Investigating Natural Language Interaction within Communities

Kelvin Luu
General Examination Document

Committee: Noah A. Smith, Hannaneh Hajishirzi, Sheng Wang , Jevin West

December 7, 2021

Abstract

Members of a community, online or offline, often share similar interests or goals. For example, members of a debate community may be interested in winning a debate and those in the research community may situate their work in terms of other existing literature to make it more understandable. This quality, along with an abundance of data, allow researchers to study interesting linguistic phenomena in how members communicate with each other premised on those shared goals.

In this work, we investigate language use in two communities. First, we study how authors of scientific texts *explain* how their paper relates to another. We propose a new task, relationship explanation generation for scientific texts, by using in-line citation text as a proxy. We introduce a new model, SciGen, empowered by pretrained language models. We perform extensive human evaluation and discuss potential shortcomings of our generations.

Second, we explore the problem of quantifying persuasive skill over time. Using data from an online debate forum, we construct a model based on on the Elo ranking model and incorporate historical lingusitic data. In order to estimate skill, we frame our prediction task to *forecast* the outcome of a debate.

Based on the above work, we describe ongoing and proposed work on investigating how NLP models degrade over time. We described our work that studied phenomena specific to particular communities. We extend these works by investigating how model performance deterioration over time differs across several tasks and domains.

1 Introduction

Interactions between members of a community provide a rich source of data for studying various language use patterns. In today's world, many of these interactions are available

as natural language text. For example, there are 2.1 million daily users generating over 500 million tweets per day on Twitter; Wikipedia has discussions over prior revisions spanning over a decade; and even peer-reviews on scientific works in submission are publicly accessible on OpenReview. The abundant data, often signifying meaningful interactions, allow for research into various linguistic phenomena within a community.

In this work, we are interested in how to represent and model particular community behaviors and phenonomena as inputs to our models. NLP models have historically benefited from input design ranging from feature selection to prompt engineering (Yang and Pedersen, 1997; Jiang et al.; Schick and Schütze, 2021). Specifically, we use the choice of input representation, such as the historical context of a text. to reveal properties from text, such as user expertise.

First, we study how computer science researchers explain the relationship of one of their papers to another. We introduce a new task, relationship explanation generation for scientific texts, and operationalize it using in-line citation text as a proxy. We build on a pretrained language model, GPT2, to construct SciGen(Radford et al., 2019a).

Due to limitations in input context windows of many pretrained language models, we cannot use entire scientific texts as input. Instead, we seek to find a dense representation of a scientific document that provides enough information to for SciGen to generate meaningful explanations. We experiment with various methods of selecting full sentences to use as input for scientific documents. Since explanations of scientific texts are highly technical, we perform a highly expert human evaluation and discuss potential future directions.

Second, we explore the problem of quantifying persuasive skill over time. We operationalize this problem by using data from an online debate forum. In this project, we build a *linguistic profile* of a debater, or a summmary of their language use. We use the linguistic profiles to build on the Elo ranking model, a system for ranking participants in two-player games. Through our analysis, we are able to draw conclusions such as which types of features are correlated with skill and how more skilled users become experts over time.

We also discuss ongoing and proposed work on investigating how NLP models degrade over time. Prior work has broadly established that language use can change over time for various reasons(Labov, 1994; Altmann et al., 2009; Eisenstein et al., 2014). However, prior research that studied language shift over time has mostly focused on a narrow set of domains or tasks (Röttger and Pierrehumbert, 2021; Cao et al., 2021; Zhang and Choi, 2021; Rijhwani and Preoţiuc-Pietro, 2020). We propose work that extends our two completed works, which invested phenomena that are characteristic of particular domains, by investigating performance degradation of NLP models for a variety of downstream tasks and text domains.

2 Explaining Scientific Documents in Relation

We now explore how members of the scientific community communicate with each other. In Luu et al. (2021), we investigate how researchers in the computer science community explain

how their work is related to another. We introduce a new task, generating natural language explanations of the relationships between two scientific papers, and operationalize it with citation sentences. As we are interested in how two papers broadly relate to each other, we generate such sentences from general representations of document content rather than the specific in-text locations where these sentences occur.

In this section, we describe the contributions of Luu et al. (2021): we establish a new task of generating relationship explanations; introduce a novel dataset for the task; release our SciGen model for describing document relationships; and provide an extensive evaluation and analysis of machine generated technical text.

