LEVERAGING CONTEXT DOCUMENTS FOR SOCIAL NATURAL LANGUAGE PROCESSING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ana Smith

May 2023

LEVERAGING CONTEXT DOCUMENTS FOR SOCIAL NATURAL LANGUAGE PROCESSING

Ana Smith, Ph.D.

Cornell University 2023

Aligning natural language processing models with stakeholder values is critical to managing the social biases that a model may learn from data. How to select and use context so that the resulting models are sensitive to these values is still an open question. I present three studies in which linguistic social context, or **context documents**, are leveraged for natural language processing tasks using **base documents** from social data (i.e., Wikipedia and conversation data). The context documents are produced by the members of the community as part of pre-existing processes. The context documents also enable an approach that trades the burden of precise annotation for noisier but value-sensitive information in those documents. I use techniques from semi-supervised learning and distant supervision to incorporate the information extracted from context documents into several inference tasks.

BIOGRAPHICAL SKETCH

Ana Smith is from Georgia and did her bachelors at the Georgia Institute of Technology. She left the sunshine and warmth of Georgia for the rolling hills and waterfalls of the beautiful New York Finger Lakes region. She likes knitting and long-form documents supported by reliable sources. Ana grew up in a family that loves to discuss everything from Star Wars to the stories and music of 20th century black musicians. This formative experience is the primary inspiration for Ana's work on discourse and dialogue.

To my parents, Edward and Monique, and my siblings, Heather, Sophia, Aviva, and Samuel, and one Brian C. Grant.

ACKNOWLEDGEMENTS

Much of this work (that involving Wikipedia data) was done with my advisor Lillian Lee. I am also grateful for the generous amounts of time others have spent with me engaged in insightful conversations. In particular, I thank Tianze Shi, Yiqing Hua, Krithika Vachali, Andy Ricci, Claire Liang, Danny Adams, Esin Durmus, Jonathan P. Chang, and the whole NLP seminar for their comments and feedback. Special thanks goes to my committee members Claire Cardie, who encouraged my questions, and Hakim Weatherspoon, who encouraged my foray into industry.

I have been fortunate to intern at a variety of places: Pindrop Security, IBM Research, Duolingo, and the New York Times. This has greatly influenced my outlook on natural language processing, its breadth of applications, and the interests of NLP practitioners in industry. Finally, I would like to express gratitude for the Cornell Colman Fellowship and Cornell Provost Diversity Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND ORGANIZATION

Modeling language using large-scale data has become the norm in natural language processing. While these models have outperformed previous models on benchmarks, there remain questions about how robust, how socially biased, or how truthful large language models are. Researchers have gone down a number of different avenues to amend mistakes these models make including the following: unit testing [230], data validation [83], post-training debiasing [31], and knowledge grounding [84]. The type of correction one might pursue depends on the task and stakeholder goals. Most approaches share this notion of accountability before correction, as bias is pervasive (and perhaps unavoidable) throughout model development [204].

That brings us to the first question of this dissertation: who are the stakeholders? One branch of natural language processing emphasizes benchmarks to measure state-of-the-art performance [35, 136]. Another branch is more interested in the broader implications of language use in social contexts [215, 148]. This dissertation addresses stakeholders may be NLP practitioners whose interests may fall somewhere in the middle of this spectrum. They may wish to adapt specific benchmark models for their own purposes. But their purposes may depend on their specific social context. This dissertation attempts to address the domain-specific needs and model expectations of these practitioners. My approach attempts to empower practitioners to introduce domain knowledge via **context documents** to their target **inference tasks** which canonically use a set of **base documents**. Throughout this work I pursue answers to the following primary research question.

Research Question: *How can* ***context documents*** *associated with* ***base documents*** *be used to inform natural language processing* **inference tasks***?*

This work focuses particularly on the interplay between dialogue and documents from a value sensitive design perspective [79, 78, 312]. Rather than attempting to train a model for immediately substituting human input, I focus on modeling human-human interactions to understand stakeholder values and goals. The following chapters will specifically discuss synchronous dialogue (i.e., meeting conversations), written asynchronous dialogue (e.g., Wikipedia talk pages), and discourse (i.e., Wikipedia articles).

In each cases, these conversations or articles serve as **base documents**. The **context document** may be a summary, IT reference material, or even a similar document on the same subject from a different linguistic and/or cultural context. It is through the **context document** that we expect the practitioners to introduce domain-specific information and stakeholder values.

The primary tasks in this work are battle outcome inference, article outcome prediction, and next-turn selection. Though they are all different styles of natural language processing tasks, they are all relatively simple in that the final label is binary, though the task setting may be one of distant supervision. This is to emphasize fitting models that inform us of the data generation process *in its natural context*. For example, we wish to understand *why* a given turn was generated more than we wish to know *how* to generate a valid or probable turn.

Not just any document may serve as context. For example, one would not expect a randomly selected email chain to be particularly informative in modeling a dialogue with an IT help desk agent. However, one would expect that a

IT troubleshooting document on a relevant topic could be. It is up to the practitioner to map the topology of stakeholder processes and find relevant **context documents**. That said, I propose three heuristics for establishing a relationship between the base document and context document that is expected to inform an NLP task:

- **Condition 1 (Semantic differences)**: The context documents contain new interpretations of information or meaning not present in the base document.
- **Condition 2 (Semantic overlap)**: The information in the context documents is relevant and can be aligned to some degree with information in the base document.
- **Condition 3 (Structural difference)**: The structure of the context document is distinct from that of the base document.

Not all conditions need to be met for a **context document** to be relevant to a **base document**. But one can use these conditions to quickly source **context documents** and form hypotheses regarding their usefulness for a particular task.

1.1 Intuition

How might the additional context affect the downstream task? The intuition is that the context provides a bridge between the base document used as input and the corresponding label. While it may not always be expedient to use the context as a bridge, the cost of checking the "bridge" for a shortcut and occasional success using it offset the cost of training the model on the task directly.

This type of bridging to create a shortcut for information flow is similar to the intuition behind highway networks [257].

1.2 Methods

Given that this work introduces a framework for applying machine learning problems to domain specific tasks, benchmarks and the models designed for them are of less help without modification. This section attempts to explain the data and task choices that led to these circumstances and the methods that were employed.

1.2.1 Data collection

The domains chosen for this work (primarily portions of Wikipedia, IT help desk chat logs, and meeting conversation corpora) were selected specifically because of three properties: public availability, document contexts, and rules and objectives that constrain discourse and conversation. Wikipedia in particular has been noted for the "ecology of [its] discussions" and "rich metadata" that enables multiple types of research investigations, particularly ones related to social acts [24]. Though there is no clear boundary between casual conversation and non-casual conversation, these settings constrain language in such a way that a conversational analyst may group them under the umbrella of *institutional talk* [104].

Why do we choose to focus on institutional talk? First, it is typically more restricted than "ordinary conversation" in that participants have defined roles or

status that specify the linguistic actions they may take. This lends itself to computational methods. Second, these roles and rules governing conversation and interaction are often specified in institutional documentation. This lends itself to the methods put forward in this work. Third, institutional talk, though highly regulated, is often found near centers of power. Understanding interactions in these settings is critical to creating accountable systems. Even in the cases when the project emphasizes non-conversation (e.g., World War Wikipedia), there is the undercurrent of institutional talk on the talk pages where edits are discussed and negotiated.

1.2.2 Inference over prediction

Though multilabel (e.g., multiple category membership) and multiclass (e.g., single category membership) inference tasks are discussed, most of the time the fundamental problem is binary. Did the British win the battle described or not? Was the article discussed promoted or not? Does this turn fit this context or not? In all these instances, the label may be a simple "yes" or "no", but it is also an *emergent property*[1] that cannot be reduced to any one part of the input. These tasks serve as fixed points to help us evaluate other questions not answered directly by the task. This approach is particularly amenable to *value-sensitive algorithm design*, a design paradigm that emphasizes the inclusion of the community whose data is being used and to whom algorithms are being applied.

For example, given that the British won the given battle, how likely will a model, specifically one trained on the English Wikipedia account of battles, pre-

[1] `https://plato.stanford.edu/entries/properties-emergent/`

dict a British win versus a model trained on the equivalent German Wikipedia account? So the goal is not to predict the outcome of a battle long past, but rather to understand to what extent documents generated by different communities allude to or imply certain conclusions through the presentation of select facts. Most of the experiments presented in this work follow a similar pattern wherein the goal is not the achievement of the downstream task; rather, the goal is a quantitative evaluation of the relationship between a base document and its context document in the setting of the given inference task.

1.2.3 Computational methods

The state of the art in natural language processing has bounded ahead multiple times in the course of this work. There was the jump from Bayesian inference to deep neural networks; deep neural networks expanded to recurrent neural networks; and feature engineering and weighting gave way to large pre-trained contextual language models. The contrastive approach presented in this work is meant to be model agnostic to an extent, but a few methods feature more prominently. The context document may be parsed using methods from information extraction to produce (subject, object, relation) tuples. Multitask learning may be used as a means of incorporating that contextual information.

1.3 Organization

The organization of this dissertation is intended to enumerate three primary approaches to incorporating context: parallel documents, references to guideline

and policy documents, and summary documents. The choice of context type is highly domain dependent. This work illustrates three possible approaches, but one is not limited to these choices.

- Chapter 2 goes over theories relevant to each and all projects. In particular, value sensitive design is discussed as the set of guiding principles for much of this work. Value sensitive design posits that technology should cater to the values and needs of stakeholders. Without this sensitivity, there is an increasing risk of designing technology with unwanted but difficult to manage social biases. Such technology impedes progress towards stakeholder goals rather than facilitating it as originally intended. Conceptual, empirical, and technological studies may be conducted under the umbrella of value sensitive design. The work presented may be categorized as empirical. For each project, a different set of values is identified and examined. Theories more specific to the tasks in these scenarios are discussed in more detail.
- Chapter 3 shows the effectiveness of using parallel documents as context when the base document is collaboratively edited text. This precedes the use of conversational data, with the intention of maintaining the same document type (i.e., Wikipedia article) and topic (i.e., World War I battle articles) when considering the benefits of augmenting a text classification task with context. Here the benefits of context come from differences in grammatical structure, information gaps, and combatant emphasis that may arise due to linguistic community alignment with battle participants.
- Chapter 4 introduces dialogue from Wikipedia's Featured Article Candidates discussion as the base documents and Wikipedia's policies and

guidelines as context. The caveat is that the context documents are present as implicit and explicit references to links in the primary document. This simplifies context representation and serves as a shorthand for a reviewer's "justification". This context is then used to investigate to what extent reviewers turn to institutional rules to justify their comments.

- Chapter 5 also employs dialogue as a base document, this time using synchronous conversations as found in the AMI and ICSI corpora. To learn about dialogue structure and flow, we emphasize this information. But here context is derived from post-dialogue summaries and pre-existing institutional documents. These documents are *not* referenced using links as the policies and guidelines of the Wikipedia discussions are conveniently referenced. This chapter focuses on methods for incorporating semi-structured context documents.
- Chapter 6 concludes the dissertation with a discussion of limitations and scope of applicability. There is also a brief discussion of ethical implications of this work, as well as a few thoughts on possible directions.

The reader is encouraged to note the **base document**, the **context document**, and the **inference task** in each chapter. The primary contribution of this work is a value sensitive NLP methodology that centers stakeholder values via **context documents**.

Why pursue a *value sensitive* approach, specifically? From a practical standpoint, this work also facilitates NLP inference tasks in domains with highly specialized or contextual knowledge that cannot be outsourced easily for annotation at scale. But more broadly, an approach that incorporates stakeholder values from the beginning is an approach that can account for model behav-

ior that undermines those values later. Chapter 2 will discuss value sensitive design in more detail along with other background information.

CHAPTER 2

THEORIES AND APPLICATIONS

This chapter covers relevant theories and recent work that provide the backdrop for the following chapters. Of linguistic morphology (language units), syntax (language structure), semantics (language meaning), and pragmatics (language use), this work is most related to the pragmatics of language [113], especially the social pragmatics of language use in institutional settings like Wikipedia. Conversation features prominently in the work done for this thesis. Though there is one chapter that focuses primarily on long-form documents, even then there are undercurrents of conversation described more fully in an appendix.

What the reader should take away from this chapter is a high-level understanding of value sensitive design and theories specific to conversation and writing. Although each chapter relies on theories specific to the interactions and data at hand, there are a few common themes uniting this work:

- **Not fully supervised learning.** A common and effective style of experiment in natural language processing requires labels well-aligned with instances, as par for the course in a fully supervised setting. The work presented here always includes a binary inference task, though this is more of a distantly supervised supervised setting than a fully supervised one. Also, semi-supervised learning is key to extracting useful information from the **context document**, be it through query-based summarization, policy invocation identification, or seeded tuple clustering. Semi-supervised and distantly supervised learning are particularly advantageous in situations where obtaining annotations might be prohibitive because of privacy reasons (e.g., internal IT help desk chats), because of the

high level of necessary expertise (e.g., familiarity with Wikipedia policy), or because of annotation labor burdens (e.g., labeling a single sentence in a long Wikipedia article as non-neutral).

- **Value sensitive algorithm design.** In this work, concerns about the unique properties of the stakeholder community eclipse generalizability. In short, specificity is traded for generalizability. The upside is the identification of a **context document** that is relevant and informative for a stakeholder's **inference task**. This is not to say that one cannot use a model on a dataset if it is trained on another. Rather, this is to say that care must be taken in identifying and justifying **context documents** that are specific to the domain. Tthe task is not simply input $\rightarrow$ output, it is $(\text{input}, \text{context}) \rightarrow \text{output}$.
- **Bias and accountability.** Related to value sensitive design, a common goal in my work is to account for *social biases* present in the data that might work against stakeholder values. The goal is *not* to correct for bias in a post hoc fashion or eliminate bias through dataset restructuring or other means. The goal is always simply to account for the *amount* of bias that may be present. Freedom from bias and user autonomy are recognized as values that should be upheld in almost any application of value sensitive design [79]. What to do about bias once it is identified is best left to the stakeholders themselves.

2.1 Writing and writing in other languages

Chapter 3 is concerned primarily with discourse in different languages regarding battles in World War I and World War II. These subjects involve a certain tension between language editions on Wikipedia that might be less prevalent in articles about dogs or "local" historical sites. Linguistic, conceptual, and citation differences may lead to variation between articles produced by different language editions [114]. It is easily observed in our data sample that battle articles available in one language are absent in other language editions. But suppose that a subset of language editions agree that a battle happened. Why would their accounts differ? Multiple social factors are the likely culprit.

2.1.1 Theories of social biases in natural language data

The variation in information available across Wikipedia language editions is indicative of social bias at the platform level [204]. Broadly, there is a distinct lack of data available for research, particularly for low-resource languages [210]. Of the datasets that are created, only a small portion are used and reused, even when another dataset might be better aligned with domain-specific problems [145]. This creates an ecosystem in natural language processing research that favors a few languages and the users of those languages [28]. To avoid the pitfalls of dataset reuse, this work attempts to cultivate suitable data.

Other social biases are also at play. Gender biases in data can lead to biased predictions with respect to gender in trained models [31, 210]. Racial biases in data can appear as stereotypes in pre-trained representations [180], and they can

negatively influence model predictions in hate speech detection [240]. Numerous studies have proposed approaches to reducing undesirable bias in natural language processing tasks [31, 87, 83, 242, 91]. Although eliminating all bias may not be an attainable or desirable goal when working with social data, accounting for bias is recommended [204].

Frameworks such as value sensitive algorithm design [78, 312] attempt to introduce accountability by advocating for the inclusion of stakeholders and their values in the design of algorithms that work on the data they produce. Though there are a "constellation" of features of a design process that may identify it as an instantiation of value sensitive design, five prominent features particularly relevant to algorithm design are as described as follows [312]:

1. **Identifying stakeholders, values, and priorities**
2. **Prototyping solutions to their problems**
3. Eliciting stakeholder feedback
4. Testing and refining models
5. Assessing reasons for failures

This thesis engages in the first two steps most prominently as this is the natural scope of natural language processing research.

Chapter 3 is about accounting for bias that appears when a Wikipedia's stated values diverge from its actual values, particularly that of Neutral Point of View across multiple language editions. There are a number of other values Wikipedia promotes across language editions. We choose to focus on the one because the juxtaposition of Wikipedia articles enables us to find segments of

text that might occur in only one language edition. The lack of *corroboration* by other language editions suggests a non-neutral point of view or undue weight being given to a topic. It is not an absolute judgement. The highlighting of such passages could be used to assist attempts to monitor problematic text in Wikipedia.

2.1.2 Neutral Point of View in multilingual Wikipedia

Neutral Point of View (NPOV) is a Wikipedia policy that says articles should maintain a neutral tone throughout the article, giving due weight to article subtopics as appropriate for the subject. Identifying neutral point of view violations in Wikipedia is a well-tread research path, but the data used is usually limited to English Wikipedia [226, 212]. There have been studies of conflict-related language within Wikipedia and Encyclopedia Britannica [239, 258], but these are limited to English. In contrast, the work documenting variation of across language editions is typically broader than geopolitical conflict and less connected to neutral point of view [16, 103]. Each individual language edition exhibits bias in the frequency and manner in which entities are mentioned in (parallel) articles about World War I and II battles. Chapter 3 introduces a method of accounting for possible sources of non-neutral point of view using other language versions (i.e., the **context documents**) of a given battle article (i.e., the **base document**). It is also shown in this chapter that these biases can affect battle outcome predictions (e.g., the **inference task**), which may be related to latent geopolitical factors [114].

2.2 Writing and talking

Chapter 4 maintains a focus on Wikipedia. In this chapter, the focus broadens to include values other than just Neutral Point of View and articles of a wider range of categories by using Featured Article Candidate (FAC) discussions. FAC is a Wikipedia peer review process which judges and provides constructive feedback on the quality of an article. Reviewers explicitly invoke Wikipedia's policies and guidelines and Featured Article Criteria (FACr) to justify their comments and decide whether to grant Featured status. This chapter focuses on how consistently this policy-conscious community cites policies in practice and possible biases in citing those policies.

2.2.1 Theories of writing iteration and writing feedback

Why does it matter whether a community cites its own policies when making judgements? It was found that destructive criticism undermines the efficacy of workers and trust in institutional structures [17]. To counteract these effects, feedback should be polite, prompt, specific, and actionable to be constructive[18]. This work assumes that the FAC process is a constructive criticism process in which policy references are used to make commentary more specific and actionable.

The constructive criticism of FAC reviews is also part of the revision process of the nominated articles. Iterative revision is essential in writing [76], and Wikipedia lends itself to understanding the dynamics of writing and feedback because of the excellent access to revision history, writing commentary, and doc-

umented values [24]. How criticism is administered may greatly affect the success of an article.

2.2.2 Applications to writing feedback models

The ultimate application of this work would be to improve writing assistants. Writing assistants typically provide feedback on writing affect, grammatical correctness, and genre fit. Wikipedia lends itself to training such writing assistants with the availability of revisions [43], metadata such as editor username [2], and public discussions about the article [120]. So why not simply approach this as a text generation task where the input is the article and the output is the review discussion? After all, a similar approach has been used on generating reviews for academic papers [286]. Even simpler, one could train a binary article outcome prediction model directly on accepted and rejected article submissions, framing it as an automated essay scoring task [265].

This is not impossible. In this circumstance, the **base document** is the article text, the **context document** is the discussion, and the **inference task** is article outcome prediction. But the main goal would be to generate sequences that align with FAC's values, not just fluent and likely text (in the case of generation) or a single score (in the case of probabilistic outcome prediction). Initial experiments in the article-based outcome prediction setting showed that the article's category was strongly correlated with outcome, even if it is known that category generally should not be the sole reason an article is accepted or rejected. Even if the reviewers are not biased by category, the reviewers themselves may be biased in how they invoke policies or justify their reviews. Chapter 4 lays

out an approach to modeling whether or not reviewers are consistent in their application of policy invocations as step towards more value sensitive writing assessment. Instead of the *article* being the **base document** and the *review discussion* being the **context document**, the *review discussion* is the **base document** and the *policies* are the **context documents**.

2.3 Talking and writing

Chapter 5 shifts to providing context for modeling of conversational data, rather than modeling conversational outcomes. Dialogue modeling has seen rapid advancement in casual and informative conversational situations [54]. While canonical dialogue modeling tasks such as the Alexa Prize [224] emphasize the development of a fully automated dialogue agent that can engage in lengthy chit chat, there is a subset of work still entirely focused on better modeling human-human conversation to understand humans and their interaction choices in conversation [282, 259, 231, 61, 80]. Chapter 5 focuses on human-human conversation to eventually inform dialogue agent applications via document grounding.

2.3.1 Theories of grounding dialogue in documents

Document grounding is an avenue of research that seeks to incorporate documents into dialogue modeling for improved turn generation, dialogue state tracking, and improved dialogue agent engagement [309, 72, 71]. The intuition is that dialogue models behave more predictably predictable behavior if they are

grounded in a pre-existing document. Theories of grounding in human-human dialogue [56] have guided the development of new multi-modal models that model language in the context of images [262]. Document grounding is similar but instead uses additional text as context.

2.3.2 Applications to conversation modeling

Experiments for modeling human-human conversation can take on a variety of forms, but a recognized experimental form involves predicting the next turn from a list of candidate turns from context turns [231, 283]. The context may be chosen specifically to reflect speaker roles [61], technical content [171], or other features of conversational context. Other experiment forms include conversation disentanglement [171] or ordering conversations [231] from shuffled turns. These experiments may be intended to probe the conversational structure rather than to improve a model's performance on a particular application. Most relevant is a set of experiments related to conversational initiative [282]. This work posits that there are dialogue behaviors related to information and planning in task-oriented conversation.

At the time, a dialogue model of help desk chat logs that more closely aligned with IT help desk procedures was considered more valuable. In this case, the **base document** is the dialogue, and **context documents** for IT conversations are available in the form of procedural documents. The **inference task** is next turn selection, where a model must choose one of eight turns to match to a given context. The work on conversational initiative inspired our approach which uses summarization to group utterances as dialogue manage-

ment (planning) utterances and/or procedural (information) utterances. Training the model to identify the type of utterance in addition to whether it fit the context improved next turn slection significantly.

CHAPTER 3

PARALLEL DOCUMENTS AS CONTEXT

Although this work emphasizes the use of dialogue transcripts as base documents, we do consider cases where the base document and context document may be non-conversational. Conversation is distinct in that the communication between the listener (or audience) and the speaker (or writer) only require that both are able to perceive and speak to one another in the same language. Discourse is a broader category encompassing conversation, speeches (or monologues), and written forms of communication. In this section apply our method to this broader category using Wikipedia articles as objects in discourse. This work was done jointly with Lillian Lee.

In this chapter, the **base document** is a Wikipedia article describing a World War I battle in a given language. The **context document** is also a Wikipedia article on the same World War I battle but written in another language. The **inference task** is labeling the entire article or parts of the article with the outcome of the battle, even if the outcome is not explicitly referenced in the article.

This chapter's experiments seek to answer questions about the composition of "facts" in the Wikipedia articles and how that composition may subtly influence the perceptions of battle outcomes even without alluding to those outcomes directly. The article corpus is derived from Wikipedia language editions linguistically associated with four of the major combatants in World War I (e.g., de.wikipedia.org and Germany would be associated). This inference task has parallels to tasks in work on stancetaking. Whereas traditional stance detection posits a triangular relationship between the subject, another subject, and an object, this work posits a relationship between a language community, another

language community, and the battle itself. The stance is the attitude of the first subject towards the object in alignment (or non-alignment) with the other subjects. Here, the attitude is measured by the extent to which the article or portion of the article implies a particular outcome of that battle, which have a positive or negative connotation for political reasons. It is for these same reasons that countries may report different casualty counts.

