on Semantic Computing, ICSC 2012*, pages 22–29.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. *Proceedings of WWW 2016*, pages 613–624.

Y. R. Tausczik and J. W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk before you type: coordination in Wikipedia. *40th Hawaii International Conference on System Sciences*, 1:1–10.

Miaomiao Wen, Keith Maki, Xu Wang, Steven P Dow, James Herbsleb, and Carolyn Rose. 2016. Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. *Proceedings of the 9th International Conference on Educational Data Mining*.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia. *Proc. ICWSM*, pages 446–455.

Diyi Yang, Miaomiao Wen, and Carolyn Rosé. 2015. Weakly Supervised Role Identification in Teamwork Interactions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1671–1680.