2.1 Problem Definition

We aim to generate an explanation: a natural language sentence which expresses how one document relates to another. Explicit examples of such sentences are nontrivial to find in corpora, especially when annotation for a highly technical task is expensive. To this end, we use in-text citations in a scientific document to prior work as proxies for relationship explanations. We use these citing sentences as partial supervision for our task, and refer to them as "explanations."[1]

We distinguish one document as the *principal* document, from which we will draw explanations that reference the *cited* document. Let t denote an explanation drawn from principal document S, and S' denote S without t. Then let

$$P(t \mid S', C) \tag{1}$$

be the probability of t given S' and the cited document C. A good generation technique should maximize this probability across a large number of $\langle t, S, C \rangle$ triples, so that at inference time the model is able to generate a sentence t^* which accurately describes the relationship between new documents $\hat{S}$ and $\hat{C}$.

Optimizing Equation 1 is made easier by modern representation learning. Pretrained neural language models like GPT2 have shown strong performance when generating sentences conditioned on a context. However, existing implementations of GPT2 limit the context window to 512 or 1024 tokens, far smaller than scientific documents. In this work, we explore ways to represent the documents' content for use with language models.

Data We use English-language computer science articles and annotation from the S2ORC dataset (Lo et al., 2020). S2ORC is a large citation graph which includes full texts of 8.1 million scientific documents. We use 154K connected computer science articles, from which we extract 622K explanations with a single reference that link back to other documents in our corpus. We omit any sentences that cite more than one reference. We hold 5000 sentences for each of the validation and test sets. Detailed statistics can be found in Table 1.

[1] Future work might seek to filter or systematically alter in-text citations to be more explanation-like, without otherwise changing our approach.

	total	average/doc.
documents	154K	–
tokens	813M	5.3K
unique tokens	7.1M	1.3K
explanations	622K	4.0

Table 1: Dataset statistics, total and per document.

Evaluation The most appropriate evaluation metric for this and many text generation tasks is human judgment by potential users of the system. Evaluating explanations of the relationships between scientific documents requires human judges with scientific expertise whose time and effort can be costly. While collecting human judgments in technical domains is relatively rare, we believe it to be an important step in evaluating our systems for this task. Thus, we conduct thorough human evaluations and analyses with expert judges. We make use of both larger scale expert evaluations yielding hundreds of judgements as well as smaller scale, deeper evaluations where we can effect a higher degree of quality control over fewer datapoints. Further, we make use of intermediate human evaluations in the development of our models, and supplement these evaluations with automatic metrics — BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) that are established in other generation tasks.

2.2 Models

We develop several models for explaining document relationships. Following current work in neural text generation, we finetune the predictions of a large pretrained language model to our task (§ 2.3). In order to bring the language model into the scientific text domain, we do additional language model pretraining over full scientific texts. We also investigate approximate nearest neighbor methods to retrieve plausible human-authored explanations from the training data as a baseline (§ 2.4).

2.3 Neural Text Generation

Recent work has shown that finetuning large pretrained language models to text generation tasks yields strong results (Zellers et al., 2019). To this end, we construct SciGen, a model based on GPT2 (Radford et al., 2019b), a transformer model trained on 40GB of internet text with a left-to-right language modeling objective (Vaswani et al., 2017). We do so by finetuning the predictions of the language model to generate explanations using different expressions of the principal and cited document as context.

To finetune GPT2 architectures for text generation, it is typical to concatenate the conditioning context $X = x_1 \dots x_n$ and target sentence $Y = y_1 \dots y_m$ with a special separator token ξ^y. To adapt this technique to our task, we construct the conditioning context X from the principal and cited documents and use the explanation as Y. We take j tokens from principal

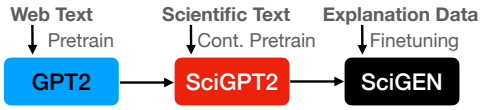

Figure 1: Overview of the construction of SciGen. We take the pretrained GPT2 and continue pretraining on scientific texts. We then finetune using data in Table 1.

document $s_1, \ldots, s_j$ along with k tokens from the cited document $c_1, \ldots, c_k$ (which tokens to draw from the two documents is an independent variable that we explore experimentally). We then condition the generation of explanation Y on $X = s_1, \ldots, s_j, \xi^x, c_1, \ldots, c_k$, where ξ^x is a token used to indicate the end of the principal document. SciGen is trained to predict the explanation one token at a time as described above.