3.1 Using Wikipedia as a resource

As a resource Wikipedia is exceptional in that it contains extensive text and image documentation of a wide variety of topics. The contributions are comprised of volunteer edits and uploads. A conglomeration of Wiki Projects, administrator groups, and review and assessment processes make Wikipedia content more organized and consistent than might otherwise be expected from a freely edited encyclopedia. The NLP community has taken note of the consistency and breadth of Wikipedia knowledge across several different languages. The result is that Wikipedia is often used for training large language models and translation models.

The drawback in using Wikipedia as a resource is one that Wikipedia itself understands and notes in their policies and guidelines: Wikipedia is not a reliable resource. In this next chapter, we will demonstrate the variability in information quality of Wikipedia using Wikipedia itself: largely in the form of different language editions.

3.2 Task introduction

Wikipedia's expansive content and multiple language editions have made it an invaluable resource, particularly for the training of large language models and translation models in natural language processing. Less work has gone into quantifying the *differences* among language editions though. In particular, military conflicts, with their political implications and charged nature due to casualties, may be described in distinct ways by different language editions.

While community guidelines ensure some quality control and consistency across articles, Table 3.1 shows that in descriptions from German (DE), English (EN), French (FR), and Italian (IT) Wikipedia articles about the World War I battle at Verdun, there is still disagreement about whether the German objective was to "bleed" the French army. Instead of glossing over this difference, we aim to quantitatively measure it.

There are challenges to measuring these differences, though. Language editions may differ because of (1) linguistic differences in expression; (2) lack of information access, especially due to language barriers; and (3) an author's subjective preferences for sources.

There is work on identifying subjectivity in Wikipedia [227, 213]. But these supervised approaches, while successful, are limited by their need for explicit annotations. This work instead uses unsupervised methods to measure reporting tendencies of Wikipedia articles about battles in World Wars I and II from four language versions — German (DE), English (EN), French (FR), and Italian (IT).

DE	*Summary: Germany did* not *intend to "bleed" France*
	In contrast to subsequent representations by the Chief of Staff of the German Army, Erich von Falkenhayn , [3] the original intention of the attack was not to "bleed" the French army without spatial targets. With this assertion made in 1920, Falkenhayn tried to give the unsuccessful attack and the negative German myth of the "blood mill" an alleged meaning.
EN	*Summary: Germany* did *intend to inflict mass casualties on France*
	Falkenhayn wrote in his memoir that he sent an appreciation of the strategic situation to the Kaiser in December 1915, "...French General Staff would be compelled to throw in every man they have. If they do so the forces of France will bleed to death." The German strategy in 1916 was to inflict mass casualties on the French, a goal achieved against the Russians from 1914 to 1915, to weaken the French Army to the point of collapse.
FR	*Summary: Germany did* not *intend to "bleed" France*
	According to the version that Falkenhayn gives of his plan in his Memoirs after the war 15 , the goal is to engage in a battle at the loss ratio favorable to the German army, and therefore to discourage France to obtain the stop of the fights... Recent historical works, notably those of the German historian Holger Afflerbach, cast doubt on the version of Falkenhayn who claimed to want to "bleed dry" the French army.
IT	*Summary: Germany* did *intend to "bleed" France*
	... [I]n Verdun the purpose of the Falkenhayn offensive was to "bleed the French army to death drop by drop." In the plans of the German Chief of General Staff , the moral and propaganda importance of an attack on Verdun would have meant that all the French effort was poured into the defense of a stronghold considered to be of primary importance for France.

Table 3.1: Segments of different-language articles that provide contrasting accounts of a supposed German strategy to "bleed" France in the Battle of Verdun. (Google Translate was used for German (DE), French (FR), and Italian (IT); English (EN) is the original.)

We narrow our scope of analysis to national entities and their contexts, posing the following computationally-amenable question about the representation of such entities:

RQ1: Do language editions focus more on their associated combatants?

Here, our "focus" is limited to the combatant entity distributions. We measure how much they vary among articles from different language editions about the same event. Although an author's preferred writing language is not equivalent to an author's nationality, language editions are known to reflect geopolitics in images [103], cultural topics [268], and community participation [248]. Therefore, we hypothesize the following:

> *H1: Languages associated with particular combatants will emphasize that combatant more than others.*

While entity distributions alone facilitate comparisons, the context in which those entities appear may also contribute to subtle differences in perspective. We incorporate context by using (subject, relation, object) *tuples* filtered for the geopolitical entities used above, asking the second question:

> *RQ2: Is there a language edition that is most corroborated by every other language edition?*

We operationalize our understanding of "corroborate" by measuring how tuples from different language editions are grouped (or not) when clustered in a semi-supervised scheme. Differences between language editions are expected, but the gap between languages associated with Germany and Italy and the languages associated with the United States, Britain and France might be expected to have more overlap in their accounts of battles, given wartime alliances:

> *H2: The German (*DE*) and Italian (*IT*) language editions of Wikipedia will overlap more in facts than the English (*EN*) and French (*FR*) language edi-*

tions. EN, *being the sizable language edition, might be expected to be most corroborated by every other language edition.*

Contributions. In a quantitative analysis of entity distributions related to language-country association, we find a language edition associated with a particular country does tend to emphasize that country more than other language editions do (H1 validated). An additional contribution is an approach to reveal conflicting or corroborating tuples by using a downstream diagnostic *battle outcome* inference task. The results of this task indicate that several factors discussed in more detail below affect representation quality.

We demonstrate that though there are more instances of standalone tuples, clustering facts based on similarity across language editions and averaging their representation yields a representation that is more linearly correlated with battle outcome. The results of our outcome prediction task suggest that different language editions provide complementary information and models benefit from using all language versions rather than just one.

In this work, we describe multilingual Wikipedia articles. But there are parallels to news articles from different broadcasters and countries that produce documents covering the same events. A possible extension is to identify domain-specific indicators of differences in opinion in scenarios where a prebuilt lexicon is not immediately available, but multiple perspectives are. Another possible application of this methodology is as a diagnostic tool to identify potential sources of bias in Wikipedia datasets.

3.3 Related work

There is prior work extracting relations between and events involving geopolitical entities from text [203, 49, 175, 98, 258] as well as a survey [115]. We focus on managing and comparing descriptions of such relations across different language communities [181, 241]. (Of course, multilingual parallel and comparable corpora have been a mainstay of machine translation since its beginnings.)

3.3.1 Multilingual Wikipedia

Our research is primarily a study of the relationship between a Wikipedia article's content and its relationship to the corresponding article in another language edition. Other work compares Wikipedia language editions from the perspective of the geography associated with an article [162], the imagery of articles [103, 220], and perspectives of colingual groups on common topics [268]. Our project is closely aligned in spirit with other analyses of how wars are described across different language communities in Wikipedia [85, 310, 38, 151]

3.3.2 Wikipedia and information extraction

Wikipedia has served various purposes outside of its obvious role as an open-edited, free encyclopedia. After years of studies on Wikipedia's information quality [260, 9, 152], more recent work focuses more on leveraging it to answer questions [51], populate knowledge bases [105, 295], and generate summary tables [168]. The former line of work more directly questions the quality of

	WWI			WWII		
Rank	Lang	No.	$\not\approx$ En	Lang	No.	$\not\approx$ En
1	EN	606	—	EN	2958	—
2	FR	373	23%	FR	1358	10%
3	IT	327	7%	IT	888	10%
4	DE	225	16%	DE	788	5%

Table 3.2: Number of retrieved distinct identifiers for Wikipedia articles listed under the WWI or WWII battle categories. (Recall that we restricted attention to Latin-script languages for countries with the most casualties.) "$\not\approx$ En" columns: % of articles in that language without an English-language equivalent.

Wikipedia content. We do not assess the quality of information directly, but rather assess the prevalence of certain pieces of information. Our work is similar to the latter line of work in that we attempt to simplify Wikipedia content to a few phrases for analysis. Our work differs from prior work in that it does not extract snippets from a larger body of text to fill in answers. Rather, it compares snippets from multiple language editions.

3.4 Data collection

Our corpus of battle descriptions is collected from multiple language editions of Wikipedia. To identify potential candidate articles for download, we take the names of articles listed under the English language categories "Battles of World War I" and "Battles of World War II"[1] and corresponding categories in other language editions (e.g., Battaglie_della_prima_guerra_mondiale) identified by interlanguage Wikilinks for German, French, and Italian. These languages were selected because they are the primary languages employing Latin script used

[1] `https://en.wikipedia.org/wiki/Category:Battles_of_World_War_I`, `https://en.wikipedia.org/wiki/Category:Battles_of_World_War_II`

by combatant countries with the largest recorded casualties.[2]

Different language editions do encompass different sets of articles, with some articles available in only a subset of data. So even if the communities are comprised of the same individuals with the same aims in every language edition, the output is non-equivalent for all languages. In total, our dataset has 765 distinct WWI battles and 3430 distinct WWII battles.

See Table 3.2 for the distribution across language editions.

WWI				WWII			
DE	EN	FR	IT	DE	EN	FR	IT
german	german	german	german	german	german	german	german
british	british	british	british	japanese	japanese	japanese	japanese
french	french	french	french	british	british	british	british
russian	germans	germans	germans	soviet	italian	french	italian
army	russian	france	russian	american	soviet	germans	soviet
germans	ottoman	russian	italian	us	french	soviet	germans
division (german empire	france	ottoman	germany	allied	germans	american	french
italian	russians	germany	france	germans	allied	us	american
german empire	belgian	italian	russians	italian	us	germany	us
austria	germany	armenian	russia	french	american	france	allied
france	allied	austro	austrian	army	germany	allied	germany
hungary	armenian	austria	austro	americans	france	italian	italy
reserve division	italian	hungary	ottoman	france	japan	japan	france
army corps	russia	turkish	turkish	infantry division	axis	americans	americans
russians	austria	somme	army	category	united states	united states	japan
category	belgium	allied	belgian	polish	dutch	soviets	soviets
austrian	uk	russians	allied	dutch	chinese	polish	polish
reserve corps	hungary	ottomans	italy	germany	the united states	dutch	axis
weblinks	romanian	italy	meuse	japan	italy	italy	army
germany	turkish	serbian	belgium	the red army	italians	category	chinese

Table 3.3: Top 20 most frequent non-pronoun, non-individual-human terms per language (after →Spanish→English translation) automatically tagged as geopolitical named entities in our World War I (left) and World War II (right) corpora.

After the names of battle articles in different languages are collected, they are disambiguated by linking them to a Wikidata item identifier known as a QID, obtained by querying the WikiData API. QIDs link articles across different language editions, and we use the reduced set of QIDs to identify all language editions of each article. Though there is still a bias for articles grouped under the

[2]We repeat that the restriction to Latin script is an attempt to minimize processing differences between languages. Moving to a larger set of more diverse languages is a direction for future work.

"Battles of World War I" and "Battles of World War II" categories, this additional step reduces the likelihood that we are collecting data only visible from English Wikipedia. For example, the DE version of Wikipedia tends to have fewer articles, possibly because they conceptualize warfare differently (e.g., campaigns instead of actions).

Full-text content is then downloaded from Wikipedia using the PetScan interface[3]. The next section discusses how this data is further cleaned and partitioned.

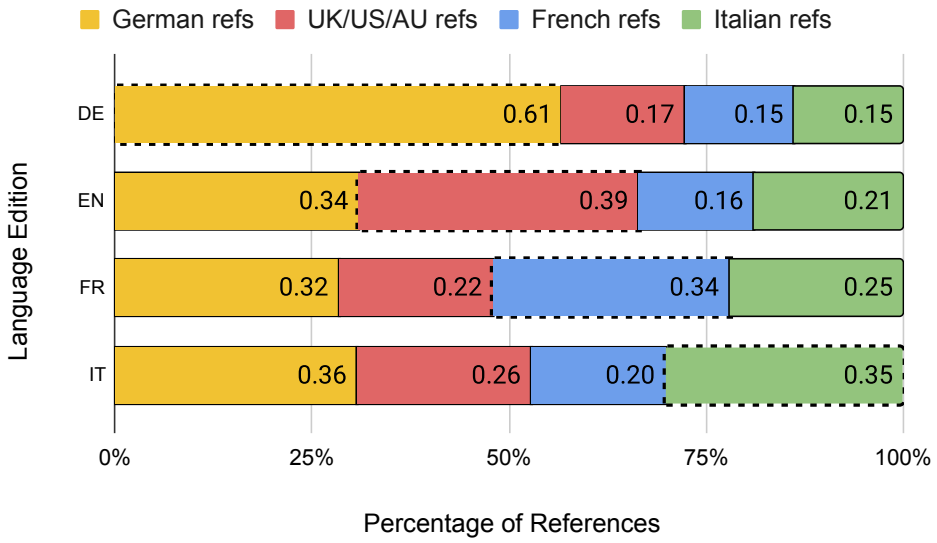
3.5 Associated languages and entities

Initially, all battle articles listed under the battle categories in each of the four languages are collected. But because this work compares language editions, only the intersection of the four language editions is used. This results in 131 articles for World War I and 414 articles for World War II. This subset is then processed as described below.

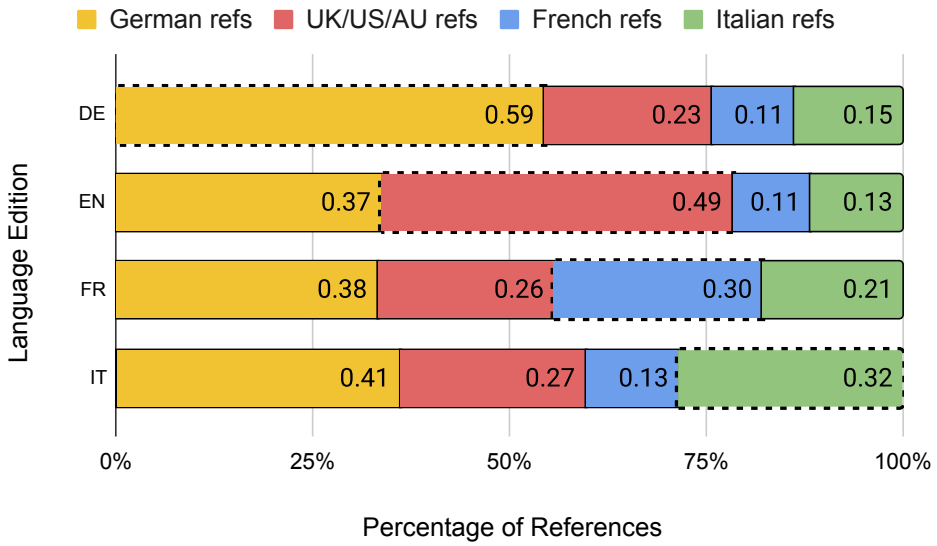
3.5.1 Processing articles

Our approach requires the use of open-domain information extraction, which has until recently been largely restricted to English, so all articles must be translated to English for our method. To compensate for translation noise in our non-English articles, all articles (including English articles) are translated to "new",

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



(a) WWI



(b) WWII

Figure 3.1: Comparison of reference proportions, as discussed in §3.5.3. All language editions mention Germany significantly more than other language editions (see yellow bars), but they all mention their associated combatant more than other language editions mention that combatant (see diagonal of dashed boxes).

fifth language, Spanish, and then to English using Google Translate.[4] Importantly, we subject English to potential translation errors to avoid privileging it as the only language under consideration that would not have undergone translation otherwise.

Translation. Using different language revisions enables us to probe differences across groups of editors employing the same language. On the other hand, although the original language of the articles is expected to give the most accurate distinctions, we choose to work with translated versions of the articles so that we can apply a standardized set of NLP tools developed for English. To avoid privileging the originally-English articles, all language versions are first translated to a *new* language (Spanish, given that there are many high-quality machine translation models between Spanish and other languages) before being then retranslated into English via Google Translate.

Text cleaning The collected articles are in xml format, complete with internal links, templates, and other artifacts. The article text is sentence- and word-tokenized; then, internal links are simplified to the alt-text only, and we remove templates including infoboxes, inline references, and text starting from the section headers "References" and "See also".

Named entity tagging. Though there are two major sides in these wars, there are numerous combatants. We use a named entity tagger to identify geopolitical entities and persons. Manual inspection of the entities in the context of the

[4]We employed only European languages to stay within a family of relatively related languages; future work can be more ambitious about language choices.

article is used to identify ties to a single political entity. Alliances of entities (e.g., Allied Powers) are considered separately.

Associating entities with nations/alliances. Recall our first research question is related to the combatant distributions across language editions.

One difficulty is associating a particular entity with a combatant nation, due to issues with granularity and type of reference: American entities may be referred to as *the United States* (nation name), *Eisenhower* (leadership), *333rd Field Artillery Battalion* (military unit), or *they* (pronoun). To address this issue, a list of nations, leaders grouped by nation, and military units by nation are collected for each article from English Wikipedia categories and pages. (We exclude pronouns and entities not clearly identified by nationality as existing coreference tools did not prove reliable enough on our data.) Though this does not encompass all entities mentioned in our corpus, it does capture prominent entities.

3.5.2 Entity-count statistics

In total, there are 88,317 entities in our WWI corpus and 274,713 entities in our WWII corpus. Table 3.3 lists the most common non-pronoun non-person grammatical subjects in our World War I and World War II data.

The most prominent national entity across all language editions by far is Germany. This is to be expected given that in both wars Germany was engaged with combatants on both the Eastern Front and the Western Front, whereas most other combatants only appear on one Front. In the World War I corpus, the British are the second most common national entity subject. In the World War II

corpus, the Japanese are the second most prominent national entity. Not shown here is a list of PERSON entities. The most frequent persons listed in those tables are, surprisingly, *battle* in WWI and, unsurprisingly, *Hitler* in WWII. Our tags do contain noise. The word *battle* should *not* be tagged as a person, but it was tagged so across all language editions. The spaCy [108] en_core_web_sm model was used to obtain named entity and part-of-speech information.

3.5.3 Associated-language test (RQ1)

In the introduction, we hypothesized that languages associated with a combatant country would reference that combatant as a subject more than any other combatant. To evaluate our hypothesis, we compare the relative proportion of counts per article of *self-references* (i.e., references to a nation by its associated language) to *other-references* (i.e., references to a nation by other languages). Each other-reference is normalized by the number of other languages (i.e., 3) for a more balanced comparison to other self-references. Though doing so reduces statistical power, instances are grouped by war for better analysis.

Figure 3.1 is a stacked barplot of the self-reference and other-reference proportions in our dataset. To test significance between populations, we use the Mann-Whitney U test implementation in scipy [278], as our population sizes differ between the "self" country reference group and the "other" countries reference group and are non-normally distributed. When using a Bonferroni correction of 2 on a p-value threshold of 0.01 since a test was run for each war, our p-values for both WWI (5.46e-6) and WWII (8.61e-4) are significant at <0.005. Though the data are not normally distributed, the self-reference distribution

suggests that our hypothesis H1 is supported (i.e., languages associated with particular nations are more likely to mention those nations than ones that are not).

A breakdown of references by language edition and country reveals more nuance, with self-references highlighted by the dashed borders. The significance of the above test may be attributed in part to DE's many self-references and other language editions' many other-references to DE. This is likely because Germany's engagement on both Eastern and Western fronts made it a more common reference overall. That said, for every language version, the proportion of self-references is greater than references to that country in other language editions. This indicates there is indeed a tendency to emphasize the countries commonly associated with these languages. We consider H1 validated.

3.6 Tuple Clusters

While the entities alone indicate a preference for language editions to reference their associated countries more, the context in which they occur may aid our understanding of why these differences in distribution occur. We hypothesized that overlap among languages may be more likely between English and French accounts and German and Italian accounts than any combination of the two. But overlap alone says little about why accounts may differ.

We simplify article text to (subject, object, relation) tuples. Solely as a means to validate the quality of representation, a domain-specific outcome inference task is used. The intuition is that a better representation should enable a linear classifier to learn a correlation between outcome and text, among other proper-

ties.

3.6.1 Extracting tuples and clustering

Tuple extraction. Once all articles are translated, (subject, relation, object) tuples are extracted with the Stanford NLP Toolkit's OpenIE implementation [6]. This system was chosen instead of a neural approach to limit the possibility that information is hallucinated or generated that was not in the original text (such problems are known to occur in neural models such as Imojie [147]).

One problem is that essentially redundant tuples may be considered distinct. Consider the following tuples:

1. EN: ('sides', 'suffered casualties with', 'numbers of soldiers succumbing to freezing')
2. EN: ('sides', 'suffered casualties with', '**large** numbers of soldiers succumbing to freezing')

The only difference between (1) and (2) is the adjective "large" in the object. To address this problem, we group tuples by subject and relation per article section (e.g., == *Aftermath* ==) and take only the tuple within each group with the longest object (in tokens). No subject should be a substring of another subject, and no relation should be a substring of another relation. Hence, tuple (2) would be retained and (1) discarded.

Tuple representation. Following prior work [150], averaged word embeddings are used to represent text content. As a baseline, we compare this against

a 1- to 3-gram bag-of-words.

We begin with a basic representation of tuple t that doesn't distinguish between subject, object, and verb (relation) status:

$$v_{sro} = \frac{1}{|t|} \sum_{w \in t} \text{emb}(w) \tag{3.1}$$

where emb() is a mapping of w to a pretrained vector. This reflects our naive hypothesis that treating an entity (e.g., France) as an object is not distinct from treating it as the subject. We also compare a pre-trained embedding (GLoVe [216]) and an embedding trained on our corpus (using fasttext) only to assess the extent to which the context of World War conflict influences a model. Though GLoVe is trained on more data, the nature of conflict may contravene typical associative assumptions and domain-specific words (especially entities) may be dropped. Both vectors are of dimension 100. This dimension was chosen because previous studies suggest that dimensions on the order of 100 are relatively similar in performance but better than those with dimensions on the order of 10 [234]. In the case of GLoVe, a random vector was assigned to out-of-vocabulary words. The fasttext embeddings were trained using a character n-gram of maximum size 3 and a learning rate of 0.05. These embeddings are trained over the combined corpus (both WWI and WWII). Words appearing in fewer than 0.1% of tuples are excluded to manage the number of features and prevent overfitting.

The first representation neglects the structure denoted by the tuple. But this may be harmful in cases where distinguishing the subject and the object tuple matters (e.g., (France, defeated, Germany) is distinct from (Germany, defeated,

1st lang	Tuples contributed to cluster
DE	('German armed forces', 'lost will', 'resist')
DE	('German positions', 'against Army is', 'United Kingdom')
DE	('British troops', 'Only announced', 'their victory at Battle of Havrincourt')
DE	('German forces', 'lost will', 'resist')
EN	('Germans', 'could consolidate', 'their positions')
EN	('American forces', 'face', 'difficult task')
EN	('Germans', 'encouraged', 'Allies')
EN	('Germans', 'were', 'weakening')
FR	('German divisions', '6 at', 'least')
FR	('German army', 'withdraw until', 'November 11 1918')
IT	('advance', 'would', 'would also backed by 300 machine guns')

Table 3.4: An example multilingual (s, r, o) cluster obtained from articles on the 1918 Battle of Havrincourt. The component tuples, while from four distinct languages, generally correspond to the "tuple" that the Germans were unable to hold their position against British troops.