At inference time, the model is provided with an unseen principal/cited document pair. An explanation of their relationship is generated one token at a time using nucleus sampling (Holtzman et al., 2020). At timestep t, output token $\hat{y}_t$ is sampled from the top 90% of the distribution $P(\hat{y}_t \mid X, \xi^y, \hat{y}_1, \ldots, \hat{y}_{t-1})$ (renormalized). The selected $\hat{y}_t$ is used to condition the prediction of subsequent tokens.

Context The primary question we investigate with the SciGen model is what kind of input is best for describing the relationship between the principal and cited documents accurately and informatively. Since models based on GPT2 have a small context window relative to the length of scientific documents, we investigate the use of abstracts, introductions, or non-citing sentences sampled from throughout the document as conditioning context. The effectiveness and description of these approaches is described in § 2.5.

Language Model Pretraining Prior work has shown that pretraining on in-domain data improves the performance of large language models on domain-specific tasks (Beltagy et al., 2019; Gururangan et al., 2020). Inspired by this, we continue pretraining the GPT2 model in the science domain to produce SciGPT2, which we use as the underlying language model for SciGen described above. SciGPT2 starts from the standard pretrained GPT2-base model and is trained for an additional 75k gradient updates at a batch size of 64 (effectively a single epoch over 4.8 million abstracts and body paragraphs) with a language modeling objective. Figure 1 illustrates the process.

We observed significant improvements in the quality of SciGen outputs after replacing the underlying GPT2 language model with the domain-specific SciGPT2 model. We saw a perplexity improvement in a held-out set and, in informal inspections, qualitative improvements as well.

When using pretrained language models, text from task-specific test data cannot be guaranteed to be absent from the large task-independent corpora upon which these models are trained, which may improve model performance compared to models without this exposure. For the experiments described in this work, we train a version of SciGPT2 only

on documents appearing in the training data, so that the principal documents and target sentences in the test data are guaranteed to be unseen by the language model. We provide this and a full-corpus version of SciGPT2 as resources for future research.[2]

2.4 Retrieval with Approximate Nearest Neighbors

While neural text generation techniques have advanced significantly in recent years, their outputs are still inferior to human authored texts. For some tasks, it is better to retrieve a relevant human-authored text than to generate novel text automatically (Fan et al., 2018). Is this also the case when generating explanations?

To answer this question, we use an information retrieval (IR) baseline. We adapt an approximate nearest neighbor search algorithm to find similar pairs of documents. The basic search procedure is as follows: Given a test instance input (S, C) for principal S and cited document C, we find the set $\mathbf{N}_C$, the nearest neighbors to C in the training data. For each document N_C from $\mathbf{N}_C$, let $\mathbf{N}_S$ be the set of documents that cite N_C. This means that each $N_S \in \mathbf{N}_S$ contains at least one citing sentence t' which cites N_C. We use the t' associated with the (N_S, N_C) pair from the training set which is closest to (S, C) as the explanation of their relationships, which we describe in more detail below.

We measure the closeness of two pairs of documents using the cosine distances between vector representations of their abstracts. The abstract of each document is encoded as a single dense vector by averaging the contextualized embeddings provided by the SciBERT model of Beltagy et al. (2019) and normalizing. The distance between (S, C) and neighbors (N_S, N_C) is computed as

$$\alpha \cos(S, N_S) + \beta \cos(C, N_C), \tag{2}$$

where α and β control the relative contribution of the two document similarities. We explore setting both α and β to 1, or tuning them to optimize BLEU on the validation data using MERT (Och, 2003).

2.5 Representing Documents with Sentence Selection

Methods for the related task of citation recommendation have made use of abstracts, which perhaps act as sufficient summaries of document content. Building on this, we represent the principal and cited documents with the first 450 tokens of either their abstracts, introductions, or sentences randomly sampled from throughout the full document.[3] In this section, we answer two questions: 1) do neural generation models with sentence-based context outperform the IR baseline and 2) does the type of sentence-based context (abstract, introduction, sampled) matter? We answer these questions by performing both automatic and human evaluations.

[2] https://github.com/Kel-Lu/SciGen

[3] We exclude any sentence with a citation from being sampled in all conditions. This context type is also only used for the cited document and not the principal document.