France)). To address this, a 300-dimensional representation is concatenated to v_{sro}. The mean vector for each word in the subject (s), relation (r), and object (o) is calculated as above and concatenated as follows:

$$v^{(t)} = [v_s; v_r; v_o] \tag{3.2}$$

Though the structure of $v^{(t)}$ ensures that the word *France* as an object is distinct from *France* as a subject, similar tuples may be written in the passive voice in one language and not another. To combat the issue of word order, $v^{(t)}$ is concatenated to v_{sro} to form the second feature vector used:

$$v^{(t)}_{final} = [v_s; v_r; v_o; v_{sro}] \tag{3.3}$$

Clustering tuples into tuples. The ultimate goal is to group similar tuples from different language versions in such a way that we minimize the size of the clusters — so that the included tuples should be more similar — while maximizing heterogeneity of within-cluster source languages, that is, the number of source languages represented in the cluster. To address both limits, we implement a hierarchical K-means clustering algorithm with thresholds for cluster sizes. Euclidean distance is used to measure (dis)similarity among instances. Clusters are recursively split until they contain fewer than 16 instances. Table 3.4 shows an example cluster.

Because word embeddings may associate words by type (e.g., tuples with *Germany* and *France* as subjects appear in the same cluster), an additional one-hot vector is prepended to $v^{(t)}_{final}$ to split tuple clusters along country lines when clustering.

$$v^{(t)}_{cluster} = [a_{\text{de}}; a_{\text{en}}; a_{\text{fr}}; a_{\text{it}}; v^{(t)}_{final}] \tag{3.4}$$

Here, $a_{\text{<language>}}$ is 1 if the associated language occurs in the subject of the tuple, otherwise 0. A single cluster can be represented by the mean of all $v^{(t)}_{cluster}$ tuple representations in the cluster. It is this mean vector that is used in the following experiments.

3.6.2 Validating representation quality

To assess the quality of the proposed representations, we use the outcome of the battle as a target to evaluate the extent these representations implicitly attribute advantages to (or minimize disadvantages of) combatants. For this task, the input is a tuple representation and the output is the *outcome* (e.g., 0 if Germans

feature	WWI tuples			WWI clusters			WWII tuples			WWII clusters		
	F1	recall	prec	F1	recall	prec	F1	recall	prec	F1	recall	prec
majority	0.372	0.500	0.297	0.286	0.500	0.201	0.378	0.500	0.304	0.317	0.500	0.232
#words	0.372	0.500	0.297	0.375	0.500	0.299	0.378	0.500	0.304	0.349	0.500	0.268
#tuples	0.372	0.500	0.297	0.375	0.500	0.299	0.378	0.500	0.304	0.349	0.500	0.268
bow_{sro}	0.467	0.523	0.560	**0.609**	**0.610**	**0.635**	0.502	0.545	0.617	0.573	0.586	0.608
bow_{final}	**0.475**	0.520	0.545	0.604	0.605	0.622	0.508	0.547	0.611	0.583	0.591	0.607
v_{sro} (G)	0.392	0.506	**0.616**	0.536	0.555	0.585	0.468	0.533	0.636	0.602	0.616	0.650
v_{final} (G)	0.431	0.516	0.581	0.531	0.539	0.548	0.512	0.553	0.638	0.606	0.621	0.660
v_{sro} (F)	0.435	0.521	**0.616**	0.562	0.573	0.604	0.537	0.569	0.658	0.633	0.642	0.675
v_{final} (F)	0.468	**0.533**	0.613	0.602	0.602	0.616	**0.567**	**0.586**	**0.663**	**0.662**	**0.669**	**0.706**

Table 3.5: Battle outcome inference results using several representations. (F) denotes the use of fasttext vectors, while (G) denotes GLoVe. On the left side of each table are the results obtained when using individual tuples as instances. On the right side are the results obtained when using the mean of a cluster's tuple representations as instances.

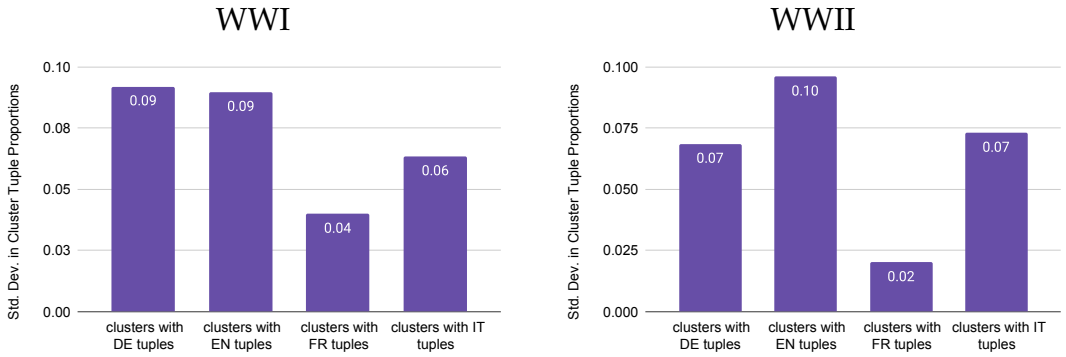


Figure 3.2: Bar chart showing the standard deviation in the proportion of language tuples in a subset of clusters defined by the presence of at least one tuple of a particular language. The cluster subsets defined by the presence of FR tuples in both WWI and WWII tend to have a balanced mix of tuples from DE, EN, and IT.

won, otherwise 1). Not every tuple is expected to directly correspond to the outcome, but any tuple that does should benefit from a better representation as indicated by an increase in model precision. In our experiments, we employ 3-fold cross-validation; for each fold, we fit a logistic regression model using the scikit-learn implementation [214]. The regularization parameter C is tuned over the range [0.01, 0.1, 0.5, 1.0, 3.0]. The results of evaluating the model on a

held-out test set are shown in Table 3.5.

Results. The bag-of-words (*bow*) representation presents a competitive baseline, particularly for WWI, as do the smaller $v_s ro$ representations. The WWII corpus benefits from the word embedding representation across the board, though. (Bear in mind that it is approximately 4 times larger than the WWI corpus.) Additionally, averaging the tuple representations per cluster yields even better outcome inference results — for example, on WWII using fasttext, F1 goes from .567 for unclustered to .662 for clustered — likely because of the larger context on which it draws in comparison to a single tuple.

Though there are fewer instances, using clusters is more advantageous in outcome inference than using individual tuples suggesting that the context derived from grouping similar tuples is useful for corroborating outcomes. Part of this effect may be due to complementary information from different language editions. Using $v^{(t)}_{cluster}$, we turn to our second research question regarding the overlap between language editions with clusters.

3.6.3 Measuring corroboration via cluster composition (RQ2)

To measure language heterogeneity in the tuples, each language (l_1) is paired with every other unique language (l_2) counted in the cluster. The count of occurrences of l_2 is then divided by the total number of tuples for that language. Correcting in this way, rather than using simple overlap, is intended to reduce the effects of population size (e.g., there being more EN tuples than FR tuples means the former are more likely to end up in any cluster by chance). This in

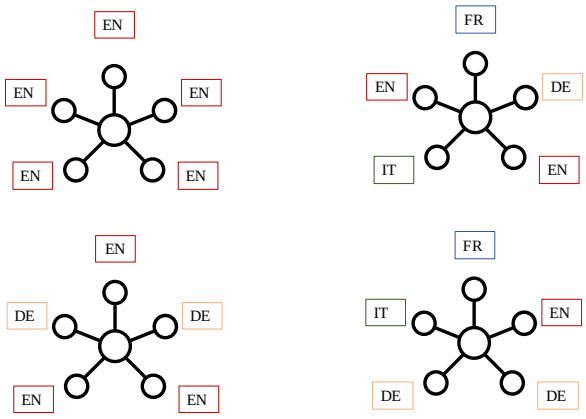


Figure 3.3: Tuple clusters may have large variation in composition (e.g., EN tuples) or little variation (e.g., FR tuples).

turn helps us to better assess semantic (dis)agreement among language editions.

Figure 3.2 shows that the tuples with FR language tend to co-exist in a balanced manner with tuples from other languages in both the WWI and WWII data; this is true even though FR has the fewest tuples of all the language editions. One possible explanation may be that though the French language version contains fewer tuples, each tuple tends to be corroborated by other language versions. See Table 3.6 for the total number of tuples. In contrast, EN tends to be the most variable in its proportions. Though FR clusters include EN tuples in a similar proportion to all other tuples, EN includes a much smaller proportion of FR tuples. These results partially contradict our hypothesis that the overlap would be greatest between FR and EN and between DE and IT. We consider H2 as not validated.

	WWI		WWII	
Lang	Tuples	Clusters	Tuples	Clusters
DE	181,456	78.5%	526,290	77.5%
EN	184,795	81.0%	504,085	69.8%
FR	107,879	69.6%	376,489	69.8%
IT	133,041	69.5%	532,223	76.3%

Table 3.6: Counts and cluster coverage of tuples extracted from the World War I and World War II corpora using the Stanford OpenIE system. The "Clusters" columns indicate the proportion of clusters in which the languages appear. Despite FR having the least number of shared clusters, it is still the least biased language edition of the four under these circumstances.

3.7 Conclusion of main experiments

In this work, we introduced a methodology for identifying information upon which language editions agree and disagree by applying open-domain information extraction and unsupervised learning to English translations of articles. Our results indicate that (1) language editions tend to mention their associated country more than other language editions mention the same country and (2) the FR language edition align with other language editions' accounts more than the reverse. Result (1) confirms other work on geopolitical tendencies of multilingual Wikipedia, though FR had the least self-emphasis. Result (2) was more surprising; it implies that FR Wikipedia may have a more limited though balanced account than other language editions *even though it is one of the smallest language editions*. The EN edition, though sizable, clearly has a large portion of tuples not corroborated by any other language edition. More qualitative analysis is needed to confirm what is in the uncorroborated clusters.

Limitations There are limitations to using machine translation for historical analysis. The noise introduced by the translation process may have still left

lingering problems. For example, the EN edition may still be the best translated language, possibly leading its tuples to clump together simply because other language edition tuples have too much noise. Additionally, good translation may be to expensive to scale to all Wikipedia articles. Information is lost in other ways as well. To avoid issues regarding nuance, articles are reduced to a set of simple (subject, relation, object) tuples. The vector representations used were also evaluated on downstream tasks before use in our second experiment. A more nuanced approach may be to include full sentences or equally sized token n-grams without any word removal.

Future work There are several possible directions for future work. Regarding tasks, it may be of interest to NLP practitioners to understand the impact the information imbalances have on downstream tasks such as translation. For the language communities themselves, it may be useful to be aware of the gaps in the accounts they are writing. To make this more useful for them, an important step would be to expand to other languages; our analysis is limited to four languages. Future work should include articles from languages correlated with combatants on the Western and Pacific front.

Ultimately, more conclusive results will require a better model of the community dynamics and citation practices of editors, especially over time as well as more qualitative analysis of the differences between language editions. We aim to continue this work with the hope it encourages interest and advances in the overlap of computational, historical, and cultural analysis.

3.8 Conclusion

Because of the way article actions and talk page interactions are publicly recorded via revision histories, Wikipedia lends itself to a number of natural language processing tasks. The similarity between Wikipedia articles in distinct languages has been known and well-studied, particularly for the purpose of training translation models. The differences between language editions has received less attention, and the attention it has received has largely been related to cultural studies. This work attempts to highlight and quantify the bias each language edition has towards associated entities. The DE, EN, FR, and IT language editions mention German-speaking, English-speaking, French-speaking, and Italian-speaking entities more, respectively. When it comes to incorporating more contextual information, the FR version (though smaller) overlaps with ever other language edition more than other language editions overlap with it. This suggests that it is the most neutral point of view in our corpus of battle articles.

Contributions. This chapter introduces a method of contrasting articles to identify aspects of language community attitudes towards World War I battles. There is no one state-of-the-art model, as research in the area is limited by the lack of multilingual datasets. The gap is only widened by the lack of multilingual open domain information extraction models at the time. To overcome both, we introduce a translation step using a commercial API. In summary, this work has the following contributions:

- an **account** of language edition relationship to combatants;

- the **identification** of battle outcome as one value check; and
- a **method** for semi-supervised corroboration which may aid article neutrality vetting.

While this approach is only applied to Wikipedia articles on World War I battles, news articles are a natural next step towards generalizing this method. It is clear that, in Wikipedia, action and dialogue are intertwined. But does this extend to other kinds of articles? In particular, what is the extent to which stance be modeled in discussions of a particular assessment community? This led to our next set of experiments investigating discussions of a wider variety of Wikipedia articles.

Broader implications. How differences between language editions should be treated requires more work on the social and political side of this problem. The latter part of this chapter attempts to do preliminary work characterizing the social interactions between multilingual users and monolingual users and the effects of those interactions on article composition. Ultimately, addressing the question of the meaning of these differences and how they should be addressed is out of scope for this work. But FR Wikipedia appears to be the most balanced account of these battles. Perhaps, if we value neutrality in our data language editions such as the FR version could be used to good effect. Possible reasons for this neutrality could be the willingness to engage with other languages (e.g., a polyglot tendency over a monoglot tendency). Being able to cite more sources than ones in a single language could make it easier to obtain more perspectives. See Appendix B for more information about the interaction dynamics that occur on the EN battle article talk pages.

CHAPTER 4

POLICY INVOCATION AS CONTEXT

In Chapter 3, the Wikipedia battle articles are compared to one another to identify possible Neutral Point of View violations; they serve as both **base documents** and **context documents**. But there are more values than neutrality in Wikipedia, but they are only implicitly available in the document form. Thankfully, these values, including Neutral Point of View, are documented in an ever growing number of policies and guidelines for article writing [42, 134]. To coordinate and create articles that adhere to these policies, editors discuss article edits in the corresponding talk pages [141]. The production of Wikipedia articles is a highly collaborative process, with coordination taking place on the corresponding article talk pages [73]. It is known that Wikipedia editors will explicitly state their reasons for modifying their article and their intention to coordinate with other editors. This has been exploited to modeling action-driving dialogue successfully [120]. It is also known that talk page conversations may devolve into arguments in this setting though [308].

To find a better alignment between article quality and article discussion, we turn the lens of investigation to Wikipedia's Featured Article community. The Featured Article community reviews articles nominated by dedicated editors who want the article to appear on English Wikipedia's front page. It provides a signal of quality (acceptance / rejection of the article) and context for that decision (nomination discussion). During this process, we observe that Wikipedia's values are invoked explicitly using links or implicitly using highly coded phrases. Article outcomes are decided on the merit of reviewer arguments rather than popular vote. So referencing Wikipedia's policies and guide-

lines is likely done in an effort to justify the importance of a reviewers claims in terms of Wikipedia's values.

What if a model was trained to score the article? In theory, it could lessen the load of reviewers. This is similar to automated writing evaluation [265] and automated reviewing [286]. The **base document** is the article itself, and the **inference task** is the article's outcome. This is the basis of work that was done with Esin Durmuş, Lillian Lee, and Claire Cardie.

But what need would we have for a **context document**? While technically a model may be trained directly on an article's submission version and its eventual outcome (rejection or acceptance), a value-sensitive approach requires investigating community values and the way they prioritize those values. To better align an article scoring with FAC values, we consider the use case of incorporating the discussions as **context documents** in an automated article scoring task.

But first, are the reviewers' policy invocations consistent with article outcomes? Jumping to using reviewer data risks replicating existing biases with little accountability. This chapter sets out to identify possible social biases in the FAC process as a step towards value-sensitive writing evaluating. To account for other kinds of reviewer-nominator bias, contextual factors such as the experience of the nominator and the chronological ordering of comments are contrasted with policy invocations. To check that the FAC nomination discussions are consistent with the stated Featured Article Criteria, the nomination discussion becomes the **base document** and the policies become the **context document**. The **inference task** remains nomination outcome prediction.

4.1 Task Introduction

The process of evaluating nominations of articles for Featured Article status in English Wikipedia is an excellent case study for examining how a long-running community has used and prioritized (or failed to appeal to) written criteria in making important evaluative judgments.

Nominated articles undergo the rigorous Featured Article Candidate evaluation process (FAC), wherein reviewers publicly discuss the article's merits and flaws. Review comments may include mention of written policy and guidelines,[1] which can be considered a form of grounding for the reviewer's judgment. We examine how much "weight" the invocation of policies, jointly and separately and explicitly or implicitly, appears to have in the ultimate acceptance/rejection decision. We summarize our investigation in the following research questions:

RQ 1: What policy invocations are invoked in Featured Article Candidate discussions?

RQ 2: How are policy invocations correlated with article outcomes?

RQ 3: When are policy invocations used in nomination discussions?

To address these research questions, we use more than 10 years worth of discussions from the Featured Article Candidates archives in English Wikipedia. RQ1 is addressed via extracting corpus statistics on policy invocations; the data

[1]Henceforth, we use the shorthand "policy" for both policy and guidelines.

Description	Nomination Text
Metadata	Wish You Were Here (Pink Floyd album); Nominator(s): Parrot of Doom (talk) 10:12, 27 August 2009 (UTC); Featured article candidates/Wish You Were Here (Pink Floyd album)/archive1
Nominator	Ok so it isn't quite the greatest rock album ever like DSotM, but its probably up there in the top ten and certainly ranks top of some people's lists. Its slightly shorter than I'd like but that's more down to a paucity of written material than anything else (DSotM has entire books written about it, WYWH does not). [REDACTED] (talk) 10:12, 27 August 2009 (UTC)
Reviewer #1	Comment. The alt text is quite good and very detailed, but it's a bit long; see WP:ALT#Brevity. Relatively unimportant details like "The sky is blue with no clouds." can be omitted. [REDACTED] (talk) 08:30, 28 August 2009 (UTC)
Reviewer #2	Comment I'll give a more thorough review later, but here are a few right off the bat: * discogs.com is not reliable in the same way that IMDb is not reliable, as they're both user-generated. You've double sourced the release information you're citing anyway, so unless I'm mistaken removing the discogs citation doesn't change anything. * Reference 45 seems to be broken; check over this real quick, seems like a simple fix. It's evident from the ref name alone exactly what you were trying to cite. * Same issue as DSotM on the formatting of the "Sales chart performance" table. * Another minor issue, which also came up with DSotM; I'm not sure how exactly this is fixed, but it bothers me that the titles of web references are italicized and not in quotes as they should be. Can someone explain why this is? –[REDACTED] (talk) 07:48, 29 August 2009 (UTC)

Table 4.1: Example excerpt from a FAC-nomination conversation. It includes explicit policy invocations (e.g., WP:ALT#Brevity) and implicit policy invocations (e.g., *not reliable* and *Reference*). Reviewer names are [REDACTED] for privacy. Our goal is to understand the implications of the pattern of usage of policy invocations for the outcome of the article.

processing and collection methods are discussed in §4.3 and §4.3.1. In §4.4, the data are then used to train a logistic regression model to address RQ 2, where our hypotheses are encoded as features of the conversation, participants, and

article metadata. Discussion features are treated as fixed effects and categories are treated as random effects; we report both correlation coefficients and model performance and discuss their implications. In §4.5, we address RQ 3 with a Cox-regression test comparing the time to first explicit and implicit policy invocations for both promoted and unpromoted nominations.

We initially hypothesize that policy invocations are generally more negatively correlated with promotion, given the history of policies being used to flag problems in articles and spur corrective action [42, 226, 212]. If so, references to Wikipedia's Verifiability policy would be anti-correlated with promotion, in line with work on characterizing escalation patterns in Wikipedia's Articles for Deletion discussions [118]. Our results indicate, though, that while the Reliable Source guideline and Original Research policy are correlated with *non*promotion, verifiability mentions are correlated with *promotion*. With respect to prediction results, we note that including invocation-context features improves article outcome prediction. We also highlight instances where a bag-of-words model outperforms one trained with policy invocations.

While the Featured Article Criteria, the rules by which FA candidates are meant to be evaluated by, might be thought to be determining factors in the eventual outcome, we also investigate other possibilities. Nominator attributes such as experience, community involvement, and choice of topic may also affect their interaction with the reviewers; similarly, reviewer experience may indicate how likely the are to assess a nomination negatively. We do find that one can predict an article's outcome within the first few reviewer comments using only policy invocation features, and that these features improve a model's performance even in the presence of features derived from the number of re-

view comments, nominator experience and history, and article category. This work validates the consistency of the Featured Articles Candidate nomination process with respect to the policies and guidelines described in the Featured Article Criteria. This may have implications for downstream tasks such as writing assessment and tools for policy automation and support.

How are policy invocations relevant to problems facing contemporary platforms? Fact-checking and claim verification has become an increasingly popular task with more data resources available for training automated fact-checking models [99, 267, 13]. Datasets for these tasks rely on the reliability of Wikipedia [92], and communities such as FAC produce the some the most reliable content using policy invocations to ground evaluation conversations [228].

4.2 Related Work

Our work builds on an extensive foundation of research on Wikipedia article quality and collaborative writing. The goal of this work is to identify how patterns of policy referencing relate to candidate outcomes, particularly in comparison to other factors such as reputation and category.

4.2.1 Policy and guidelines in action in online communities

Why focus on references to *policies*? There is a strong precedent for working with policies in Wikipedia research. Early on there were calls for more policies and vandalism prevention for quality control [55]. Around 2007, a sharp decrease in the number of users and editors was identified and followed by

recommendations for policy and future directions for Wikipedia [221]. Other work followed, touting policies as a necessity for community coordination and function [42, 26, 118]. One branch of later studies homes in on the significantly slowed growth in Wikipedia that follows 2007 and new user enculturation [261, 207, 193]. Another branch of work focuses on specific policies and the edits that follow in the wake of their invocation [226, 114, 212, 228]. The confluence of these branches is to define users by their edits in Talk pages and articles [8, 10, 299]. While there has been work on community norms outside of Wikipedia [50, 74, 288], Wikipedia lends itself to analysis with its transparent policies and revision tracking. Furthermore, there is significant overlap in rules even across language editions [117], which is encouraging for the potential to generalize beyond English.

Perceptions of fairness Delivering decisions and negative feedback in a manner that is conducive to work and improvement has long been studied in applied psychology [17, 18, 156]. This research generally finds that specific, prompt, conscientious, and actionable feedback is most effective and perceived as more fair. Wikipedia guidelines offer advice for structuring sentences on a semantic and syntactic level, thereby representing more specific and actionable feedback. Policy invocations may occur throughout a review, but earlier invocations may be more valued. In a grounded analysis of talk pages of English Wikipedia, [149] have observed that ambiguity in how policy is invoked can give rise to powerplays. But they find that overall policy generally facilitates collaboration. [274] finds that gender plays a role in how articles are nominated for deletion in AfD discussions; specifically, in Wikipedia's biography articles, women are more likely to be labeled as non-notable than biographies about men

though both articles may meet Wikipedia's standards for inclusion.

Policies for evaluating contributions Recent work has found that fixed attributes such as author reputations are correlated with writing evaluation outcomes [110, 272]. But FAC are promoted on the basis of the arguments for acceptance or rejection, not popular vote. In a study of Wikipedia's Articles for Deletion process, the extreme opposite of FAC, there is prior work on identifying references to policies as both a source of contention and argument strength [118]. Building upon this work has enabled the development of more tools for automated policy enforcement [306]. The research into evaluating article content directly is a continuing effort to aid editors' assessment of their work and ease the burden on reviewers [2, 62]. But the Value-Sensitive Algorithm Design framework highlights the adverse effects of modeling problems without accounting for community values, such as the loss of new users who are deterred by bot edits [312]. Based on this framework, an analysis of Wikipedia stakeholder values with respect to machine learning systems makes the following recommendation: "[Recommendation 20:] To aid in community governance efforts, algorithmic systems should provide mechanisms to assess the trustworthiness of community members based on their community contributions and behaviors" [255]. We hope our analyses can provide scalable accountability for assessment processes on Wikipedia such as FAC.

4.2.2 Policies for conversational grounding

Why focus on policy *references*? We hypothesize that policy invocations are critical for *grounding* conversations in the *Featured Article Criteria*. Grounding is

theorized to facilitate cooperation in conversation [56, 37, 25]. Designing mechanisms for grounding in impoverished contexts such as online text-online communication such as Wikipedia is particularly critical [37]. In discussions found on Wikipedia, invoking policy invocations is a salient way Wikipedia editors may ground their discussion of article changes in the policies and guidelines of Wikipedia. Prior work has looked into characterizing link-based referencing patterns in terms of grounding behavior in collaborative code writing environments [55]. We argue that when hyperlinked, these policy invocations similarly allow for easy access to the detailed pages of various rules, which are continually updated [117].