	Method	Context	BLEU	Rouge-1	Rouge-2	Rouge-L
Sentence-Based	SciGen	principal abs × cited abs	9.82	10.7	0.6	8.4
		principal abs × cited intro	9.39	10.7	0.6	8.4
		principal abs × cited sample	9.60	10.7	0.7	8.5
		principal intro × cited abs	9.92	11.1	1.0	8.7
		principal intro × cited intro	9.80	11.1	1.1	8.8
		principal intro × cited sampled	9.81	10.9	0.9	8.7
	retrieval	principal abs × cited abs	9.93	14.2	0.7	9.7
		+ MERT (BLEU)	10.23	14.3	0.7	9.8
		no principal × cited abs	9.79	14.1	0.6	9.6
IE-based	SciGen	principal intro × cited tfidf	13.17	15.0	1.3	12.0
		principal abs × cited entities	13.10	14.3	0.8	11.4
		principal intro × cited entities	13.41	14.7	1.4	11.8
	+Ranking	principal intro × cited tfidf	13.50	15.5	1.6	12.3
		principal abs × cited entities	13.28	14.7	1.0	11.6
		principal intro × cited entities	13.16	15.0	1.3	11.8

Table 2: Automatic evaluation of generated texts for all of our systems. Our best models, the IE-based ones, are omitted for space reasons. Please see our work (Luu et al., 2021) for details.

2.6 Representing Documents with Information Extracted Contexts

We found in our work that generations using selected sentences as context can miss important phrases such as unique model or dataset names and other lower-frequency terms. We investigated using IE-based contexts, such as lists of salient words and phrases, as dense representations for our task. We report results and examples in Table 2 but omit details for space reasons. Please see our work, Luu et al. (2021) for more details.

2.7 Automatic Evaluation

We compare the SciGen and IR systems using BLEU (Papineni et al., 2002) and ROUGE (specifically L; Lin, 2004). The "Sentence-based" rows of Table 2 show the test set performance of the IR system and the best SciGen models when provided with the different sentence-based input context combinations.

We assess statistical significance as well by bootstrapping with 1000 samples in each of 100 iterations. We find that context *does* make a difference for SciGen, and that a slight but statistically significant performance improvement comes from using the introduction of the principal document rather than the abstract.[4] We do not, however, find enough evidence to reject the null hypothesis that any particular representation of the cited document's content (abstract, intro, or random sample) yields any significant difference in performance.

We find that using the introduction of the principal document paired with the abstract of the cited document performs best, and so we select these for human evaluation. The

[4] $p < 0.01$ after Bonferroni correction.

	Specific	**Correct**	**S&C**	*agr*
SciGen	72.3	64.0	55.0	*70.5*
IR	74.8	46.3	40.0	*77.5*
Gold	81.4	72.1	68.0	*83.8*
agreement	*69.8*	*71.4*	*63.1*	

Table 3: Human evaluation of SciGen (intro $\times$ abs) and IR (abs $\times$ abs) systems compared with gold explanations in percent. S&C represents those that were both specific and correct. All differences significant at $p < 0.01$ except SciGen vs. IR specific.

IR systems perform well, obtaining slightly better scores in some settings. We choose the MERT-optimized version for human evaluation.

2.8 Human Evaluation

We conduct a human evaluation to determine, given a particular pair of principal and cited abstracts, how *correct* and *specific* the generated explanation of their relationship is. By "correct" we mean: does the explanation correctly express the factual relationship between the principal and cited documents? Because generic explanations such as "This work extends the ideas of Chomsky and Halle (1968)", while possibly factual, do not express a detailed understanding of the documents' relationship, we ask judges whether the explanation describes a specific relationship between the two works. An explanation can be specific even it is incorrect.

We compare the *principal intro* $\times$ *cited abs* SciGen setting against the tuned IR system. For calibration, we also elicit judgments for the gold explanations extracted from principal documents along with the correct principal and cited abstracts. In all three cases, we ensure that the principal document appeared in the ACL anthology to ensure annotator expertise.

To ensure no annotator sees the output of more than one system on each datapoint, we randomly select 50 datapoints for each system (*principal intro* $\times$ *cited abs*, IR, and Gold explanations) from the subset of our test data whose principal documents appear in the ACL anthology. Each judge is given 15 datapoints for each of the specificity and correctness qualities. Judges are shown a table of datapoints and asked to mark whether each meets ("Yes") or fails to meet ("No") the condition. Judges are permitted to label "?" or skip examples they feel uncertain about or unqualified to judge, which we ignore. In total we solicit 37 NLP researchers and collect over 800 judgments, with over 100 for each system/quality dimension combination.