4.2.3 Featured Article Candidates

Why focus on policy invocations in the *Featured Article Candidates (FAC) nomination discussions in English Wikipedia*? Though studies have been done on academic papers [132, 272], privacy policies and conference resubmission make it difficult to track specific submissions. Additionally, conference papers contribute groundbreaking work. FAC specifically forbids original research and discourages the submission of articles likely to be updated over time, two factors which make the evaluation of academic papers difficult. FAC is also one of the more active assessment processes on Wikipedia, and the decisive outcome for a large corpus of articles is useful for understanding the implications of policy invocations. English Wikipedia was chosen because of its size (in both user population and article population) [139], the availability of natural language processing tools in English (which we acknowledge limits generalizability) [22], and its accessibility to the authors' whose shared language is English. This is

not to say that our results are only applicable to English Wikipedia, but that we choose depth over breadth in this work. Our method is primarily dependent upon the consistent use and identifiability of policy invocations. Additionally, special care must be taken to account for differences among language editions [189, 117], particularly as related to Featured Article review criteria.

Though we focus on this one community, we note that there is a significant amount of work on socialization in Wikipedia [141, 287, 311, 256, 193, 194]. WikiProjects often serve as incubators for article development, wherein editors can coordinate with one another around subjects of shared interest [256]. Editors can solicit feedback and recruit help through a WikiProjects, and certain types of assessments (e.g., A-class article assessment) may be defined by a specific WikiProject[311]. An article may be of interest to multiple WikiProjects and may be marked as such. We use Category as assigned by the Featured Article listing as a proxy for understanding WikiProject influence, though we acknowledge the limitations of assigning an article to a single representative category. One category assignment may not capture the fact that multiple WikiProjects may be invested in a single article, and identifying article topics is an area of active research [140, 126]. Other work has looked into socialization and the role of norms through Wikipedia's Teahouse sub-community established for answering the common questions of newcomers [193, 194].

4.3 Featured Article Nominations

Nominators typically should have significantly contributed to an article under consideration for Featured Article status, but any Wikipedian may participate

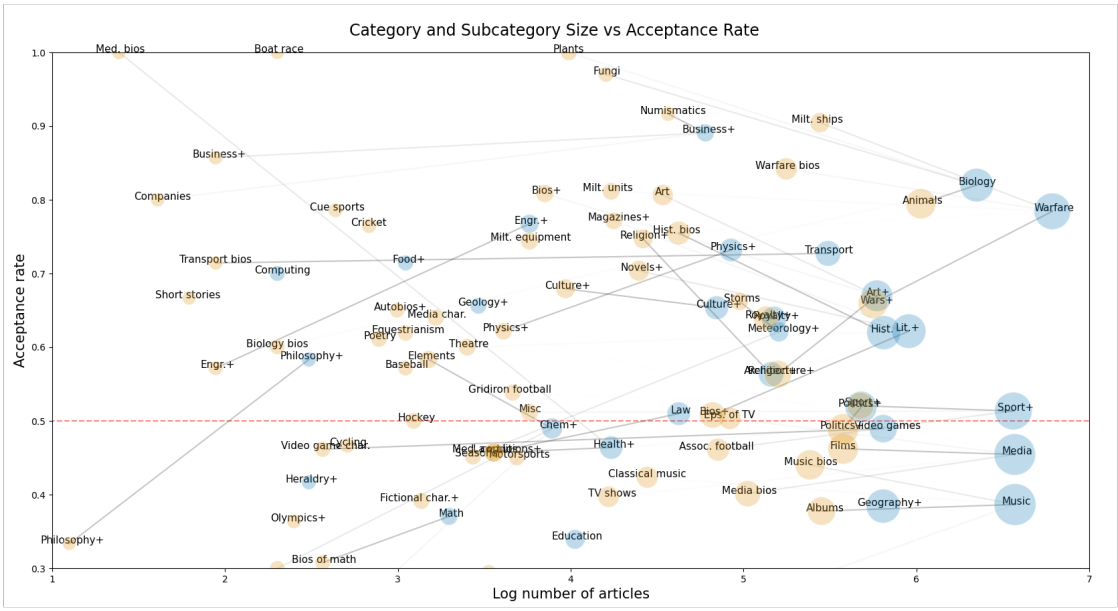


Figure 4.1: Percent of nominated articles in a given Category (dark circles)/Subcategory (light circles) versus the log number of nominators in that category. "+"s indicate truncation of category name. Dot size denotes number of nominations. Gray lines link Categories to Subcategories; line weight is correlated with category proportion. Red dashed line indicates 50% acceptance rate.

as a reviewer. For a nominated article to be accepted, it must not only satisfy general content policies, but also meet a special set of *Featured Article Criteria*. During the nomination period, reviewers discuss the article (see Figure 4.1 for a conversation snippet); nominators may participate and submit revisions. Acceptance requires consensus among the reviewers that the FA criteria have been met; whether consensus has occurred is determined by a small group of FAC coordinators.

Our corpus is drawn from Wikipedia's archive of Featured Article Candidates reviews, dating back to 2004; we also make use of the fact that nominators often record their submissions on their Talk pages. Because of the highly overlapping style of Wikipedia Talk pages, we use the diffs of review revisions to ob-

tain a time-linear representation of the review conversation. While prior work has found methods for extracting a tree structure from Wikipedia talk pages [154], we choose to emphasize the chronological ordering to study nomination survival (see § 4.5).

We extract policy invocations p, which may be explicit (i.e., mentioned directly by name) or implicit (see Figure 4.1 for examples) from review comments by regular-expression matching (see §4.3.1). Note that p may be part of a description of a policy violation or part of a validation that there is no such violation, depending on context.

We may consider the outcome of a nomination to be a function of nominator experience or *history* h_n, reviewer experience h_r, article category c, article subcategory c', and set of policy references $\{p_i\}$. We may also elect to include a policy reference's context, thus representing a policy invocation as a bag-of-words vector p_{bow}. A subset of policy invocations may also be used to separately study implicit (p^i and p^i_{text}) and explicit usage (p^j and p^j_{text}).

Licensing Wikipedia data is available via the Creative Commons Attribution-ShareAlike 3.0 Unported License [291]. Our code for data collection, processing, and analyses will be made available upon article acceptance.

4.3.1 Parsing Discussions

The review discourse is a multi-turn dialogue between the nominators and reviewers. During the course of conversation, the article may be updated continuously as the reviewers re-read the article, make requests of the nominator,

and assess the article's progress. This culminates in a reviewer's assertion of supporting (e.g., "support') or opposing (e.g., "oppose"/ "object") the article's nomination, together with a brief explanation. We now describe our method for parsing the discussions.

4.3.2 Piecing together conversation

A review is comprised of review comments. The comment text is extracted from the review by applying a line-level diff method[2] to subsequent review revisions obtained through the Wikimedia API. The timestamp and editor's username are critical metadata used to identify the participant and time of the comment. Though an informal thread structure may be created using indents, for the purposes of this work comments are treated in a chronologically sequential manner. Editors frequently edit their text with strikethrough tags to indicate that it no longer applies in the review. Because struck text is often a duplicate of earlier comments, it is removed. We emphasize, though, that the original comment text — which appeared earlier in the review process — is retained.Review discussions range in number of comments from as few as 3 to as many as 700. To exclude outliers in length which may include desk rejects and extremely controversial or complicated subjects, articles more than 3 standard deviations from the mean in log number of comments are removed. Comments made by FACbot or indicating promotion status are also removed.

[2]From the difflib library in Python.

Policy	explicit	implicit
Article_size	48	21
Citing_sources	307	45
Featured_article_criteria	521	5820
Manual_of_Style	4430	35362
Neutral_point_of_view	187	2625
No_original_research	317	3079
Non-free_content_criteria	382	1758
Notability	25	1976
Reliable_sources	427	6844
Summary_style	74	2007
Verifiability	369	1780

Table 4.2: Counts of implicit and explicit references retrieved from the training set of nomination reviews under stricter. Implicit references are derived from alt text phrases co-occurring with an explicit policy invocation > 1 time.

4.3.3 Identifying policy invocations

Policy invocations are a means by which reviewers can strengthen their argument for support or opposition. Figure 4.1 shows some example comments with and without policy invocations. Table 4.2 lists the most common policy invocations at FAC, including counts for implicit and explicit mentions. We addresses RQ1 by highlighting that policies like Verifiability and Neutrality rank highly, but guidelines such as Manual of Style, Citing Sources, and Reliable Sources are more pervasive. Implicit mentions outnumber explicit mentions by at least a magnitude.

Explicit policy invocations Policy invocations are first identified by a regex search for internal Wikipedia short links of the form [[WP:TAG]], [[Wikipedia:TAG]], or [[MOS:TAG]]. Multiple tags may be used to link to each policy; for example, the Wikipedia:Neutral_point_of_view page may be linked to by [[WP:NPOV]], [[WP:POV]], [[WP:WEIGHT|undue weight]],

Policy	Implicit keywords
Manual of Style	accessibility, accessible, alt, alt text, alt texts, alternative text, alternative texts, avoid self references, captions, click here, consistency, context, dash, dashes, easter egg (+61 more)
Verifiability	policy, reliability, reliable source, reliable sources, self-published, self-published source, self-published sources, self-publisher, self-publishing company, verifiability, verifiable, verification, verify
Citing sources	bundling, citing sources
Neutral point of view	due weight, fairly and without bias, neutral, neutral point of view, neutrality, npov, undue weight, weight
Featured article criteria	1(e), 1a, 1b, 1c, brilliant, crit 2c, criteria, criteria 3, criterion 1a, criterion 1c, criterion 2a, criterion 3, fa criteria, fa criteria #3, fa criterion (+11 more)
Non-free content criteria	criteria for inclusion of non-free content, fair-use criteria, nfcc, non-free content criteria, policy, wp:nfcc
No original research	no original research, or, original research, original synthesis, primary, primary source, primary sources, synthesis
Reliable sources	high-quality reliable source, reliability, reliable, reliable source, reliable sources, reliably sourced, rs
Notability	notability, notable
What Wikipedia is not	what wikipedia is not
Summary style	summary, summary style, summary-style
Article size	article size

Table 4.3: Keywords that serve as implicit policy invocations. These words were extracted from the alt text of explicit invocations, and manually inspected for quality of content.

and [[WP:FALSEBALANCE]], inter alia. To reduce tagset size, we normalize policy invocations to a policy-page name by following the WP link, recovering the policy url from page metadata, and then extracting the name from the url. Thus, [[WP:POV]] and [[WP:FALSEBALANCE]] both become "Wikipedia:Neutral_point_of_view". This results in our set of *explicit policy invocations*.

Implicit policy invocations Only approximately 15% of comments contain explicit policy links, and manual inspection reveals that it is common for users to mention policies without links. To recover other likely policy invocations, we use the alt text co-occurring with the WP links (e.g., "undue weight" in [[WP:WEIGHT|undue weight]]) to bootstrap a larger set of implicit policy invocations. The full alt text phrase must be matched to count; only alt text phrases co-occuring with the same explicit policy invocation at least twice are counted. The policy invocation link text itself is also included in the set (i.e., *Neutral point of view* is considered a possible implicit mention even if not observed in alt text). Stopwords, spaces, and unfinished WP links are filtered out. Once these words are identified in the alt text from the link collection pass on the training set, a regex of variants is used to identify implicit references in a second pass over the data. This increases the number of identified comments with rule invocations significantly. See Table 4.2 for counts of explicit/implicit policy invocations. Note that the total number of policy invocations to several of the selected policies/guidelines increase by a magnitude because of implicit policy invocations being included. See Table 4.3 for a portion of the final keywords list. The complete list will be made available along with the data and code if the paper is accepted.

Explicit and implicit policy invocation features For our POLICY_ONLY features, each policy invocation is counted in each nomination discussion. Policy invocations appearing in fewer than 1% of reviews are excluded. For both implicit and explicit policy invocation features, there are a total of 13 features, with one feature marking the absence of any policy invocations.

Policy invocations in context We also collect the context of implicit references; the intuition is that "there are npov issues" are "there are no npov issues" express opposite opinions regarding criteria satisfaction. Specifically, we expand policy invocation regex matches to include up to 20 characters of content preceded and followed by word boundaries. Whereas POLICY_ONLY would map these all to [[WP:FAC]], the context features treat the policy invocation and its context as a bag-of-ngrams ranging from 1-grams to 4-grams derived from the text of all policy invocation contexts for a single nomination discussion concatenated. To avoid overfitting and to facilitate comparison, only the 100 most frequent n-grams are used. The terms "support", "oppose", and "object" are filtered out to prevent the outcome label leaking into the train/test instances. A stopword list from spaCy [107] are used for filtering. The resulting features are tf-idf weighted.

4.3.4 Reviewers and nominators

Although nominator comments are part of the process, we are particularly interested in the reviewers' comments. To filter out nominator comments, we identify nominators by the username meta-data for the first comment and by a username regex search of the first comment with a "Nominator(s)" tag.

Participant experience Though evaluating the quality of the article content itself is out of scope of this work, the experience of the participants in the process may be established by reviewing the nominators' and reviewers' nomination and review histories. Presumably, a nominator with more nomination experience will be more likely to succeed at FAC. Reviewers with more experience

may be more consistent than those with less experience. We model experience as treated as a binary variable with respect to the distribution of the natural log of the number of previous nominations or reviews. If there are multiple nominators or reviewers, the natural log of the median experience is used.

The success of a nomination is expected to vary with nominator experience, though less with reviewer experience. Nominator experience and reviewer experience are treated as a binarized version of the number of nominator past nominations, nominator past reviews, reviewer past nominations, and reviewer past reviews. The threshold for binarization is determined by the median value of experience across the corpus.[3]

4.3.5 Category assignment

Though the invocation of rules may be important to the outcome of the nomination, they are conditioned on the article itself. A follow-up hypothesis is that certain categories of articles may be easier to cultivate for Featured Article status.

While any article may be labeled with several category tags, accepted articles belong to a definitive category shown by the Wikipedia Featured Articles listing. Poorly applied category tags or unclear category membership may be responsible in part for rejected articles' outcome. There are methods for content-free category labeling that rely on article linking patterns [126], but we find anecdotally that these labels are too noisy for our purposes and may not align with

[3]Though more features such as number of nominators and continuous features did improve the performance of the features as a whole in this task, we exclude these results to maintain clarity of analysis.

the categories on the Wikipedia Featured Articles pages. For consistency and simplicity, we apply category labels that best align with the Featured Article categories available based on content similarity.

Cosine similarity between each unlabeled article and a representative article from a category is calculated. The most specific label is used (e.g., Warfare/Biographies) and treated as independent of the broader label (e.g., Warfare). Let a_i be a rejected uncategorized article where $\tilde{c}$ is the category to which we assign a_i. Let a_j^c be an accepted article sampled from known category c according to a uniform distribution. For every unassigned article, we let $\tilde{c} = argmax_c cossim(a_i, a_j^c)$. Each article is represented as a bag-of-words vector of unigram and bigram counts. N-grams that appear in fewer than 1% of articles are removed. While a full analysis of *article* content is out of the scope of this work, we can use category and subcategory information as a proxy. See Figure 4.1 for an overview of category and subcategory sizes, number of nominations, and acceptance rates. This figure suggests that broad category membership is correlated with article outcomes.

Categories and subcategories are treated as separate binary categorical features. Subcategory features are expected to better capture the subtleties of category information: for example, articles about wars tend to be less successful than articles about military ships, as seen in Figure 4.1, though both occur under the Warfare category (in the "northeast" section of the plot).

4.3.6 Baseline features: length

Though we do not aim to evaluate the article itself here, *article length* has been used as an indicator of article quality in previous work [111, 29]. We measure it by the log of the number of characters. We hypothesize that greater *review length* indicates more problems (and is hence correlated with failure), although one could also imagine that it indicates more help from reviewers (and hence correlates with success).

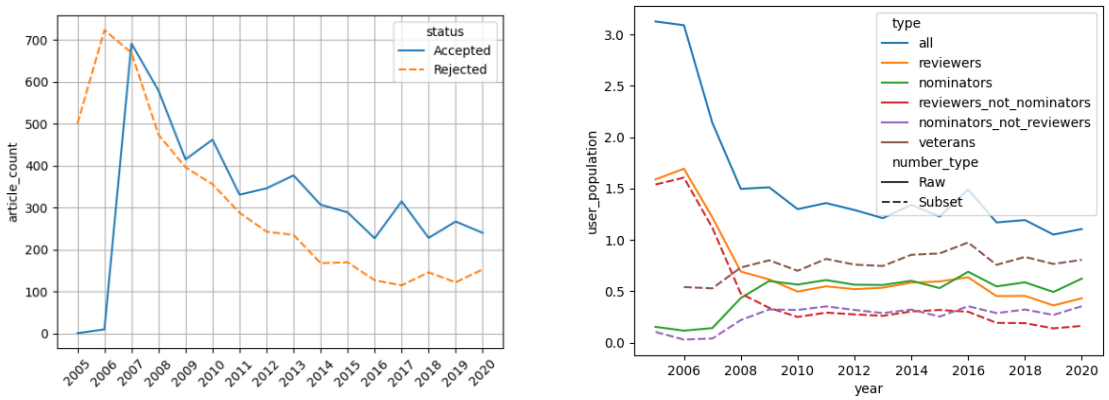
4.3.7 Upperbound features: bag-of-words of review content

We include in our second experiment a comparison to a bag-of-words feature set. This feature set is simple yet effective upper bound on our dataset. To limit overfitting we limit the maximum number of tf-idf weighted features to 1000. Stopwords and instances of "support" or "oppose" comments are removed, but more interesting forms of processing are left to future work.

4.3.8 Macrotrends

We examine the context from which our data originate for macrotrends in the FAC process. Some statistics of interest are user population, submission count, and acceptance rate. Exploratory data analysis reveals that there is a sharp change in the FAC community around 2007, stabilizing in 2008; hence, submissions that precede 2008 are excluded except when gauging user experience.

Earlier years see a markedly higher submission rate, with the population



(a) Submissions per year.

(b) Users per article per year.

Figure 4.2: The number of article submissions trends downwards over time, although the population of nominators increases. The reviewer population also decreases substantially, though this is likely due to changes in Wikipedia. For 2005, we were not able to recover accepted nominations. To compensate for the early instability and inconsistency in forum archiving, we only take data from 2008 and after.

and number of nominations growing throughout 2005-2007 (The selection of Featured Articles began in mid 2004). The acceptance rate varies slightly, but is relatively stable between 50 to 64%. Generally, the number of explicit invocations is stable, if increasing over time. The Manual of Style, being a compendium of numerous guidelines, is referenced in nearly every article from the beginning.

The number of people per nomination participating in FA selection process also stabilizes around 2008, after a sharp decline (Figure 4.2b). This corresponds with general observations of Wikipedia's slowed growth. Though it cannot be established from these figures alone, the increase in rule invocations occurs around the same time as the decline in total reviewers and an increase in nominators.

4.4 Outcome Prediction

To address RQ 2 (relationship between policy invocations and promotion outcome), we set up our outcome prediction task as follows with features describe below encoding our hypotheses about policy invocations. Our instances are defined as the concatenation of the reviewers' comments in chronological order, including edited sentences. The label is the outcome of the nominations as indicated by the promotion result typically found in the final comments. The promotion result may be expressed as a template or a phrase, or indicated by the link at which the nomination was found (promoted and nonpromoted nominations are archived separately). Our positive label is *rejection*, as this is the minority class and facilitates interpreting results in Table 4.4.

4.4.1 Training details

We selected the set of reviews from 2008 to 2020, filtering out reviews from the more tumultuous period before. The review comments are anonymized before feature extraction. The train/dev/test split (5289 training instances, 588 dev instances, and 1469 test instances) is created by randomly sampling 20% of the uniquely identifying nomination titles and is constant over all experiments. A logistic regression (LR) model is tuned on a dev set, with the best regularization parameter C chosen from the set $\{.1, 1.0, 3.0\}$ and selected from the model with the best AUC. All features are scaled by their maximum absolute value in the training data.

feature	auc	F1	Prec	Recall	#F
article length	0.524	0.000	0.000	0.000	1
review length	0.652	0.234	0.932	0.134	1
category	0.660	0.451	0.576	0.371	30
subcategory	0.678	0.414	0.587	0.320	95
explicit POLICY	0.525	0.016	0.714	0.008	13
implicit POLICY	0.602	0.310	0.601	0.209	13
expl+impl POLICY	0.613	0.378	0.604	0.275	26
context POLICY	0.723	0.558	0.631	0.500	100
user stats	0.679	0.580	0.606	0.557	4
[ALL]	0.816	0.674	0.730	0.626	257
-article length	0.815	0.679	0.745	0.624	256
-review length	0.790	0.652	0.694	0.614	256
-(sub)category	0.791	0.639	0.727	0.570	132
-all POLICY	0.776	0.637	0.716	0.574	131

Table 4.4: Test-set outcome-prediction AUC scores for logistic regression using the indicated feature set. A minus-sign indicates that the feature set was ablated away from ALL. The feature all POLICY indicates implicit, explicit, AND context policy invocation features. #F indicates number of features

4.4.2 Results

ROC AUC is a convenient summary of precision-recall characteristics across many prediction thresholds, and it is used here to evaluate and compare model performance across settings. Table 4.4 shows that the feature set that performs the best in isolation is POLICY CONTEXT, and removing all POLICY features is the feature set whose deletion from the set of all features hurts performance the most. The explicit mentions are too sparse to provide a strong signal, although they do have high precision. The implicit mentions are far better for predicting outcome, likely because of how common they are. But is is the policy invocations in context that obtain the highest AUC. Care was take to limit the small number of features to avoid overfitting given the size of our dataset.

Though article length has been shown to be predictive of quality in previous

work, its removal has less of an impact on nomination outcome than other features.This contradicts the findings of previous work, but this may be because other work includes articles found in Wikipedia at large, rather than the FAC setting. Review length is *negatively* correlated with rejection, suggesting that nonpromoted nominations are quickly weeded out and survivors are given additional advice for article improvement.

As an aside, we note that the coefficients of the trained models indicate that the most successful subcategories are largely Biology and Warfare related (e.g., Biology/Plants and Warfare/ships), while Music and Chemistry subcategories are correlated with failed nominations. This corroborates Figure 4.1.

Table 4.5 depicts the learned coefficients for individual policy-related features and helps us answer RQ2. The importance of source reliability is clear: "Reliable Sources" shows up with a highly positive coefficient in the POLICY ONLY column. Mentions of reliable sources and original research are correlated with rejection, suggesting these policies are most strongly enforced. Somewhat surprisingly, Verifiability and Neutral Point of View mentions are correlated with promotion, not rejection. Given that these are serious policies on Wikipedia, we believe that this correlation is due in part to affirmative practices. For example, in Table 4.5 the affirmation that the article "meets" some criteria is correlated with positive outcomes. In Table 4.6, "1a good ." likely goes against the usual usage of "1a" but the nuance in affirmative tone is captured by the bow classifier. Additional experiments omitted for brevity indicate that NPOV severity of reference may depend on whether it co-occurs with Reliable Source references or Manual of Style references.

POLICY ONLY		POLICY CONTEXT	
coeff	feature	coeff	feature
1.76	NO implicit-i	1.63	article
1.56	reliable sources-e	1.58	high quality
1.53	non-free content criteria-i	1.41	wpwiafa
1.32	featured article criteria-i	1.26	review
1.10	reliable sources-i	1.15	issues
1.00	notability-i	1.14	wprs
0.93	no original research-i	1.11	please
0.52	summary style-e	1.03	prose
0.49	what wikipedia is not-i	0.94	wpalt
0.40	notability-e	0.87	b
-0.20	summary style-i	-1.27	image
-0.36	neutral point of view-i	-1.33	context
-0.38	verifiability-e	-1.47	captions
-0.45	citing sources-e	-1.68	dash
-0.49	verifiability-i	-1.68	seems
-0.51	featured article criteria-i	-1.68	consistency
-0.66	NO explicit-e	-1.80	reliability
-1.14	article size-i	-1.86	accessible
-3.56	manual of style-i	-2.64	meets

Table 4.5: Coefficients of models trained with policy invocation features. The "-i" suffix indicates features derived from implicit features, while the "-e" suffix indicates features derived from explicit mentions. NO implicit/explicit indicates that no features of each respective type were found in the review.

4.5 Deliberation Duration Effects

In this section, we investigate whether there are effects based on the number of comments into the discussion. The intuition is that there may be survivor bias due to the fact that certain comments may be "fatal" to the nomination.

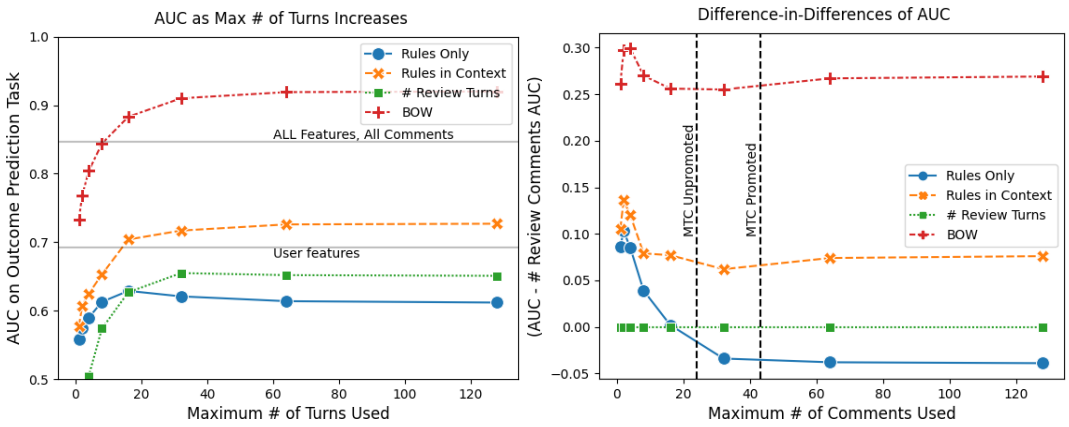
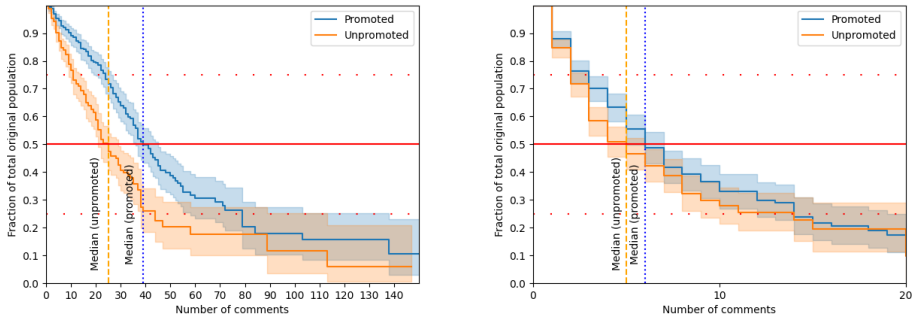


Figure 4.3: (a) AUC when restricting to the first n comments; (b) difference in AUC between a given feature set and the feature set "number of review comments"

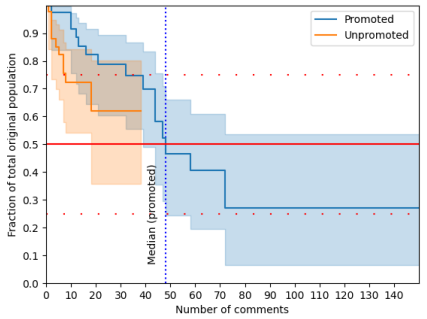
4.5.1 Cumulative comments and feature efficacy

As foreshadowed by our previous finding that review length is predictive of outcome (Table 3.5), there is a notable difference between promoted and unpromoted articles in median number of review comments: 43 vs. 24. But when we plot performance *restricting the test instances to the first 1, 2, 4, 8, . . ., or 128 comments*, the non-prefix-length features provide the greatest performance gains over the prefix-length features in the 8- to 16-comment-prefix range, that is, relatively early into the discussion, and well before the median number of comments. Figure 4.3a shows the AUCs for four feature sets (the number of review comments, or prefix length, is the curve with green squares); Figure 4.3b, a difference-in-difference analysis, shows the delta in AUC of each of these feature sets in comparison to prefix length.

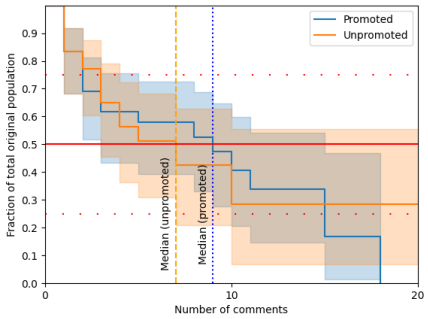


(a) Explicit, inexperienced

(b) Implicit, inexperienced



(c) Explicit, experienced



(d) Implicit, experienced

Figure 4.4: Kaplan-Meier survival curves [68] that chart the number of comments a nomination survives until its first policy invocation. Nominations are grouped by policy invocation type (explicit / implicit) and nominator experience (inexperienced / experienced). The first invocation of any policy occurs earlier for unsuccessful nominations.

4.5.2 Time to first explicit policy invocation

One can be more certain of a nomination's outcome as a review progresses, but how does time to the first policy invocation and only that policy invocation relate to a nomination's outcome? Figure 4.4 demonstrates that the *first* invocation of a reliable-source policy occurs earlier for unpromoted nominations. For example, 50% of unpromoted nominations by inexperienced[4] nominators encounter explicit policy invocations before the 25th comment, whereas for nom-

[4]"Inexperienced"= ln(number of nominations +1) > 2. When multiple nominators are involved, we use the median experience.

	Ppl	SHORT SENTS	Ppl	LONG SENTS
promoted	1.00	e.g .	1.17	otherwise , sources look okay , links checked out with the link checker tool .
	1.02	ing .	1.28	comments you ve mixed using the template citation with the templates that start with cite such as template cite journal or template cite news .
	1.04	tag .	1.32	random check of three other online citations showed them as fine .
	1.04	apologies .	1.34	odnb only mentions july to october ⟨unk⟩.
	1.04	1a good .	1.39	ref 100 30 september 2001goes against your date consistency .
unpromoted	1.01	ugly .	1.16	25 million place the non-breaking space between 25 and million .
	1.02	questionable .	1.19	otherwise , sources look okay , links checked out with the link checker tool .
	1.02	rephrase .	1.20	would uru required five years and 12 million to complete .
	1.03	george yes .	1.23	apparently not even a see also for visual art , despite an ⟨unk⟩good series of period articles .
	1.04	seriously .	1.25	even the importance of ⟨unk⟩values section which addresses applications to industry remains vague as to specific examples of its usage i.e .

Table 4.6: Sample sentences from instances in the intersection of a full-content classifier's correct predictions and the policy invocation-based classifier's incorrect ones. Sentences are sorted by perplexity score (Ppl) assigned by a Kneyser-Ney trigram language model trained on the data subset. The most likely 3+ token (SHORT SENTS) sentences are on the left, while the most likely sentences after dividing perplexity by the log number of sentence tokens are shown on the right (LONG SENTS). Note that the sentence "otherwise, sources look okay…" is frequent in both promoted and unpromoted articles.

inations that are promoted, 50% do not receive an explicit policy mention until the 40th comment. The difference between promoted versus non-promoted first-encounter time is significant only for explicit policy invocation encounters by inexperienced nominators (Cox regression test); the difference for first encounters with *implicit* policy invocations are non-significant. Our curated list

may sacrifice precision for recall. Because of ambiguity, implicit policy invocations may require more context to use effectively.

For experienced nominators, we observe that there is no significant difference in time to first explicit policy invocation with respect to outcome (Figure 4.4c). This may be because the population is smaller or because experienced nominators are more adept at avoiding obvious policy violations. This suggests that more explicit policy invocations are primarily used in an outcome-consistent way with inexperienced nominators. More work needs to be done to understand the interplay of experience and the act of referencing policies.

4.5.3 Policy invocations vs. full review content

We have observed that outcome can be predicted using policy invocation features. But experiments with a bag-of-words representation of the entire review outperforms the policy invocation features by a significant margin. If a reviewer is not using policy invocations in a way linearly consistent with outcome, *what are they doing?*

To address this question, we compare the true predictions of the content model to the false predictions of the policy invocation model both trained on the first 8 comments of a review. Though the AUC is lower per model at this point than at 16 comments, the difference between these features and the review length features are greatest at this point (Figure 4.3). This intersection of instances leaves us with a data sample that shows aspects of the review process perhaps not reflected by a policy "adherence checklist". For ease of manual inspection, we sentence- and word-tokenize these instances, filtering for some

non-alphanumeric characters. To rank the instances, we train a Kneyser-Ney ngram language model with maximum ngram order of 3 and a vocabulary of size 2000. Words not in the vocabulary are replaced with "⟨unk⟩" (unknown-word token). The trained ngram model is then used to calculate the (1) perplexity of sentences and (2) the perplexity of sentences normalized by the log length of the sentence in tokens.[5] The most likely comment sentences under these conditions are shown in Table 4.6.

A cursory qualitative inspection of the SHORT SENTS column suggests that subjective judgment (e.g., *1a good .* and *ugly .*) plays a greater role in the content model. Meanwhile, the LONG SENTS column suggests that domain knowledge (e.g., *apparently not even a see also for visual art , despite an ⟨unk⟩good series of period articles*) and minor issues (e.g., *ref 100 30 september 2001 goes against your date consistency*) are also better identified by the content model.

We leave the task of addressing article content issues to work on expert writing such as [132]. As for the subjectivity, it is unclear to what extent the Wikipedia community may want subjectivity to play a role in FA decisions. But our analysis and methodology may aid future work in identifying the extent to which subjectivity plays a role in the evaluation process.[6]

4.6 Conclusion of experiments

Over the past 15 years, Wikipedia has changed tremendously. Given an increasing interest in using Wikipedia, as a resource for fact checking data, our goal

[5]The correction reveals likely yet complex comments that might otherwise be downranked because of length.

[6]Observation, not criticism of reviewers, is the intent of this work.

in conducting this study was to understand how policy invocations are used to ground the internal assessment process at FAC and to what extent they correlate with outcomes. (RQ1:) We find that guidelines such as Manual of Style and Reliable Sources dominate the number of policy invocations, though Verifiability and Neutral Point of View policies are also prominent.

(RQ2:) Our results indicate that policy invocations are more predictive of outcome than some static features of the article such as category or nominator experience at time of submission, though not as predictive as full content features. Correlation coefficients indicate mentions of Reliable Sources are highly weighted (towards non-promotion with respect to mentions ignoring context; different contexts for Reliable Sources can be correlated with promotion or with non-promotion), while mentions of Manual of Style and Verifiability are correlated with success. Our results indicate that context is critical for disambiguating policy invocations.

(RQ3:) Additional analysis reveals that implicit policy invocations may arrive early in conversation. But using survival analysis techniques, we see that explicit policy invocations regarding reliable sources occur significantly earlier for unpromoted articles in comparison to promoted ones, in the case of inexperienced nominators' nomination deliberation.

4.6.1 Limitations

While our analysis is restricted to English Wikipedia's Featured Article Candidate discussions, it may be possible to extend this to other systems of review. Peer-reviewed scientific articles have been the subject of work on automated

writing feedback[132]. While articles on scientific advances require technical expertise, which is difficult to encode as criteria, the program committee may use our approach in combination with keywords or aspects (e.g., *originality* or *technical depth*) instead of policy invocations to gauge whether reviewers are consistent in how they apply this language to justify their reviews. This may still be limited to presentation details more than technical details.

Our own work would benefit from methods for better implicit policy invocations extraction. Aspect mining and other information extraction methods may prove useful here (we designed our methodology to select for a small, high-precision set of words rather than a high-recall set). Another possible application would be to provide accountability for content moderation, promoting trust in the communities who use it. We acknowledge that much of the work done here is observational or quasi-experimental. Before applications for this can be developed, further human evaluation, especially input from Wikipedia editors, is needed to contextualize these results. Work done on policy implementation in communities may be relevant to next steps for incorporating Wikipedia editors' input [179, 306]. This work may be best used to inform more questions about the *when* policies are invoked and their consequences. Making this work available for other language editions in addition to English Wikipedia, on which this work is based, is also a direction for future exploration.

4.6.2 Ethics statement

The data we use are publicly available through Wikipedia's API, and researchers are encouraged to use it for the purposes of developing the body of

knowledge about Wikipedia and for the development of tools aiding in the outcome.

One concern readers may have is whether the nomination process could be manipulated by bad-faith actors. Though one could picture that a hypothetical Wikipedia reviewer could try to apply our results in such a way, we purposefully choose to frame this work as a descriptive study of how reviewers already choose to *ground their comments* with policy invocations (a positive behavior). This may give not only reviewers, but also admin at FAC insight into trends in nomination discussion language. We also actively investigate where our policy invocation features fall short with comparisons to category, user, and content features. Though Wikipedia's policies and guidelines are constantly growing and being updated, we intend this work to be used in ways that encourage editors to engage constructively within the civil governance structures of Wikipedia and its processes. More work can also be done to see how much policy content alone can be used to infer article outcomes, though care should be taken to understand and account for the *review process*, not simply to predict the *review outcomes*.

4.7 Follow-up work

Modeling the discussion of the article's quality is important for understanding the FAC process itself. This is of interest to people who study peer-review processes such as PeerRead and perhaps those running the process itself. But what does it do for the nominators of the article? The nominators of the article may benefit from knowing what might be the biggest factors that could affect their

article's success. But nominators may prefer a more specific, immediate, and private means of receiving feedback.

This set of experiments showcase which factors contribute the most to the decision making process and how they reflect positively or negatively on the article. Follow-up work may tie this knowledge in with a writing feedback system. While a binary outcome score informs the user of the likelihood of their article being accepted overall, a score for likelihood of policy invocation can be used to give more nuanced feedback directly related to policies and guidelines.

4.8 Conclusion

The connection between **base document**, **context document**, and **inference task** was tested in this chapter. Context need not be a complete document nor does it need to be a document. Context may be introduced through a simple link or lexical keyword. And context make take on the form of an entire discussion, including those links and keywords.

This chapter also shows the traps of simply treating a text classification task as one of single text input and binary output. While inferring the outcome of an article from the article text itself is feasible, providing valuable feedback requires situating that model in context. Category, though correlated with outcome, is not as easily modifiable as the prose of an article. The majority of this chapter was not about classifying text better, in which case category would certainly help. Rather the majority of this chapter was spent evaluating how the content used as context in our later task was generated and in what was it was well-justified by policies and guidelines (the **context documents**) or biased by

subjective preferences.

To summarize, the contributions of this chapter are as follows:

- an **account** of non-text factors such as article category and discussion length in FAC reviews and how they relate to outcomes;
- the **quantification of FAC community values** via policies invocations; and
- an **overview** of applications to writing feedback.

CHAPTER 5

SUMMARIES AND SOURCE DOCUMENTS AS CONTEXT

In the other chapters of this dissertation, the source of data has been overwhelmingly Wikipedian in nature. Wikipedia is certainly one of the largest publicly accessible sources of data with the qualities the methods here require. But there are others. In this chapter, we discuss document-grounding of conversation in other contexts. This work was done at IBM Research in collaboration with Song Feng, Vera Q. Liao, and Luis Lastras.

In contrast to Wikipedia discussions, the conversations serving as our **base documents** are synchronous, meaning turns happen in real-time. The **context documents** in each of these instances is a summary describing important highlights **of** the conversation or is being described **by** the contents of the conversations. In both cases, the working assumption is that (1) procedural information is contained in the **context document** and (2) modeling procedural information well is valuable in an IT help context. Oftentimes, the (human) help desk agent must walk a customer through a procedure that is explained in the document. However, conversational elements are far more frequent. Therefore they are more likely to be generated than procedural steps by a dialogue model naively trained on the data.

Hence, we hypothesized that grounding the dialogue model using the procedural documents as **context documents** could improve the likelihood the model generating relevant procedural steps. The **inference task** is dialogue turn selection. In this experiment setting, the model takes one or more conversation turns as input, scores several candidate turns, and labels the highest scoring turn as the correct turn for the context. In our scenario, an auxiliary task is introduced

to score how likely a turn is to appear in whole or in part in an extractive summary of a dialogue. The intuition is that a summary will contain more critical procedural information and the model learns when to expect a procedure critical turn versus a conversational management turn. A limitation of this work is that having a good model of human-human dialogue is not necessarily the model for good human-human dialogue outcomes. Call quality outcomes are not available, but this complexity better reflects the dynamics of data and institutions outside of Wikipedia.

5.1 Introduction

There is an increasing need for dialogue agents particularly in the customer care domain, as customer service agents may be overwhelmed by a deluge of customers experiencing similar problems. One valuable resource for training automated agents is the existing goal-driven human-human dialogues typically between human agents and customers. However, using such data poses many challenges for dialogue models. It is both financially costly and time-consuming to label goal-oriented dialogue data at a larger scale, particularly for sensitive data. Even with an abundance of labels, dialogue modeling of human-human conversations is challenging.

This is especially true when dealing with longer conversations as the models typically suffer from the accumulation of errors. This *drifting* phenomenon holds even in models that learn to condition on history using various mechanisms such as attention. Thus, we hypothesize that a dialogue model may benefit from being informed by how critical a turn is to the global conversation

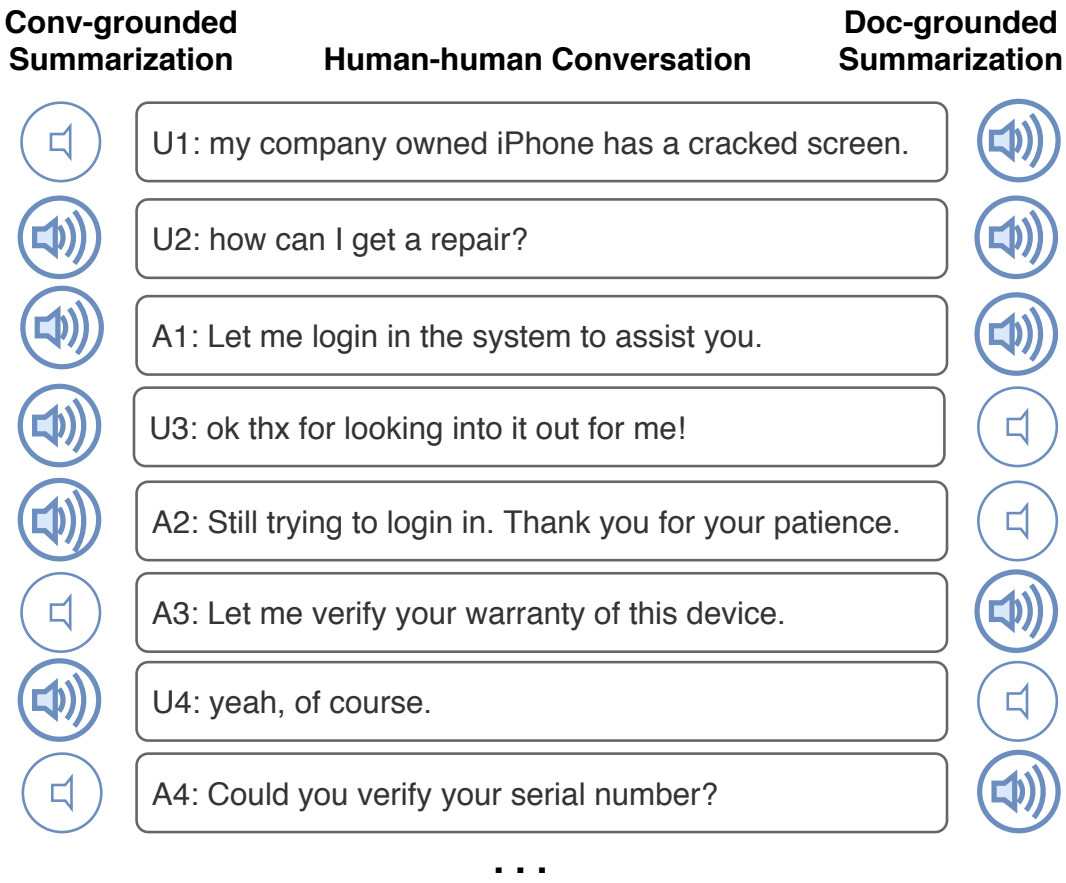


Figure 5.1: A sample conversation from HELPDESK with the predicted labels by conversation-grounded and document-grounded summarization respectively. Higher/lower volume icons illustrate how critical the turns are for fulfilling the task.

structure. We ask whether a model may better learn to predict the next turn in a conversation if it knows whether or not that turn is critical. For instance, Figure 1 presents the beginning of a conversation from our dataset, HELPDESK, of chat IT help chat-based interactions. Here, a (human) HELPDESK agent and an end user attempt to resolve problems with a broken device. There are certain dialogue behaviors, such as turn **A1**, **A2** and **U3** in Figure 5.1, that are common in the chat logs, but they may be unnecessary for fulfilling the goal-oriented tasks.

Inspired by the previous work on document summarization [205, 88], this

work applies summarization approaches to identify the critical turns for a given dialogue, as illustrated in Figure 5.1 (critical turns are indicated by "louder volume" icons), and utilizes the summarization information for distantly supervising a dialogue model based on existing human-human goal-oriented conversations. Specifically, we apply query-based summarization techniques to dialogue, which is more suited for extractive summarization [294, 205], than abstractive summarization [88]. In our work, summaries are generated for the sake of dialogue modeling, similar to work using structured summaries to inform models of dialogue[283]. The intuition is that using summarization as context in a dialogue turn-selection task can mitigate the effects of topic drift over the course of conversation. This could eventually aid automated dialogue agents in goal-oriented conversations. We generate extractive summaries using semi-supervised methods. The status of the utterance as extracted (or not) is then used as a binary label in a distantly supervised auxiliary task and as input in an oracle task.

Goal-oriented dialogue typically involves diagnostic processes for task completion, such as turn **A3** in Figure 5.1. In addition, they also include social acts (e.g., "hi, how are you today?") or clarifications (e.g., "where is the button again?"). In this work, we consider the former as conversation management turns [290] and the latter as information management turns. In our framework, we create different queries to a criteria-based semi-supervised extractive summarization algorithm for differentiating the conversation management turns and information management turns based on conversational data and the corresponding domain documents. In our case, conversation-based queries are effective in identifying conversation management turns such as **A1**, **A2**, **U3** and **U4**; while documentation-based queries are effective in identifying information

management turns for the summarization, such as **U1**, **U2**, **A3** and **A4**.

In particular, we expect that such a setting can be applied to other customer care domains since customer care services are often supplemented by online documentation. This documentation can provide (1) troubleshooting structures in the form of preconditions and solutions and (2) problem-resolution concepts. Many task-oriented dialogue systems and QA agents leverage a knowledge base or grounding documents, but our approach puts fewer constraints on the documents. In particular, the solution need not be contained within the documents, so long as solution-specific *concepts* are. These semantic concepts are used to form the queries and estimate the importance of utterances in a summary.