Table 3 shows the percentage of "yes" judgments versus the total of "yes" and "no" judgements for each system/quality combination, along with pairwise agreement rates. Gold texts received the highest scores for all dimensions of text quality from the evaluators as well

as the highest agreement rate. We can also see that IR systems tend to produce incorrect explanations more often than not.

The SCIGEN system performs quite well in this analysis, with a majority of outputs deemed correct. We observe a larger difference in specificity between SCIGEN and gold texts, indicating that SCIGEN, like many neural text generation systems, often generates vague and generic sentences. These generations tended to be vacuous such as "(CITED) This work is an extension of the paper." Specificity is key for future downstream applications such as automated literature review and will need to be improved for those tasks.

2.9 Discussion

1	Principal:	*A Syllable-based Name Transliteration System*
	Cited:	*A Joint Source-Channel Model for Machine Transliteration*
	SCIGEN:	Following *Cited* , Chinese characters are considered as Pinyin sequence.
2	Principal:	*Recovering discourse relations: Varying influence of discourse adverbials*
	Cited:	*The Benefits of a Model of Annotation*
	SCIGEN:	The two text collections provided by *Cited* were used for training, and the other two text collections were used for evaluation.
3	Principal:	*Coreference Resolution for Swedish and German using Distant Supervision*
	Cited:	*Collective Cross-Document Relation Extraction Without Labelled Data*
	SCIGEN:	It is one of the most widely used distant supervision techniques and is inspired by techniques proposed by *Cited*.
4	Principal:	*Neural Text Generation in Stories Using Entity Representations as Context*
	Cited:	*Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer*
	SCIGEN:	The authors of *Cited* proposed a model that combines neural generation with user interaction to create an object-centric reading experience.

Table 4: Example explanations. The given texts are the document titles and the SCIGEN outputs. In the last example, the two documents **do not** cite each other.

Example system outputs for selected test datapoints are shown in Table 4. The first example illustrates a case where the model identifies a correct relationship between the two documents. In this instance, they both use the pinyin representation for Chinese characters in their transliteration models.

Output 2 demonstrates a failure of the explanation generation system. The principal document deals with the topic of discourse relations, the automatic identification of which is a long-standing machine learning task. However, this particular document is an analysis paper, and does not involve any training.

Output 3 is an example of a "Not specific (but not incorrect)" case. Here again the explanation generated by SCIGEN is topical, dealing with the concept of "distant supervision"

that is key to both input documents. However, this sentence fails to capture the specific use that the principal makes of the research described in cited document.

The final example, output 4, showcases potential for our system to explain concurrent work. The generated text summarizes the *cited* and implies that *principal* will build on that work. However, selected papers are both concurrent generation papers published in the same venue and do not cite each other. This appears to be a weakness in using citation sentences as proxies for relationship explanations. Citations of contemporaneous work occur less frequently, so these types of sentences appear less often in training. Similarly, relationship explanations between papers with more distant connections (e.g., "multi-hop" in the citation graph) are missing in our training data.

In addition to missing some relationships, not all citation sentences are useful as explanations. As pointed out by other work, citation sentences can often be simple summaries of the cited work (Qazvinian and Radev, 2008; Cohan and Goharian, 2017). Alternatively, they can be too specific to be useful, as seen in Output 1, where a higher-level summary might be more useful. Future work could focus on curating better training sets for our task.

It is notable that the SciGen model usually outputs syntactically correct and topical explanations, even given the difficulty of the vocabulary in this domain. This is consistent with many recent findings using domain-specific language models.

3 Measuring Persuasive Skill Over time

In this section, we explore skill estimation of members of a community evolve over time. In (Luu et al., 2019), we found that online debate communities offer an opportunity to investigate *persuasive skill*. These communities feature users who participate in multiple debates over their period of engagement. Such debates involve two parties who willingly and formally present divergent opinions before an audience. Unlike other media of persuasion, such as letters to politicians, there is a clear signal, a win or loss, indicating whether or not a debater was successful against the adversary. This work aims to quantify the skill level of each debater in an online community and also investigates which factors contribute to expertise.