We evaluate the proposed framework on two types of real human-human dialogue data for the next-turn selection tasks. One corpus is the online chat logs of between HELPDESK agents and end users for troubleshooting IT issues; the others are meeting corpora AMI [47] and ICSI [122]. We show that including the summary generated by our framework as an auxiliary task improves next-turn selection accuracy significantly.

Our main contributions can be summarized as follows:

- We propose a framework to generate and leverage summaries of human-human conversations in a semi-supervised fashion using domain-specific documents (e.g., IT procedure manuals) for the purpose of training dialogue agents.
- We demonstrate the value of these labels by conducting experiments using summary labels in two settings, customer care and multi-party meetings.

- We provide a detailed analysis of the summaries to better understand how they aid conversation modeling.

5.2 Related work

Our framework relies on techniques derived from extractive summarization. It also largely related to human-human dialogue analysis and dialogue modeling.

5.2.1 Summarization tasks

Most of the prior work on text summarization mainly focused on single-speaker documents or monologues or single-speaker documents such as news articles and scientific publications [19, 200], as opposed to the dialogic data. For the work on dialogue data [47, 121, 88] the goal is to generate an overview or summary of the dialogue content. In contrast, our target output is a subset of utterances given a dialog. In this work, we also evaluate our approach using the online forum data. Though there has been some work on summarizing online forum threads for question-answering tasks [301], the applications to more extensive dialogue systems have not yet been explored. Our work is closely related to extractive summarization [205]. Unlike the prior attempts on leveraging extractive summary to improve group efficiency [137] or abstractive summarization [285], our goal is to obtain the summary of the conversations that is mostly useful for goal-oriented dialogue tasks.

5.2.2 Summarization

Summarization annotations naturally complement supervised approaches to summarization [44, 197]. This approach requires a large corpus of labeled instances, and the result is that much work has been done for document abstractive summarization but not so much for dialogue summarization. There is promising contemporaneous work on using structured summaries to inform dialogue models [283]. Unsupervised and semi-supervised approaches, particularly for extractive summarization, have flourished, though [163, 247]. In this work, we mainly investigate how to develop an unsupervised approach, which is more feasible to generalize to various customer care domains in practice.

5.2.3 Structure in human-human conversation

Our work is also related to previous work on the analysis of human-human conversations, especially as it relates to detecting structure via unsupervised methods. The approach to modeling conversation structure here is most similar to work on dialogue management and conversational initiative [282]. Inspired by this work, we seek to label turns as informational (i.e., procedural, in line with the IT troubleshooting document) or intended for dialogue management (e.g., initiating conversation or pausing for clarification). The unit of analysis in this work is the utterance. This is done instead of labeling the entire dialogue like other work [201]. Ours is also different than the utterance-level labeling or using dialogue acts for summarization [205, 88, 125] since our target output is not obtaining dialogue act tags or a summary of the dialogue content. Our targeted output is a subset of dialogue utterances that are the center for task

completion.

5.2.4 Dialogue agent models

Ultimately, we want to know if using summary labels to mark utterances as critical or not aids a dialogue model by providing a proxy signal for dialogue structure. There has been some success with this contemporaneously with this work [283]. Though the summaries might work for many types of dialogue agents, we experiment with goal-oriented dialogue agents are more constrained than chat agents but less constrained than task-oriented agents. Additional constraints have proven to be useful; much work has focused on training dialogue agents to use knowledge graphs [4, 102]. [4] is interesting in that the knowledge base is updated during gameplay, but the world is completely simulated. [102] provides an interesting scenario in which cooperative dialogue agents are trained in an agent-agent scenario, leveraging only a few human-human conversations. Our "knowledge corpus" lacks a clear knowledge base structure though, and our tasks are unconstrained [165]. As a result, we are more likely to benefit from models trained to incorporate monologic information as well as dialogue data [263].

5.3 Data

To evaluate our approach, we experiment with three corpora. The first (HELPDESK) is a proprietary corpus of chat logs between HELPDESK agents and customers. Because of the nature of the corpus, we cannot include the ex-

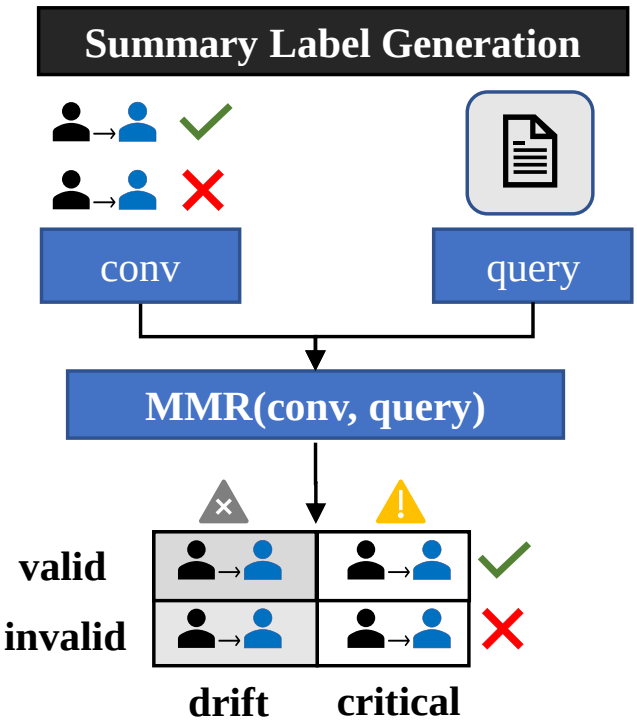


Figure 5.2: An overview of the query-based summarization system.

act stats. We also utilize a set of about 300 documents supporting the work of HELPDESK agents. While the chat logs and documentation may overlap in some topics, it is not expected that the documentation perfectly with all conversation problems and solutions. We merely expect that the documentation will allow us to strengthen priors on certain words that typically arise in the course of troubleshooting. There are no gold-label summaries for the chat.

The other two corpora are spoken dialog, the meeting corpora AMI and ICSI[1]. They are comprised of annotated transcripts from team meetings. The meetings have gold annotations for extractive summaries, abstractive summaries, reported in-meeting problems, actions taken, and dialogue acts. In comparison to HELPDESK with 2 participants per dialog, the conversations of the meeting corpora are multi-party with 4 to 10 participants per dialog.

[1] http://groups.inf.ed.ac.uk/ami/

For the HELPDESK dialogues, we tokenize individual messages. The shortest unit is either a sentence or an individual message, henceforth referred to as an utterance in the chat context. Documents are matched to the conversations using simple unigram content overlap. In practice, the document selection portion would be a separate application.

AMI and ICSI, which are spoken dialog, lack a strict speaker ordering because it is possible for multiple speakers to talk simultaneously. Each word is annotated with a start time and an end time, making it possible to chronologically order words. First though, we group words by their dialogue act segments rather than the default segmentation because the extractive summaries are tied to the dialogue act units. We then order the dialogue act segments ordered by the start time of the first word.

For both corpora, if adjacent segments have the same speaker, they are grouped as turns. Utterances are the units used for summarization, while turns are used for selection.

5.4 Methodology

Figure 5.2 presents the overview of our framework, which consists of the dialogue data processing and document data processing module, whose outputs are fed to the summarization module. In particular, the pipeline consists of the following subtasks: (1) determining salient concepts from conversations or their paired documents; (2) creating queries using a document or conversation representation and the extracted concepts; (3) labeling utterances as summary relevant or not via query-based extractive summarization.

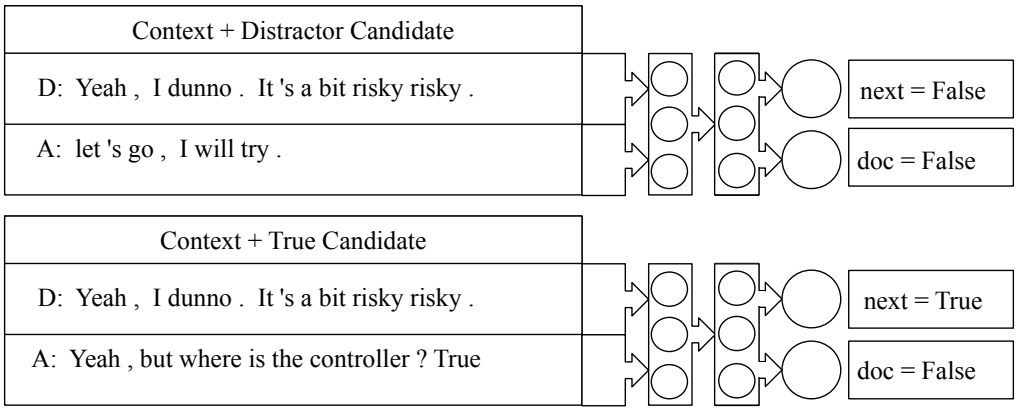


Figure 5.3: Above we show an example of our task set-up in the case of a distractor candidate and the case of the true candidate. Input is two embedded representations of a context turn and a candidate turn. The task is to predict the next turn, and the auxiliary task is document-grounded extractive summarization.

The reader may find the concept of the "query" confusing. What is the query and why was it chosen? To be clear, the query is either the troubleshooting document or the concatenation of a conversation sample set. Using the former query is equivalent to asking what utterances are similar to the troubleshooting document (i.e., what utterances are procedural). This query is known as the *doc* query. Using the latter query may be thought of as asking what utterances are likely to be found across all conversations (i.e., what utterances are for dialogue management). This will be denoted as a *conv* query. It is possible for an utterance to be considered both procedural and managerial.

While we considered selecting one summary to use as a "filtered" version of a conversation to train an agent from, we found it difficult to justify the total exclusion of turns dropped from the summary. Instead, we frame our next-turn selection task such that one may expect all utterances to be retained and the extractive summary labels are used to weakly supervise an auxiliary task.

Sample Utterances	Extracted for Summary gold	conv	doc	doc_k	Conv Progress
so we want it to be trendy um	**True**	False	False	False	0.2
So do not m make a new file .	**True**	False	False	False	0.2
but yes , we did ,	False	**True**	False	False	0.1
As you can see , this is the same tool bar uh as is located here .	False	**True**	False	False	0.2
And I count them like this .	False	False	**True**	False	0.6
I 've wrote down some examples here of what we can can speak about .	False	False	**True**	False	0.5
um because you have to make its prototype ,	False	False	False	**True**	0.9
Whether it looks like wood ,	False	False	False	**True**	0.3
maybe you have one or two stratig strategically placed lights	False	False	False	False	0.8
Then we still have some questions .	False	False	False	False	0.9
Um as a little training um I will ask Ruud first to draw uh uh your own animal on a new slide uh with uh a different colour and a different line width than the one uh now selected .	**True**	**True**	**True**	**True**	0.3
but because the case is transparent so it gives it a little bit of a glow , doesn't make it freaky .	**True**	**True**	**True**	**True**	0.8

Table 5.1: Above are example utterances and their associated summary memberships. Conversation progress indicates how early in the conversation an utterance appears (0.1 conv progress == first 10% of the conversation). The semi-supervised summaries we generate tend to have broader coverage of the conversation. Turns not in any summary are indicative of local drift.

5.4.1 Concept identification

In order to score utterances, one must first define concepts and their corresponding weights. Concepts may be a unigram, n-gram, or more abstract structures. For the purpose of these experiments, we equate concepts to keywords found in the grounding conversations or documents. We experiment with tf-idf weighting as a baseline and k-core decomposition weights as described in [270]. The value of k is selected chosen during the tuning process.

The latter approach weights words according to a centrality measure within a document graph. Our document graph is constructed by linking non-stopword unigrams with an edge if they occur within the same window of

text (12 words in total). Stopwords are specified using the NLTK list of 100 high-frequency English modals. Using the k-core decomposition of the resulting "textgraph", the words are weighted according to their core-rank. Higher core-ranks correspond to more central and possibly more important unigrams. For example, unigrams of rank 12 each have degree 12 or higher. We use the networkx implementation of a k-core decomposition to find the core ranks of words.

Note, while recent advances have been made in neural representations for long text sequences, we specifically opted for the above representations for two reasons: (1) our small document corpus may not support a neural model's data volume needs for accurate representation and (2) the above lend themselves to interpretability.

5.4.2 Summary optimization

Maximum marginal relevance (MMR), an established summarization objective, [45] is used to generate ordered summaries. The value of each utterance is conditioned only on past information in the conversation. Here, the set of utterances extracted by MMR is known as the summary S.

Let the following equation hold:

$$q_{sim} = sim_1(u_i, Q) \tag{5.1}$$

where q_{sim} is the similarity between the i^{th} utterance and the query Q.

$$r_{cost} = \max_{u_j \in S} \; sim_2(u_i, u_j) \tag{5.2}$$

where r_{cost} is the redundancy cost as characterized by the similarity between utterance t_i and the most similar utterance already included in the summary, t_j. Then we can define MMR as follows:

$$MMR := \text{a}rgmax_{u_i \in R \setminus S}[\lambda\, q_{sim} - (1 - \lambda)\, r_{cost}] \tag{5.3}$$

We use cosine similarity for sim_1 and sim_2, and the above described weights are used to create vectors for the query and the text. For each unchecked utterance in the remaining conversation R, we include that utterance in the summary S if that utterance has an MMR score above threshold. For the threshold, we use $\lambda = 0.75$ to more heavily bias the summaries towards their respective queries, at the cost of higher redundancy.

Two types of queries are evaluated: conversational and document. Conversational queries result in summaries that resemble qualities of conversations and agent behavior (*e.g.*, "hi. how may I help you today?"). Qualitatively, document queries result in dialogue summaries that more resemble the procedural and informative qualities of trouble-shooting documents (*e.g.*, "my email client keeps crashing after an OS update"). To use a document query, conversations are first matched to the most similar document in the document set. For AMI and ICSI, we used abstractive summaries as documents. For the remainder of the paper, these two types of summaries will be referred to as **conv** and **doc**, respectively. We only use the k-core decomposition weights with the documents; these summaries are denoted as $\textbf{doc}_k$. AMI and ICSI include gold annotation summaries; these summaries are denoted as **gold**. Summarization requires removing utterances. We experiment with two levels of granularity, retaining 33% of all utterances in the conversation and retaining 66% of utterances. But we found the differences minor. Unless otherwise noted, the 66% summaries are

used.

Table 5.1 lists several contrasting examples of utterances included in one type of summary, all summaries, or none. We anecdotally note that the rare utterance included in all summaries tend to be longer. This is likely because more words make it more likely to contain concepts all summaries weigh highly.

Note, we use these semi-supervised extractive summary assignments **only as labels, never as input**.

5.5 Evaluation

We hypothesize that leveraging summary information may provide useful information with respect to dialogue flow. To evaluate this hypothesis, we employ the summary annotations as weak supervision labels in a multi-task prediction setting. The primary task is next-turn selection: given a candidate turn and a context, a model must label that turn as valid or not. The auxiliary tasks are to predict whether or not that turn, regardless of whether it is next or not, is a summary turn in some conversation. Though we experimented with scoring approaches, we found that a boolean label scheme resulted in better results. This may be a result of the deterioration in negative sample similarity is so sharp it is more akin to a discrete labeling task. We discuss this more later.

We formalize the task below as a next-turn selection task with a context x and a candidate c and a set of summary turns S as follows:

$$\text{argmax}_\theta p(y = c, c \in S | x, c; \theta) \tag{5.4}$$

The parameters θ are the parameters of our model to be optimized such that the probability is correlated with the likelihood that the candidate is the true turn y and that the candidate turn is in the summary set of "query-relevant" turns S. This will bias the model towards selecting turns that are either conversation management turns (if using the conversation query) or more procedural turns (if using the document query). Given that conversational turns are simply more likely because there are more types of procedures than there are types of conversations in our dataset, the document queries (doc and doc_k) are expected to be more helpful at recovering informative (and therefore less likely) turns.

Candidate mining and instance sampling. To obtain the candidates, we sample up to 100 random turns from other conversations for HELPDESK and up to 50 turns for AMI and ICSI. We calculate the cosine similarity between the tf-idf weighted n-gram representation of the true next turn and all sampled candidate turns. The 7 most similar candidates are used for comparison. If any of these candidates exceeds a similarity threshold (HELPDESK threshold = 0.5, AMI/ICSI threshold = 0.75), it is considered a positive sample. In our internal corpus, all candidates share the same speaker (the agent), as other-speaker turns (from the client) tend to be easily distinguished from the true turn. We sample 10% of all instances for HELPDESK and AMI and 25% all instances for ICSI as sampling yielded fewer instances. Instances were grouped by conversation, and conversations were divided into train and held-out. A sample of train conversations were used for validation. Random seeds were fixed, and results are averaged over three seed choices (seed=1,2,3).

Text representation and model. For our text representation, we leverage recent work in transformer-based language modeling. The last layer of the base BERT embedding model [65] is used to represent the candidate text c and the context text x, with up to 400 tokens or 8 utterances, whichever is shorter. These representations are then fed to a multi-layer feedforward neural network, optimized using SGD. Each layer has 1000 units and a linear activation function. Initial experiments with fewer layers led us to conclude that more layers were preferable for the HELPDESK corpus. But we found only one hidden layer to be necessary for the other corpora. The learning rate is tuned over the values 10^{-4} to 10^{-1}, with 0.001 being the best learning rate for HELPDESK and 0.01 working well for AMI and ICSI. Binary cross-entropy with logits is used as the loss criterion. The models are trained for at most 100 epochs, but early stopping is used when loss ceases to decrease for 3 to 10 epochs. Usually, the models converged in 10 to 20 epochs. Figure 5.3 shows our experiment set-up and an example context and candidate from AMI.

5.5.1 Next-turn selection results

Next-turn selection results are shown in Table 5.2. These scores are the calculated precision-@1, assuming that the true candidate is the best candidate, even if there are equivalent or near-equivalent other candidates. We vary both the granularity of the summaries and the weight of summary concepts based on conversations, documents, and k-core decomposition of documents.

These results show that using document-grounded summary prediction as an auxiliary task produces better next-turn selection results than a model

Corpus	Model	Prec @1	Δ Prec @1
All	Random	0.1250	0.0000
AMI	BERT	0.3391	0.2141
	+conv	0.3407	0.2157
	+doc	**0.4164**	**0.2914**
	+doc$_k$	0.3359	0.2109
	+gold	0.4009	0.2759
ICSI	BERT	0.4868	0.3618
	+conv	0.3982	0.2732
	+doc	0.3316	0.2066
	+doc$_k$	0.2380	0.2380
	+gold	**0.4965**	**0.3715**
HELPDESK	BERT	0.1653	0.0403
	+conv	0.1727	0.0477
	+doc	**0.2085**	**0.0835**
	+doc$_k$	0.1569	0.0319

Table 5.2: Top-candidate accuracy (precision @1) for next-turn selection task given 8 candidates averaged over 3 instance samples. Performance above random baseline (0.1250) is shown in Δ Prec @1 column.

trained with a conversation-grounded summary auxiliary task or no auxiliary task. Surprisingly, the document grounding that uses high-frequency words and tf-idf weights yields the best results, while document grounding that uses k-core decomposition weights and no high-frequency words tends to perform even worse than the baseline model. This suggests that the advantage of document-grounding lies in a functional / procedural style of communication, which may best be communicated by high-frequency function words. Other work on conversation modeling that shows de-emphasizing content words can improve performance in dialogue tasks [125].

Summary 1	Summary 2	AMI	ICSI
gold	conv (0.33)	0.06	**0.09**
gold	doc (0.33)	**0.12**	0.09
gold	doc_k (0.33)	**0.25**	0.13
conv (0.33)	doc (0.33)	0.38	0.55
conv (0.33)	doc_k (0.33)	0.10	0.33
doc (0.33)	doc_k (0.33)	0.18	0.35
gold	conv (0.66)	**0.14**	0.07
gold	doc (0.66)	**0.17**	0.08
gold	doc_k (0.66)	**0.27**	0.12
conv (0.66)	doc (0.66)	0.61	0.66
conv (0.66)	doc_k (0.66)	0.27	0.34
doc (0.66)	doc_k (0.66)	0.36	0.41

Table 5.3: Overlap between between summaries in AMI and ICSI corpora. AMI's Unsupervised summaries tend to overlap more with gold summaries than ICSI's, and **doc** summaries overlap more with **gold** than **conv**.

5.5.2 Unsupervised vs. gold-annotated auxiliary tasks

Not only have we shown the value of gold-annotation summaries, but applying our approach to a domain-specific internal corpus reveals that unsupervised summary annotations may be helpful as auxiliary tasks. This is key when handling certain types of data as summary annotations for conversations are rare and high inter-annotator agreement is difficult to achieve. Obtaining qualified annotators for domain-specific and/or highly sensitive data may be particularly difficult. While one can still learn using the gold annotation summaries (as shown with AMI and ICSI), our approach may also leverage unsupervised summaries as labels for an auxiliary task.

Using document-grounded queries to guide the production of unsupervised summaries increases overlap between gold-annotated summaries and generated summaries. This is most pronounced when using k-core annotations. There exist models for jointly learning to model both conversation structure

(at the dialogue act level) and topic throughout a conversation. But our approach is orthogonal in that we do not focus on a particular definition of topic but rather relevance. However, leveraging k-core annotations in the case of the internal documents led to *decreased performance*. Possible reasons for this include poor alignment between documentation and conversations or less need for high-relevance concepts and more need for functional/procedural information.

5.5.3 Summary analysis

We observe that most types of summaries do affect the prediction outcomes. But each summary is generated differently and may inform the primary next-turn selection task differently. To better understand possible differences that we might exploit later, we investigate the overlap between different summary types and their unique linguistic properties. Table 5.3 shows the overlap between different types of summaries (calculated by Jaccard similarity). We emphasize overlap rather than precision or recall as the gold-label annotation summary is not necessarily the most useful summary for training dialogue agents.

Disparity analysis. Gold label summaries are more similar to summaries extracted with the help of document-grounded concepts, particularly when text is represented using k-core decomposition. The two tf-idf based summaries are both more similar to one another than to either two summaries. Summary overlap increases across all pairs when the number of unsupervised summaries increases, but the amount by which overlap increases varies significantly. Though the gold labels only increase in similarity by at most 0.08 with the larger 0.66

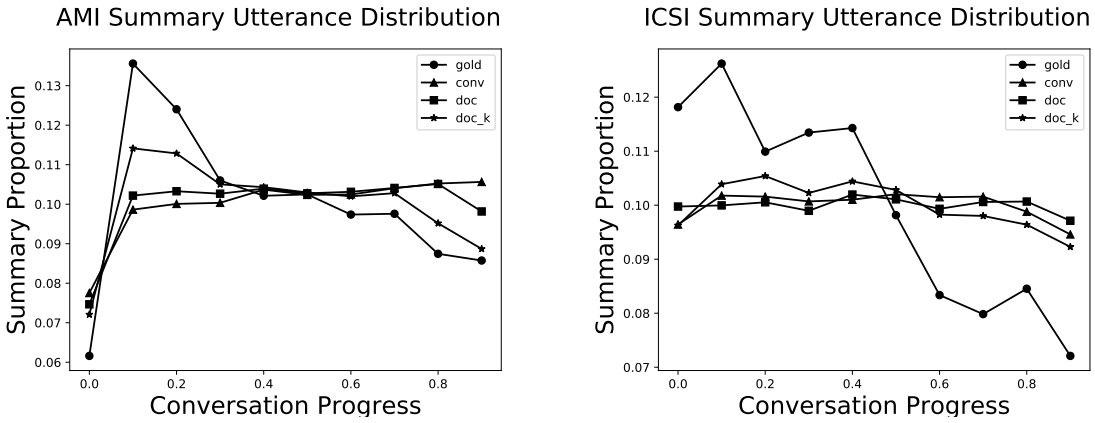


Figure 5.4: Above is a chart of the number of summary utterances at a particular part of conversation. Utterances at x=0.1 and x=0.9 are at the beginning and end of the conversation, respectively. **gold** summary utterances are skewed towards the beginning.

summaries, all unsupervised summaries increase in similarity to one another significantly, in some cases doubling in similarity.