While persuasion has generated interest in the natural language processing community, most researchers have not tried to quantify the persuasiveness of a particular speaker. Instead, they estimate how persuasive a *text* is, using linguistic features such as the author's choice of wording or how they interact with the audience (Tan et al., 2014, 2016; Althoff et al., 2014; Danescu-Niculescu-Mizil et al., 2012). Previous research has also established that debaters' content and interactions both contribute to the success of the persuader (Tan et al., 2016; Zhang et al., 2016; Wang et al., 2017). These works have found textual factors that contribute to a debater's success, but they do not emphasize the role of the individual debater.

There has been some recent work on studying individual debaters. Durmus and Cardie (2019) analyze users and find that a user's success and improvement depend on their social

network features. We take another approach by estimating each user's skill level in each debate they participate in by considering their debate history. These estimates reveal features correlated with skill and the importance of particular debates over time.

We extend the Elo (1978) rating system, a model designed for rating players in two-player games, for the debate setting using linguistic features. We construct a family of models based on Elo and decompose them into two design choices that align with two questions we hope to answer: 1) the features we use and 2) how we might choose to aggregate those features from past debates.

To validate our skill estimates, we introduce a **forecasting** task (§3.1). Previous work predicted the winner of a debate using the text of the current debate. In contrast, we aim to predict winners using our skill estimates *before* the debate (ignoring the current debate's content). This design ensures that we are modeling skill of the debater, as inferred from past performance, not the idiosyncrasies of a particular debate.

We also investigate the predictive power of our estimates through an analysis of the results (§3.5). We show that our full model outperforms the baseline Elo model, approaching the accuracy of an oracle that *does* use the text of the current debate. Moreover, we find that not all past debates are equally useful for prediction: more recent debates are more indicative of the user's current level of expertise. This adds support to our conjecture that individual debaters tend to improve through the course of their time debating. Finally, we track the linguistic tendencies of each debater over the course of their debating history. We show that several features correlated with high skill, such as length of their turns, increase over time for the best debaters, but stay static for those with less skill. These findings give further evidence that debaters improve over time.

3.1 Problem Formulation

Our aim is to estimate a debater A's persuasive ability after observing a series of debates they participated in (denoted $d_1^A, \ldots, d_{t-1}^A$ if we are estimating ability just before the tth debate). We wish to take into account the content of those debates, so as to understand what factor reveals a debater's skill levels. We formulate a prediction task: estimate p_A for A's tth debate, given $d_1^A, \ldots, d_{t-1}^A$ and also the opponent's debate history (which might be of a different length). Unlike previous work (Zhang et al., 2016; Potash and Rumshisky, 2017; Tan et al., 2018), we do not use the content of the *current* debate (d_t^A) to predict its outcome; rather, we forecast the outcome of the debate as if the debate has not yet occurred.

By observing debate outcomes alongside the two participants' histories, we can estimate the parameters of such a probability model. We seek to *explain* the probability of winning through linguistic features of past debate content.

3.2 Data

We use both debate and user skill data from the Debate.org dataset introduced by Durmus and Cardie (2018). Any registered user on the website can initiate debates with others or vote on debates conducted by others.

Registered users can create a debate under a topic of their choosing. The person initiating the debate, called the **instigator**, fixes the debate's number of rounds (2–10) and chooses the category (e.g., politics, economics, or music) at the start of the debate. The instigator then presents an opening statement in the first round and waits for another user, the **contender**, to accept the debate and write another opening statement to complete the first round.[5] We define a debater's **role** as being either the instigator or contender.

To determine the debate winner, other Debate.org users vote after the debate ends. In this phase, voters mark who they thought performed better in each of seven categories (see Figure 2).[6] This phase can last between 3 days and 6 months, depending on the instigator's choice at the debate's creation. While Debate.org defines a win by the highest number of points, we would like to model how convincing each debater is. Therefore, we count the number of times each debater was rated as more convincing to a voter *despite* presenting a viewpoint that the voter disagreed with before the debate. The debater with the higher count of such votes is considered the "winner" in the remainder of this paper.

From an unfiltered set of 77,595 debates, we remove all debates where a user forfeits, that lack a winner, or that do not have a typical voting style with seven categories. Of the remaining debates, we focus on debates where the participants have completed at least five debates. This leaves us 4,486 debates and 1,284 users.