The magnitude of overlap between summaries gives us some insight into which summaries might be used to substitute the gold-label summary when no gold-annotations are provided for a corpus. But how do these summaries overlap? In Figure 5.4, the distribution of summary utterances across conversation segments is plotted for (a) AMI and (b) ICSI. Conversation segments are 10% of the length of the entire conversation. In both corpora, the gold-label summary utterances are concentrated toward the beginning of the conversation. The document-grounded, $\mathbf{doc}_k$ summaries follow a similar distribution trend as the GOLD summaries, as would be expected from the overlap scores shown in Table 5.3. The auxiliary summarization models' relatively poor performance on ICSI may be attributed in part to poorer unsupervised summary quality. All generated summaries show less dramatic fluctuation in utterance density across the course of the conversation.

There is a large disparity between **gold** summaries and **conv** summaries. While gold-annotation summaries are themselves valuable for training summarization models, the next-turn selection task may benefit from having two contrasting types of summaries. This hypothesis was evaluated by training a model with two auxiliary summary prediction tasks: the **gold** and the **conv** summary prediction. Using these two labels in auxiliary tasks, we see *a 6% absolute gain (precision @1 = 0.4720)* over the best single-auxiliary task model (**doc**) in the AMI next-turn selection task. Similar experiments with ICSI show decreased performance, which is likely related to how ICSI models benefit less from summaries overall.

Linguistic analysis. What are the linguistic disparities between the different types of summaries? To investigate, we perform a fighting words analysis [192], comparing utterances from each sample type to utterances of all other summary types.

The **gold** annotation and $\mathbf{doc}_k$ summaries are primarily associated with more content words (adjectives and nouns), while the **conv** and **doc** summaries contain more backchannels, pronouns, and function words. Together, different summaries may inform the multitasking model with both topical and functional information. These distinctions align with [290]'s ideas on information management (information quality) and conversation management (plan quality). They observe that turns marked as *repetition* and *summaries* more frequently occur within topic. These turns are also more likely to be filtered out by our summary methods, so the turns in our extractive summaries may occur more frequently at topic boundaries. A model aware of summary information may therefore be able to negotiate topic switches better than one without.

5.6 Conclusion of experiments

In this work, we present a framework that minimizes the effects of conversation drift on dialogue agent training by using auxiliary summarization tasks. We evaluate the proposed approach on three datasets for the next-turn selection tasks. The results show that including extractive summary prediction as an auxiliary task improves next-turn selection accuracy of fine-tuned transformer models, here represented by BERT. This suggests that the summarization is related to conversation structure. Grounding summaries in a set of domain-specific documents can further improve next-turn selection accuracy, as shown for two of our three corpora.

Surprisingly, document summaries benefited next-turn selection models more than gold-annotation summaries in the AMI corpus. Furthermore, combinations of two worse performing auxiliary summary tasks beat out the best performing single auxiliary task on AMI. Overall, the benefit of the summarization auxiliary tasks seems to depend on the quality of summarization. ICSI benefits only from the **gold** auxiliary labels, and its associated unsupervised summary labels also show far less overlap with the **gold** set. For future work, we intend to investigate other interactions between documents and conversation and ways of mitigating conversation drift in dialogue models.

5.7 Conclusion

This chapter is distinct from the others in that our **base document** is dialogue. Given that our dialogues do not necessarily have clearly defined outcomes, the

research questions and **inference tasks** are related more to dialogue structure. It aligns with work on document grounding for chatbot research, but it is distinct in that our primary interest is in modeling the underlying topical structures and behaviors of human-human dialogue.

The **context documents** enable the use of prior knowledge or expert observation while not restricting those experts to constrained annotation forms that would necessitate a separate annotation process. In this way, we can actually bias our data in a way that does better reflect the values of the stakeholders (e.g., greater adherence to procedural information). Not all bias is bad [204]. We argue that the **context documents** in combination with semi-supervised machine learning is preferable because (1) annotated data for conversational exchanges are expensive and time-consuming to obtain and (2) it allows a more flexible understanding dialogue that is better fit for the domain thanks to the expert input.

However, it is clear that in the case of ICSI, context was not necessarily helpful. This may be for a number of reasons, but most notable is that ICSI conversations are natural meetings as opposed to the more constrained environments of IT help chat logs and the AMI lab-based experiments. It may be necessary to try other (yet unknown) queries to capture conversation dynamics better.

Obtaining sufficient context documents is a fundamental challenge though. Summaries of conversations might be made available through news or reporting, but those require special licensing typically. Using documents that serve as a template for conversation may be more tenable, and this is where document grounding typically finds its documents. Even then, the documents may not cover all possible variations of a problem.

This work was documented in part by the patent *Adapting dialog models by relevance value for concepts to complete a task* (US11443119B2). In summary, we achieved the following:

- improved dialogue model performance;
- introduced a generalized approach to using document context; and
- used extractive summary membership as dialogue management labels.

CHAPTER 6

CONCLUSIONS AND DIRECTIONS

To recap, it was posited that incorporating context documents could aid modeling discourse using semi-supervised methods. So long as the context documents and the base documents shared vocabulary, contained distinct vocabulary, or had structural differences, the context documents can be leveraged to inform related inference tasks. The inference tasks chosen throughout this work were chosen to help probe the discourse structure of the base document. This work is particularly relevant to domains where obtaining annotations is expensive or time-consuming.

6.1 Fundamental limitations

In all the examples we have shown, there is the assumption of contextual documents written by fairly well-studied people or supported by references or an internet community. The transformative properties of context are an important caveat. As a thought experiment, we consider the case of using the **base document** as our **context document** after shuffling the content of the document. Technically, this satisfies the condition that the structure be different. While this may improve the drawing of connections across long-form documents, it does not introduce any new information that constitutes synthesis. It is equivalent to a (poorly) trained attention model.

Consider also the case of substituting tokens in the **base document** with unique non-word tokens to form a new **context document**. This satisfies the vocabulary separation condition, but new information is not introduced, just pos-

sible reinterpretations of the information. This is not unlike a masked language model where the word's representation depends on the immediate context.

Finally, consider the case where a randomly sampled set of turns or sentences are labeled for extraction much in the manner of a baseline in Chapter 5. There, the information introduced is a sort of shortcut.

In all these cases, one problem remains: The naive choice of context only leads to an exponential increase in parameters. While approaches from the past decade of NLP research have introduced models that learn to incorporate context, it requires increasing amounts of data and compute. While that is excellent for training so-called foundation models that can be adapted to other uses, the fundamental thesis of this work is that the introduction of skillfully selected context can improve model output at a fraction of the compute. Even when compute is available, data may not be. The foundation models trained on large general corpora adapt poorly to more narrow domains. In contrast, the limitation of this work is that it is best suited for more narrow domains in which an expert group is highly active.

6.2 A summary of contributions

The reader may recall three themes that were set forth in the Chapter 2. It is the author's hope that the following work has demonstrated the importance and/or effectiveness of the following:

- **Less than full supervision.** Manually annotating data for fully supervised natural language processing tasks is a fraught process that may not be vi-

able due to privacy reasons or expertise gaps. While full supervision has advanced the field, it typically requires the development of benchmark datasets that may be poorly suited to help in adapting models to new domains [35, 210]. This work demonstrated an approach for including text context in social natural language processing tasks in Wikipedia and beyond.

- **Value-sensitivity.** This work also provides an approach to including implicitly and explicitly stated values in social NLP tasks, particularly those related to Wikipedia. Critical to this approach was linking a **base document** to a **context document** and assessing the effects in an *inference task*. In Chapter 3, parallel documents from other language editions of Wikipedia were used to assess the extent to which articles adhered to the Wikipedia value of Neutral Point of View. In Chapter 4, we studied the context of writing assessment in Wikipedia's Featured Article nominations. And in Chapter 5, IT documents were used to improve the modeling of help desk conversations. The hope is that it empowers the stakeholders to more easily audit machine learning models trained on their data.
- **Bias and accountability.** To aid work on mapping potential sources of biases in the common data sources (e.g., Wikipedia), this work emphasizes understanding the social context and possible influences on the data generation process. Chapter 3 highlights potential geopolitical bias, while Chapter 4 identifies possible cases of reviewer bias. Chapter 5 focuses more on intentionally biasing the model towards learning the less frequent but problem relevant procedural information and away from learning the more frequent but less informative social content.

A primary theme in this work is the reversal of norms in natural language

processing. Instead of seeking to overcome gaps in data, our work seeks to understand what these gaps may imply about the datasets in which they occur. The result is a framework for approaching task framing in applied natural language processing. What is usually considered blemishes in the NLP research of late — small specific datasets, discrepancies in data alignment, and smaller models — are here considered advantages to be used for more value-sensitive language modeling.

To be clear, this work benefits substantially from large language modeling that is pre-trained on a large diverse corpus. But to achieve more domain-sensitive objectives and to audit data and models successfully, we treat tasks not as pairs of (input, output) but as tuples of (input, social context, output). The additional context is not always available or of high enough quality to lend itself to modeling. But we find our approach is particularly amenable to corpora comprised of significant amounts of supporting documents.

APPENDIX A

GLOSSARY

Word	**Definition**
base document	The document that is the canonical input in a given task.
context document	A document related to a base document that encodes domain knowledge and stakeholder values.
inference task	A definition of a process in which a statistical machine learning model is used to label text input. There is no assumption that the input precedes or follows the output chronologically or otherwise.
practitioner	A person who implements natural language processing models and pipelines on behalf of or as a stakeholder.
stakeholder	A person or group with a vested interest in applications of natural language processing.

Chapter 3	
Word	**Definition**
battle	an event described by a Wikipedia article that is identified as a member of the World War I or World War II battle categories in English Wikipedia
cluster	in this context, a group of tuples (fewer than 16) that are semantically similar
corroboration	the agreement between two texts that a specific event has occurred
tuple	a (subject, relation, object) ordered grouping that is extracted from article text

Chapter 4	
Word	**Definition**
editor	any user who edits a Wikipedia page
FAC	Featured Article Candidate review process that evaluates articles for showcasing on English Wikipedia's front page.
nomination	an article which one or more editors (usually a primary contributor) that has been posted for consideration
reviewer	any user who is not one of the named nominating editors who provides feedback on the article and makes an argument for its promotion or rejection

Chapter 5	
Word	**Definition**
AMI	a corpus of multiparty spoken conversations coordinated in an in-lab experiment setting; participants are assigned to roles in a team, and each conversation is one of a sequence with that team.
ICSI	a corpus of natural spoken conversations taken from transcripts of meetings
HELPDESK	a corpus of internal IT chat logs from client-agent interactions at a corporation
MMR	maximal marginal relevance, a criteria for iteratively updating a summary set in an extractive summarization setting; it maximizes the relevance of the set while minimizing the redudancy.
query	a troubleshooting document or the concatenation of a conversation sample used to construct a summary set
summary	a set of utterances from a given conversation denoted as S

APPENDIX B

APPENDIX FOR CHAPTER 3

This appendix attempts to introduce additional experiments and data statistics. First, we introduce a broad analysis of content variation across language editions. Second, we consider patterns of user interactions and contributions that can inform our speculation regarding the mechanisms that lead to the different content in different language editions.

B.1 Language variation

There is clearly variation in language across the language editions even in translation. This is expected. But the nature of the variation is important to understanding possible alternative reasons for our results. We present a brief bayesian fightin' words analysis for transparency and additional insight.

In Table B.1, the words significantly associated with particular language editions are presented. It is clear that the EN WWII data contains incomplete translations, which is surprising given that we expected it to be most similar to its original version. Outside of translation effects, there does appear to be a greater emphasis on casualties in the EN language edition. Linguistically, the presence of the word "were" suggests a particular linguistic pattern.

In contrast, the DE language edition is more concerned with groupings of soldiers (e.g., "division", "wing", "reserve", "corps"). "Should" and "could" are function words that are distinctly DE in this context. This may be a result of linguistic patterns in the German language or it could be related to characterizing the groups of soldiers mentioned before.

WWI				WWII			
DE	EN	FR	IT	DE	EN	FR	IT
division	casualties	will	general	should	para	will	general
wing	were	must	category	troops	fueron	must	armored
reserve	line	region	sector	could	force	during	marshal
corps	would	located	enemy	section	casualties	takes	sector
section	fifth	somme	departments	corps	alemanes	latter	situation
under	fourth	armenian	team	wing	como	battle	forces
group	light	kingdom	giardino	area	battalion	offensive	maneuver
should	force	battle	situation	reserve	ejército	begins	panzerkorps
could	ottoman	takes	operations	army	fuerzas	place	field
area	armies	type	force	leadership	were	region	departments
here	ordered	during	powerful	able	japoneses	vice	mechanized
empire	defenses	soyécourt	japanese	associations	también	kingdom	team
leadership	beatty	offensive	cary	again	would	their	would
already	captured	come	lanrezac	soldiers	habían	operation	anglo
iran	second	confrontation	difficulty	connection	ottoman	admiral	category

Table B.1: Most indicative words across language editions as computed by "fightin' words" Bayesian analysis [192].

The FR language edition is primarily concerned with the time and place of battle (e.g., "region", "located", "during", "begins", "place"). The function words that stand out are "will" and "must". The former is interesting as it suggests future tense and possibly inevitability, which is not expected in an article written about a past event. But there could also be conflation with another meaning of the word "will" (definition: *the faculty by which a person decides on and initiates action.*) The use of "must" indicates a sense of inevitability though.

Finally, the IT language edition seems more concerned with leadership (e.g., "general", "lanrezac", "cary", "marshal", "giardino") and calvary/equipment (e.g., "armored", "panzerkorps", "mechanized"). There seems to be a slight lexical difference, possibly due to translation. For example, the DE edition uses "section" while the IT edition uses "sector". But notably, the IT language edition uses "enemy" more than other language editions.

B.2 User behavior

In the preceding experiments, the articles are described proxies for the opinions of a language community. As they are collaboratively written, one may indeed expect the article to reflect a mix of the opinions of the community of editors. But this overlooks the proportional contributions of specific editors and the influence of non-editors on the articles content. To understand the social dynamics that might influence the article narrative in one direction or another, an additional quantitative analysis of the associated talk pages in English is presented here.

This work focuses on English Wikipedia's talk pages to change the axis of analysis (talk page and article vs. article and article). This also enables a closer assessment by the authors of this work, whose primary language is English.

B.2.1 Users and user contributions

The users in this study are drawn from editors of the battle articles and contributors to the corresponding talk pages. While there is overlap, not all editors participate on talk pages and not all talk page contributors are editors. For each user found in the union of both sets of contributors, a profile of their contributions to other Wikipedia editors is downloaded. While it is possible for a user to contribute to many different Wikipedia language editors (e.g., via a script), more contributions to other language editions may indicate linguistic preferences.

Again, it is stressed that language is not nationality. But a willingness or ability to contribute to another language edition of Wikipedia suggests that lan-

guage may be correlated with nationality or at least knowledge of one or more countries associated with that language. With this knowledge of language communities a user frequents, it is possible then to ask to what extent a language community might influence the composition of the article. In particular, contributions to references are investigated here.

Table B.2 shows the 20 most prolific comment writers in the battle articles' talk pages. We obscure the usernames to protect the users' privacy even though this is publicly available information. There is a wide range of language editions in the top editors' other language edition contributions. But this table does not indicate the magnitude of the edits to those language editions compared to that of the contributions to the English language edition. It is possible that these edits to other language editions are minimal (e.g., editing an image used in multiple language editions) or extensive (e.g., writing an entire article in another language).

The top contributor to English battle talk pages across WWI and WWII is the user Kei***, but this contributor had no significant edits to other language editions at the time of data collection. See Figure B.1 for an overview of the number of comment contributions. WWI battle talk pages are clearly dominated by one editor with 3 or so editors trailing. Meanwhile, WWII battle talk page conversations are more evenly shared by the top 3 editors. This is likely because the larger number of articles in WWII make it more difficult for a single editor to dominate.

WWI		WWII	
User	**Non-en wiki**	**User**	**Non-en wiki**
Kei***	N/A	Kei***	N/A
San***	it.wikipedia.org	Tre***	it.wikipedia.org
Tir***	sv.wikipedia.org	Eni***	es.wikipedia.org
Sim***	es.wikipedia.org	Wan***	sv.wikipedia.org
See***	ru.wikipedia.org	Mrg***	ru.wikipedia.org
Eag***	ceb.wikipedia.org	Wdf***	de.wikipedia.org
Par***	de.wikipedia.org	DMo***	ceb.wikipedia.org
Alp***	zh.wikipedia.org	Par***	zh.wikipedia.org
Pau***	nl.wikipedia.org	Pau***	nl.wikipedia.org
InternetArchiveBot	pt.wikipedia.org	Dap***	fr.wikipedia.org
Lab***	fr.wikipedia.org	InternetArchiveBot	pt.wikipedia.org
Ada***	ja.wikipedia.org	Dam***	vi.wikipedia.org
Tho***	vi.wikipedia.org	The***	war.wikipedia.org
Ano***	pl.wikipedia.org	Bin***	uk.wikipedia.org
Ros***	war.wikipedia.org	Fol***	ja.wikipedia.org
Hud***	sr.wikipedia.org	SineBot	pl.wikipedia.org
Hub***	uk.wikipedia.org	Gre***	no.wikipedia.org
Pat***	tr.wikipedia.org	Kra***	sr.wikipedia.org
Aus***	fa.wikipedia.org	Mar***	fa.wikipedia.org
SineBot	no.wikipedia.org	Cap***	tr.wikipedia.org

Table B.2: Top contributors and the non-English Wikipedia language edition they contribute to the most.

B.2.2 Monoglot and polyglot interactions

User statistics are aggregated according to editor/commentor role. Editors make changes to the articles, while commentors have no history of editing the article. and top two language edition contributions. We specify top two because a number of editors may have majority of their edits in English Wikipedia, but they regularly contribute to other language editions. This is expected as we focus only on English edition talk pages. Table B.3 indicates an overwhelming number of interactions occur between monoglot (English only) participants. English language editors are also more active than commentors. Editors who have participated in the German and French language editions of Wikipedia are the most prominent polyglot editors.

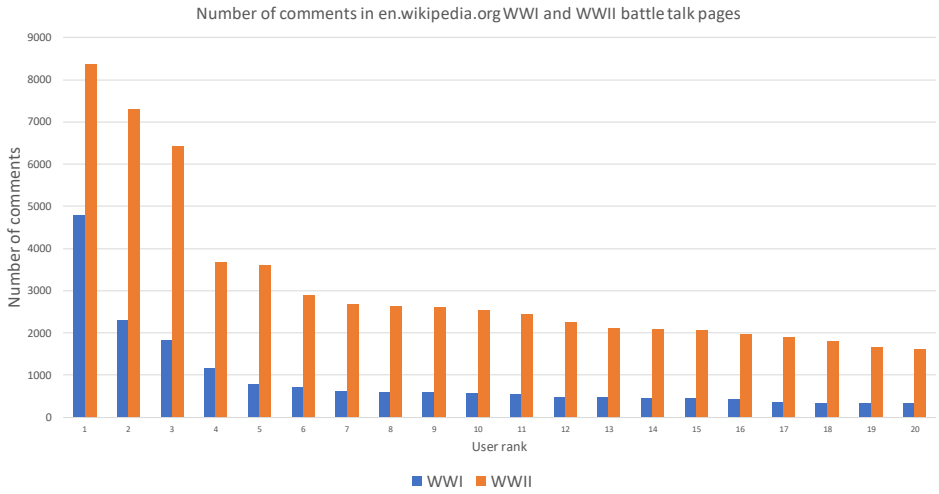


Figure B.1: Number of comments of ranked contributors.

	WWI		WWII	
Rank	Interaction	Count	Interaction	Count
1	Editor_en and Editor_en	3831	Editor_en and Editor_en	17601
2	(Starting) Editor_en	2067	(Starting) Editor_en	8613
3	Commentor and Editor_en	1131	Commentor and Editor_en	6666
4	Editor_en and Editor_en+fr	930	(Starting) Commentor	4479
5	Editor_en+fr and Editor_en	873	Editor_en and Commentor	4287
6	Editor_en and Commentor	789	Editor_en and Editor_de+en	4077
7	(Starting) Commentor	705	Editor_de+en and Editor_en	3771
8	Editor_de+en and Editor_en	516	(Starting) Editor_de+en	1929
9	Editor_en and Editor_de+en	480	Editor_en and Editor_en+fr	1782
10	(Starting) Editor_en+fr	351	Editor_en+fr and Editor_en	1719

Table B.3: Top 10 interaction types by role and language edition contributions.

The conversational initiative seems to lie with the English monoglot editors, as they initiate a large number of conversations. This is indicated by the (Starting) label. While English monoglot and French language editor interactions are more prominent in WWI battle articles, English monoglots and German language editors interact more in WWII battle articles.

BIBLIOGRAPHY

[1] Eytan Adar, Michael Skinner, and Daniel S. Weld. Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 94–103, New York, NY, USA, 2009. Association for Computing Machinery.

[2] B. Thomas Adler and Luca de Alfaro. A Content-driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA, 2007. ACM.

[3] Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312, 2016.

[4] Prithviraj Ammanabrolu and Mark O Riedl. Playing text-adventure games with graph-based deep reinforcement learning. *arXiv preprint arXiv:1812.01628*, 2018.

[5] Maik Anderka, Benno Stein, and Nedim Lipka. Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of SIGIR*, pages 981–990, 2012.

[6] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *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 344–354, 2015.

[7] Jun Araki and Teruko Mitamura. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, 2018.

[8] Ofer Arazy, Oded Nov, and Felipe Ortega. The [Wikipedia] world is not flat: on the organizational structure of online production communities. In *ECIS*, 2014.

[9] Ofer Arazy, Oded Nov, Raymond Patterson, and Lisa Yeo. Information quality in wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4):71–98, 2011.

[10] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. Functional roles and career paths in wikipedia. In *Proceedings of CSCW*, pages 1092–1105, 2015.

[11] Wenceslao Arroyo-Machado, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. Science through wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, 15(2):1–20, 02 2020.

[12] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[13] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*, 2019.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[15] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, 2013.

[16] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: Bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1075–1084, New York, NY, USA, 2012. Association for Computing Machinery.

[17] Robert A. Baron. Negative effects of destructive criticism: Impact on conflict, self-efficacy, and task performance. *Journal of Applied Psychology*, 73(2):199–207, 1988.

[18] Robert A. Baron. Countering the effects of destructive criticism: The relative efficacy of four interventions. *Journal of Applied Psychology*, 75(3):235–245, 1990.

[19] Araly Barrera and Rakesh Verma. Combining syntax and semantics for automatic extractive single-document summarization. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 366–377. Springer, 2012.

[20] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D.

Vohs. Bad is stronger than good. *Review of General Psychology*, 5(4):323–370, December 2001.

[21] Linda H. Bearinger. Beyond objective and balanced: Writing constructive manuscript reviews. *Research in Nursing & Health*, 29(2):71–73.

[22] Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

[23] Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[24] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Workshop on Language in Social Media*, pages 48–57. ACL, June 2011.

[25] Luciana Benotti and Patrick Blackburn. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, 2021.

[26] Ivan Beschastnikh, Travis Kriplean, and David W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *Proceedings of ICWSM*, 2008.

[27] Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, 2014.

[28] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[29] Joshua E Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096, 2008.

[30] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[31] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[32] Jonathan Borak and Frederick R Sidell. Agents of chemical warfare: sulfur mustard. *Annals of emergency medicine*, 21(3):303–308, 1992.