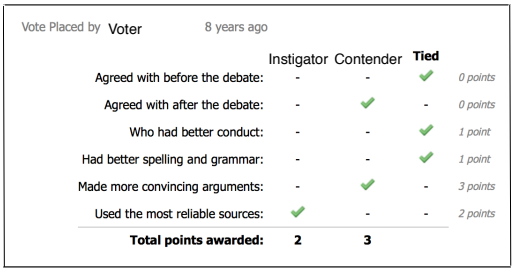

Figure 2: An example of the Debate.org voting system.

3.3 Expertise Estimation

In order to explore debater expertise and discover what contributes to a user's expertise over their time on Debate.org, we begin with a conventional approach to skill estimation, the Elo rating system, which serves as both a baseline and the basis for our final model.

[5] While many debaters use their first round to make an opening statement, some use it only to propose and accept debates. If the first turn in an n-round debate is under 250 words, we merge each debater's first two turns and treat the debate as an $(n-1)$-round debate.

[6] Another system of voting lets voters choose who they thought performed better over the entire debate. While these appear in the dataset, we do not use them in this paper.

3.3.1 Elo Model

Elo originated as a ranking system for chess players; it has been adapted to other domains, such as video games(Herbrich et al., 2007). It is one of the standard methods to rate players of a two-player, zero-sum game (Elo, 1978).[7] Elo assigns positive integer-valued scores, typically below 3,000, with higher values interpreted as "more skill." The difference in the scores between two debaters under a logistic model is used as an estimate of the probability each debater will win. For example, consider a debate between A and B. A has an Elo rating of $R_A = 1900$, and B has an Elo rating of $R_B = 2000$. Using the Elo rating system, p_A, the probability that A wins is[8]

$$p_A = \frac{1}{1+10^{0.0025(R_B-R_A)}} = \frac{1}{1+10^{0.25}} \approx 0.36 \tag{3}$$

Ratings are updated after every debate, with the winner (equal to A or B) gaining (and the loser losing) $\Delta = 32(1-p_W)$ points. (32 is an arbitrary scalar; we follow non-master chess in selecting this value.)

Note that the magnitude of the change corresponds to how unlikely the outcome was. While the Elo ratings traditionally take only a win or loss as input, there have been adjustments to account for the magnitude of victory. One such method would be to use the score difference between the two players to adjust the Elo gain (Silver, 2015). If we let S_A and S_B be scores for A and B respectively the modified gain Δ' is

$$\Delta' = \log(|S_A - S_B| + 1) \times \Delta \tag{4}$$

Under this model, we represent a user's history and skill level as a single scalar, i.e., their Elo rating. The Elo system ignores all other features, which include the style a debater uses in the debates and the content of their argument. We therefore view this model as a baseline and extend it.

3.3.2 Incorporating a Linguistic Profile into Elo

Elo scores are based entirely on wins and losses; they ignore debate content. We seek to incorporate content into expertise estimation by using linguistic features. If we modify the exponential base in Equation 3 from 10 to e, we can view Elo probabilities (e.g., p_A) as the output of a logistic regression model with one feature (the score difference, $R_A - R_B$) whose weight is 0.0025; that is, $p_A = \sigma(0.0025 \cdot (R_A - R_B))$. It is straightforward to incorporate more features, letting

$$p_A = \sigma\left(\mathbf{w} \cdot (R_A - R_B)\right) \tag{5}$$

[7] The Elo model is a special case of the Bradley-Terry model (Bradley and Terry, 1952).

[8] The base of the exponent, 10, and the multiplicative factor on the difference $R_A - R_B$, 0.0025, are typically used in chess.

where $\mathbf{w}$ is a vector of weights and R_U is user U's "profile," a vector of features derived from past debates. In this work, the linguistic profiles are designed based on extant theory about the linguistic markers of persuasion, and the vectors are *weighted averages* of features derived from earlier debates.

3.3.3 Features

We select features discussed in prior work as the basis for our linguistic profiles (Tan et al., 2016, 2018; Zhang et al., 2016). We extract these measurements from each of the user's debates. For a given debate and user, we calculate these values over the rounds written by the user. For example, if we were interested in a debate by a user as the instigator, we would only calculate features from the instigator rounds of that debate (since the contender rounds of that debate were written by their opponent). Table 5 shows the full list of features.