[33] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[34] Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. Societal controversies in wikipedia articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 193–196, New York, NY, USA, 2015. Association for Computing Machinery.

[35] Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics.

[36] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[37] Susan E Brennan. The grounding problem in conversations with and through computers. *Social and cognitive approaches to interpersonal communication*, pages 201–225, 1998.

[38] Matt Bridgewater. History writing and Wikipedia. *Computers and Composition*, 45:36–50, 2017.

[39] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.

[40] Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580, 1992.

[41] Halvard Buhaug. Dude, where's my conflict? lsg, relative strength, and the location of civil war. *Conflict Management and Peace Science*, 27(2):107–128, 2010.

[42] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but

we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *Proceedings of CHI*, 2008.

[43] Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. Robust systems for preposition error correction using wikipedia revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517, 2013.

[44] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[45] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[46] Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

[47] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2005.

[48] Ian J Cawood and David McKinnon-Bell. *The First World War*. Routledge, 2002.

[49] Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[50] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):32:1–32:25, November 2018.

[51] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.

[52] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234. ACM, 2018.

[53] Zoey Chen and Jonah Berger. When, Why, and How Controversy Causes Conversation. *Journal of Consumer Research*, 40(3):580–593, October 2013.

[54] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, 2018.

[55] Ashish Chopra, Morgan Mo, Samuel Dodson, Ivan Beschastnikh, Sidney S. Fels, and Dongwook Yoon. "@alex, this fixes #9": Analysis of referencing patterns in pull request discussions. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[56] Herbert H Clark and Susan E Brennan. Grounding in communication. 1991.

[57] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, 2019.

[58] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[59] Jonathan Cohen. Trusses: Cohesive subgraphs for social network analysis. *National Security Agency Technical Report*, 16:3–1, 2008.

[60] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.

[61] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 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, 2012.

[62] Paramita Das, Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Soumya Sarkar, and Animesh Mukherjee. Quality change: norm or exception? measurement, analysis and detection of quality change in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–36, 2022.

[63] Johannes Daxenberger and Iryna Gurevych. Automatically classifying edit categories in wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, 2013.

[64] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[67] Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, 2021.

[68] William N Dudley, Rita Wickham, and Nicholas Coombs. An introduction to survival statistics: Kaplan-meier analysis. *Journal of the advanced practitioner in oncology*, 7(1):91, 2016.

[69] Robert Englebretson. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, volume 164. John Benjamins Publishing, 2007.

[70] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[71] Song Feng. Dialdoc 2021 shared task: Goal-oriented document-grounded dialogue modeling. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 1–7, 2021.

[72] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, 2020.

[73] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 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, Avignon, France, April 2012. Association for Computational Linguistics.

[74] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[75] Katja Filippova. Multi-sentence compression: Finding shortest paths in word graphs. In *COLING*, 2010.

[76] Linda Flower and John R Hayes. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387, 1981.

[77] Dayne Freitag. Machine learning for information extraction in informal domains. *Machine learning*, 39(2-3):169–202, 2000.

[78] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, pages 55–95, 2013.

[79] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, jul 1996.

[80] Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, 2020.

[81] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, 2019.

[82] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tür. Clusterrank: a graph based method for meeting summarization. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[83] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[84] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[85] Robin Gieck, Hanna-Mari Kinnunen, Yuanyuan Li, Mohsen Moghaddam, Franziska Pradel, Peter A. Gloor, Maria Paasivaara, and Matthäus P. Zylka. Cultural differences in the understanding of history on Wikipedia. In Matthäus P. Zylka, Hauke Fuehres, Andrea Fronzetti Colladon, and Peter A. Gloor, editors, *Designing Networks for Innovation and Improvisation*, pages 3–12, Cham, 2016. Springer International Publishing.

[86] Dan Gillick and Benoit Favre. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics, 2009.

[87] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, 2019.

[88] Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE, 2018.

[89] Allen Louis Gorin, Irene Langkilde Geary, Diane Judith Litman, Marilyn Ann Walker, and Jeremy H Wright. Method and system for predicting problematic dialog situations in a task classification system, September 6 2005. US Patent 6,941,266.

[90] R Chulaka Gunasekara, David Nahamoo, Lazaros C Polymenakos, Jatin Ganhotra, and Kshitij P Fadnis. Quantized-dialog language model for goal-oriented conversational systems. *arXiv preprint arXiv:1812.10356*, 2018.

[91] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, 2022.

[92] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02 2022.

[93] Sonal Gupta and Christopher Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[94] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[95] Pentti Haddington. Stance taking in news interviews. *SKY Journal of Linguistics*, 17:101–142, 2004.

[96] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[97] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[98] Xiaochuang Han, Eunsol Choi, and Chenhao Tan. No permanent Friends or enemies: Tracking relationships between nations from news. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long and Short Papers), pages 1660–1676, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[99] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[100] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[101] Chikara Hashimoto and Manabu Sassano. Detecting Absurd Conversations from Intelligent Assistant Logs by Exploiting User Feedback Utterances. pages 147–156. ACM Press, 2018.

[102] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, 2017.

[103] Shiqing He, Allen Yilun Lin, Eytan Adar, and Brent J Hecht. The_tower_of_babel. jpg: Diversity of visual encyclopedic knowledge across wikipedia language editions. In *ICWSM*, pages 102–111, 2018.

[104] John Heritage and Steven Clayman. *Talk in action: Interactions, identities, and institutions*. John Wiley & Sons, 2011.

[105] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 541–550. Association for Computational Linguistics, 2011.

[106] Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1127–1136, 2011.

[107] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[108] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[109] Jill M. Hooley, Greg Siegle, and Staci A. Gruber. Affective and Neural Reactivity to Criticism in Individuals High and Low on Perceived Criticism. *PLOS ONE*, 7(9):e44412, September 2012.

[110] Meiqun Hu, Ee-Peng Lim, and Ramayya Krishnan. Predicting outcome for collaborative featured article nomination in Wikipedia. In *Proceedings of ICWSM*, volume 3, 2009.

[111] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information*

and knowledge management, CIKM '07, pages 243–252, New York, NY, USA, November 2007. Association for Computing Machinery.

[112] Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2160–2170, 2018.

[113] Yan Huang. *The Oxford handbook of pragmatics*. Oxford University Press, 2017.

[114] Christoph Hube. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 717–721, 2017.

[115] Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Osman Mutlu, Deniz Yuret, and Aline Villavicencio. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online, August 2021. Association for Computational Linguistics.

[116] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media*, 2014.

[117] Sohyeon Hwang and Aaron Shaw. Rules and rule-making in the five largest wikipedias. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 347–357, 2022.

[118] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24, 2018.

[119] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one*, 9(8):e104880, 2014.

[120] Kokil Jaidka, Andrea Ceolin, Iknoor Singh, Niyati Chhaya, and Lyle Ungar. Wikitalkedit: A dataset for modeling editors' behaviors on wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2191–2200, 2021.

[121] Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Macias-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting project: Resources and research. In *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*, 2004.

[122] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE, 2003.

[123] Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009*

Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, pages 1250–1259. Association for Computational Linguistics, 2009.

[124] Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342, 2016.

[125] Yohan Jo, Michael Yoder, Hyeju Jang, and Carolyn Rose. Modeling dialogue acts with content word filtering and speaker preferences. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2179–2189, 2017.

[126] Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. Language-agnostic topic classification for wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 594–601, 2021.

[127] Barbara Johnstone. Linking identity and dialect through stancetaking. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, pages 49–68, 2007.

[128] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[129] Gareth J. F. Jones, Ying Zhang, Eamonn Newman, Fabio Fantino, and Franca Debole. Multilingual search for cultural heritage archives via combining multiple translation resources. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 81–

88, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[130] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018.

[131] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 07 2018.

[132] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of NAACL*, pages 1647–1661, June 2018.

[133] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2:397, May 2012.

[134] Brian Keegan and Casey Fiesler. The evolution and consequences of peer producing wikipedia's rules. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 112–121, 2017.

[135] John Keegan. *The First World War*. Random House, 2014.

[136] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik

Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics.

[137] Joseph Kim and Julie A Shah. Improving team's consistency of understanding in meetings. *IEEE Transactions on Human-Machine Systems*, 46(5):625–637, 2016.

[138] Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics, 2010.

[139] Suin Kim, Sungjoon Park, Scott A Hale, Sooyoung Kim, Jeongmin Byun, and Alice H Oh. Understanding editing behaviors in multilingual wikipedia. *PloS one*, 11(5):e0155305, 2016.

[140] Aniket Kittur, Ed H Chi, and Bongwon Suh. What's in wikipedia? mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1509–1512, 2009.

[141] Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46, 2008.

[142] Aniket Kittur and Robert E Kraut. Beyond wikipedia: coordination and

conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224, 2010.

[143] Dan Klein and Christopher D Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478. Association for Computational Linguistics, 2004.

[144] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[145] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*, 2021.

[146] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[147] Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Soumen Chakrabarti, et al. Imojie: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886, 2020.

[148] Emiel Krahmer. Last words: What computational linguists can learn from psychologists (and vice versa). *Computational linguistics*, 36(2):285–294, 2010.

[149] Travis Kriplean, Ivan Beschastnikh, David W McDonald, and Scott A Golder. Community, consensus, coercion, control: cs* w or how policy

mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 167–176, 2007.

[150] Victor Kristof, Aswin Suresh, Matthias Grossglauser, and Patrick Thiran. War of words II: Enriched models of law-making processes. In *Proceedings of the Web Conference 2021*, WWW '21, page 2014–2024, New York, NY, USA, 2021. Association for Computing Machinery.

[151] Jakub Kubś. Historical narratives in different language versions of wikipedia. *Academic Journal of Modern Philology*, (12):83–94, 2021.

[152] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

[153] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[154] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[155] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge*

discovery in data mining - KDD '05, page 177, Chicago, Illinois, USA, 2005. ACM Press.

[156] Kwok Leung, Steven Su, and Michael W. Morris. When is criticism not constructive? the roles of fairness perceptions and dispositional attributions in employee acceptance of critical supervisory feedback. *Human Relations*, 54(9), September 2001.

[157] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308, 2014.

[158] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[159] Chen Li, Xian Qian, and Yang Liu. Using supervised bigram-based ilp for extractive summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1004–1013, 2013.

[160] Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy, July 2019. Association for Computational Linguistics.

[161] Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Par-

mar, and Simon Tong. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, 2019.

[162] Michael Lieberman and Jimmy Lin. You are where you edit: Locating Wikipedia contributors through edit histories. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2009.

[163] Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920, 2010.

[164] Hui Lin, Jeff Bilmes, and Shasha Xie. Graph-based submodular selection for extractive summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 381–386. IEEE, dec 2009.

[165] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[166] Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1957–1965. ACM, 2019.

[167] Feng Liu, Qirong Mao, Liangjun Wang, Nelson Ruwa, Jianping Gou, and

Yongzhao Zhan. An emotion-based responding model for natural language conversation. *World Wide Web*, June 2018.

[168] Tianyu Liu, Fuli Luo, Pengcheng Yang, Wei Wu, Baobao Chang, and Zhifang Sui. Towards comprehensive description generation from factual attribute-value tables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5985–5996, 2019.

[169] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[170] Charles Lovering and Ellie Pavlick. Unit testing for concepts in neural networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208, 2022.

[171] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, 2015.

[172] Ruiling Lu and Linda Bol. A Comparison of Anonymous Versus Identifiable E-Peer Review On College Student Writing Performance and the Extent of Critical Feedback. page 17.

[173] Khyati Mahajan and Samira Shaikh. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, 2021.

[174] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[175] Peter Makarov. Automated acquisition of patterns for coding political event data: Two case studies. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 103–112, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics.

[176] Senthil Mani, Neelamadhav Gantayat, Rahul Aralikatte, Monika Gupta, Sampath Dechu, Anush Sankaran, Shreya Khare, Barry Mitchell, Hemamalini Subramanian, and Hema Venkatarangan. Hi, how can i help you?: Automating enterprise it support help desks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[177] Adam Martin. The very last world war i veteran has died. 2012.

[178] Seth A. Marvel, Jon Kleinberg, Robert D. Kleinberg, and Steven H. Strogatz. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108(5):1771–1776, February 2011.

[179] J Nathan Matias and Merry Mou. Civilservant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[180] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 622–628, 2019.

[181] Arya D. McCarthy, James Scharf, and Giovanna Maria Dora Dore. A mixed-methods analysis of western and Hong Kong–based reporting on the 2019–2020 protests. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 178–188, Punta Cana, Dominican Republic (online), November 2021. Association for Computational Linguistics.

[182] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100, 2005.

[183] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[184] Tamir Mendel and Eran Toch. Susceptibility to social influence of privacy behaviors: Peer versus authoritative sources. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 581–593, 2017.

[185] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[186] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Effi-

cient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[187] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[188] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[189] Marc Miquel-Ribé and David Laniado. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 620–629, 2019.

[190] Volodymyr Miz, Joëlle Hanna, Nicolas Aspert, Benjamin Ricaud, and Pierre Vandergheynst. What is trending on wikipedia? capturing trends and language biases across wikipedia editions. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 794–801, New York, NY, USA, 2020. Association for Computing Machinery.

[191] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, 2011.

[192] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

[193] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 839–848, 2013.

[194] Jonathan T Morgan and Anna Filippova. 'welcome' changes? descriptive and injunctive norms in a wikipedia sub-community. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26, 2018.

[195] Jonathan T Morgan, Michael Gilbert, David W McDonald, and Mark Zachry. Editing beyond articles: diversity & dynamics of teamwork in open collaborations. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 550–563, 2014.

[196] Greg Myers. Stance-taking in public discussion in blogs. In *Self-Mediation*, pages 63–75. Routledge, 2013.

[197] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016.

[198] Courtney Napoles, Maria Nădejde, and Joel Tetreault. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566, 2019.

[199] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, 2018.

[200] Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*, 2018.

[201] Michael Noseworthy, Jackie Chi Kit Cheung, and Joelle Pineau. Predicting success in goal-driven human-human dialogues. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–262, 2017.

[202] Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[203] Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1104, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[204] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

[205] Tatsuro Oya and Giuseppe Carenini. Extractive summarization and di-

alogue act modeling on email threads: An integrated probabilistic approach. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 133–140, 2014.

[206] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[207] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: A study of power editors on wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, page 51–60, New York, NY, USA, 2009. Association for Computing Machinery.

[208] Ashwin Paranjape, Austin R. Benson, and Jure Leskovec. Motifs in Temporal Networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 601–610, New York, NY, USA, 2017. ACM.

[209] Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, 2021.

[210] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[211] Umashanthi Pavalanathan, Jim Fitzpatrick, Scott F Kiesling, and Jacob Eisenstein. A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of ACL*, pages 884–895, 2017.

[212] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. Mind your POV: Convergence of articles and editors towards Wikipedia's neutrality norm. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):137:1–137:23, November 2018.

[213] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. Mind your POV: Convergence of articles and editors towards Wikipedia's neutrality norm. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[214] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[215] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[216] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[217] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu,

Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

[218] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

[219] Edward L. Platt and Daniel M. Romero. Network Structure, Efficiency, and Performance in WikiProjects. In *Twelfth International AAAI Conference on Web and Social Media*, June 2018.

[220] Emily Porter, P. M. Krafft, and Brian Keegan. Visual narratives and collective memory across peer-produced accounts of contested sociopolitical events. *Trans. Soc. Comput.*, 3(1), feb 2020.

[221] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of GROUP*, 2007.

[222] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

[223] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 525–536, 2021.

[224] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.

[225] Kristopher W Ramsay. Settling it on the field: Battlefield events and war termination. *Journal of Conflict Resolution*, 52(6):850–879, 2008.

[226] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of ACL*, pages 1650–1659, August 2013.

[227] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 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, 2013.

[228] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference*, WWW '19, page 1567–1578, New York, NY, USA, 2019. Association for Computing Machinery.

[229] T. Antal P. L. Krapivsky S. Redner. Social Balance on Networks : The Dynamics of Friendship and Enmity, 2006.

[230] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020.

[231] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics, 2010.

[232] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1104–1112, New York, NY, USA, 2012. Association for Computing Machinery.

[233] Samuel Ritter, David G. T. Barrett, Adam Santoro, and Matt M. Botvinick. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. *arXiv:1706.08606 [cs, stat]*, June 2017. arXiv: 1706.08606.

[234] Pedro Rodriguez and Arthur Spirling. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *Journal of Politics*, 2021.

[235] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[236] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[237] Shinsaku Sakaue, Tsutomu Hirao, Masaaki Nishino, and Masaaki Nagata. Provable fast greedy compressive summarization with any monotone submodular function. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1737–1746, 2018.

[238] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[239] Anna Samoilenko, Florian Lemmerich, Maria Zens, Mohsen Jadidi, Mathieu Génois, and Markus Strohmaier. (don't) mention the war: A comparison of wikipedia and britannica articles on national histories. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 843–852, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[240] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[241] James Scharf, Arya D. McCarthy, and Giovanna Maria Dora Dore. Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 43–52, Online, August 2021. Association for Computational Linguistics.

[242] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[243] John R Searle and John Rogers Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

[244] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point:

Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

[245] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.

[246] Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263, 2020.

[247] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, 2018.

[248] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. The wisdom of polarized crowds. *Nature human behaviour*, 3(4):329–336, 2019.

[249] Xiaolin Shi, Jure Leskovec, and Daniel A. McFarland. Citing for high impact. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, page 49–58, New York, NY, USA, 2010. Association for Computing Machinery.

[250] Jonathan Shimshoni. Technology, military advantage, and world war i: A case for military entrepreneurship. *International Security*, 15(3):187–215, 1990.

[251] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 2004.

[252] Xin Shuai, Zhuoren Jiang, Xiaozhong Liu, and Johan Bollen. A comparative study of academic and wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, page 25–28, New York, NY, USA, 2013. Association for Computing Machinery.

[253] Valerie J. Shute. Focus on Formative Feedback. *Review of Educational Research*, 78(1):153–189, March 2008.

[254] Alastair Smith. Fighting battles, winning wars. *Journal of Conflict Resolution*, 42(3):301–320, 1998.

[255] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[256] Jacob Solomon and Rick Wash. Critical mass of what? exploring community growth in wikiprojects. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[257] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[258] Niklas Stoehr, Lucas Torroba Hennigen, Samin Ahbab, Robert West, and Ryan Cotterell. Classifying dyads for militarized conflict analysis. In *Pro-*

ceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7775–7784, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[259] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[260] Besiki Stvilia, Les Gasser, Michael B Twidale, and Linda C Smith. A framework for information quality assessment. *Journal of the American society for information science and technology*, 58(12):1720–1733, 2007.

[261] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: Slowing growth of wikipedia. In *Proceedings of WikiSym*, 2009.

[262] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics.

[263] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. Dream: A challenge dataset and models for dialogue-based reading comprehension. *arXiv preprint arXiv:1902.00164*, 2019.

[264] Ashish Sureka and Atul Goyal. Insights on transferability of dialog-act cue-phrases across communication domains, modality and semantically

similar dialog-acts. In *8th International Conference on Natural Language Processing (ICON)*, pages 28–37, 2010.

[265] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

[266] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

[267] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

[268] Yufei Tian, Tuhin Chakrabarty, Fred Morstatter, and Nanyun Peng. Identifying distributional perspectives from colingual groups. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 178–190, 2021.

[269] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[270] Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. A graph

degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1860–1870, 2016.

[271] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1499–1509, 2015.

[272] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

[273] Khoi-Nguyen Tran and Peter Christen. Identifying multilingual wikipedia articles based on cross language similarity and activity. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1485–1488, New York, NY, USA, 2013. Association for Computing Machinery.

[274] Francesca Tripodi. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, page 14614448211023772, 2021.

[275] Flavian Vasile, Elena Smirnova, and Alexis Conneau. Meta-prod2vec: Product embeddings using side-information for recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pages 225–232, 2016.

[276] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion

Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[277] Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, 2016.

[278] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[279] Luis von Ahn and Laura Dabbish. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[280] Ellen M Voorhees and Hoa Trang Dang. Overview of the trec 2002 question answering track. In *Trec*, volume 2003, pages 54–68. Citeseer, 2003.

[281] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5788–5793, 2019.

[282] Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *28th Annual Meeting of the*

Association for Computational Linguistics, pages 70–78, Pittsburgh, Pennsylvania, USA, June 1990. Association for Computational Linguistics.

[283] Borui Wang, Chengcheng Feng, Arjun Nair, Madelyn Mao, Jai Desai, Asli Celikyilmaz, Haoran Li, Yashar Mehdad, and Dragomir Radev. Strudel: Structured dialogue summarization for dialogue comprehension. *arXiv preprint arXiv:2212.12652*, 2022.

[284] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, 2016.

[285] Lu Wang and Claire Cardie. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1395–1405, 2013.

[286] Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, 2020.

[287] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 743–756, 2015.

[288] Galen Weld, Amy X Zhang, and Tim Althoff. Making online communi-

ties' better': A taxonomy of community values on reddit. *arXiv preprint arXiv:2109.05152*, 2021.

[289] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[290] Steve Whittaker and Phil Stenton. Cues and control in expert-client dialogues. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 123–130. Association for Computational Linguistics, 1988.

[291] Wikipedia. Wikipedia:Text of the Creative Commons Attribution-ShareAlike 3.0 Unported License — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Wikipedia%3AText%20of%20the%20Creative%20Commons%20Attribution-ShareAlike%203.0%20Unported%20License&oldid=1130883578`, 2023. [Online; accessed 16-January-2023].

[292] Edward L. Wilson, Ming-Wu Yuan, and John M. Dickens. Dynamic analysis by direct superposition of Ritz vectors. *Earthquake Engineering & Structural Dynamics*, 10(6):813–821, November 1982.

[293] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Hugging-

face's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[294] Kam-Fai Wong, Mingli Wu, and Wenjie Li. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 985–992. Association for Computational Linguistics, 2008.

[295] Fei Wu and Daniel S Weld. Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, pages 635–644. ACM, 2008.

[296] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? pages 203–212. IEEE, June 2016.

[297] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 975–985, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[298] Han Xiao, Bruno Ordozgoiti, and Aristides Gionis. Searching for polarization in signed graphs: A local spectral approach. In *Proceedings of The Web Conference 2020*, WWW '20, page 362–372, New York, NY, USA, 2020. Association for Computing Machinery.

[299] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Who did what: Editor role identification in wikipedia. In *Proceedings of ICWSM*, volume 10, 2016.

[300] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Who Did What: Editor Role Identification in Wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):446–455, 2016. Number: 1.

[301] Jin-Ge Yao, · Xiaojun Wan, and Jianguo Xiao. Recent advances in document summarization. *Knowl Inf Syst*, 53:297–336, 2017.

[302] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of Conflicts in Wikipedia. *PLoS ONE*, 7(6):e38869, June 2012.

[303] Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*, pages 365–368, 2010.

[304] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W Bruce Croft, and Mark Sanderson. Document summarization for answering non-factoid queries. *IEEE transactions on knowledge and data engineering*, 30(1):15–28, 2017.

[305] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online communities using coarse discourse structures. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[306] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. *PolicyKit: Building Governance in Online Communities*, page 365–378. Association for Computing Machinery, New York, NY, USA, 2020.

[307] Amy X Zhang, Grant Hugh, and Michael S Bernstein. Policykit: building

governance in online communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 365–378, 2020.

[308] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, 2018.

[309] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, 2018.

[310] Yiwei Zhou, Alexandra Cristea, and Zachary Roberts. Is Wikipedia really neutral? A sentiment perspective study of war-related Wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 160–168, Shanghai, China, October 2015.

[311] Haiyi Zhu, Robert Kraut, and Aniket Kittur. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 935–944, 2012.

[312] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–23, 2018.