Hedging with fightin' words. We introduce one novel feature for our work: hedging with fightin' words. "Fightin' words" refer to words found using a method, introduced by Monroe et al. (2008), which seeks to identify words (or phrases) most strongly associated with one side or another in a debate or other partisan discourse.[9] We are interested in situations where debaters evoke fightin' words (their own, or their opponents') with an element of uncertainty or doubt. We use each debater's top 20 fightin' words (unigrams or bigrams) as features, following Zhang et al., 2016, who found this feature useful in predicting winners of Oxford-style televised debates. We also count cooccurences of fightin' words with hedge phrases like "it could be possible that" or "it seems that." An example of this conjoined feature is found in the utterance "Could you give evidence that **supports** the idea that married couples are **more likely** to be committed to [other tasks]?" where hedge phrases are emboldened and brackets denote fightin' words (which are selected separately within each debate). We use a list of hedging cues curated by Tan et al. (2016) and derived from Hyland (1996) and Hanauer et al. (2012). The conjoined feature is the count of the user's sentences in a debate where a fightin' word cooccurs with a hedge phrase in a sentence.

3.3.4 Aggregating Earlier Debates

Since we consider the full history of a debater when estimating their skill level, we opt to aggregate the textual features over each debate. We do so by taking a weighted sum of the feature vectors of the previous debates. We consider four weighting schemes, none of which have free parameters, to preserve interpretability. Let f be any one of the features in the linguistic profile, a function from a single debate to a scalar.

[9] The method estimates log-odds of words given a side, with Dirichlet smoothing, and returns the words with the highest log-odds for each side.

Feature	Description	
Elo Score	Traditional Elo score calculated and updated. Updated traditionally, not averaged as in §3.3.4.	n/a
Length	Number of words this user uttered in the debate.	↑↑↑
Part of speech	Count of each noun, verb, adjective, preposition, adverb, or pronoun from the participant in the entire debate.	Noun:(↑↑↑) Adj:(↑↑↑)
Flesch reading ease	Measure of readability given the number of sentences in a document and the number of words in each sentence (Kincaid et al., 1975).	↑↑
Emotional words	Cues that indicate a positive or negative emotion (Tausczik and Pennebaker, 2010).	Pos:(↑↑↑) Neg:(↑↑↑)
Links	Links to external websites outside of `debate.org`. This feature operationalizes the number of sources a debater used.	
Questions	The number of questions the user asked in the debate.	↓↓↓
Quotations	The number of quotations the user included in the debate.	
Hedging	The number of phrases that soften a statement by adding uncertainty (Hyland, 1996; Hanauer et al., 2012).	
Fightin' words	The number of instances of words most strongly associated with either debater (Monroe et al., 2008).	↑↑↑
H∧FW	The number of cooccurences of hedging and fightin' words, described in §3.3.3.	↑↑

Table 5: Debate-level features used in estimating skill levels. Aside from Elo, the features are a part of the user's linguistic profile. The third column represents statistical significance levels in comparing winners and losers' features (independently) with Bonferroni correction: ↑ is $p < 0.05$, ↑↑ is $p < 0.01$, ↑↑↑ is $p < 0.001$.

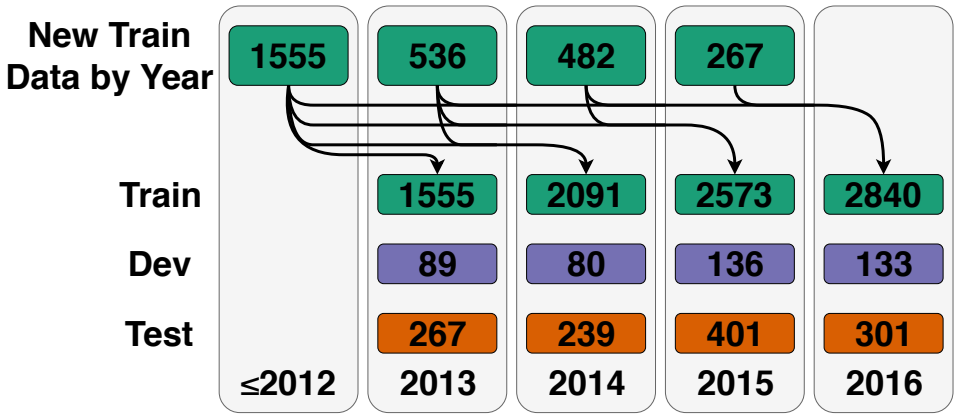

Figure 3: We split data into training/development/test based on year. This chart shows the number of debates in each subset of the data. We note that, for training, we use all the training debates from previous years (e.g., if we were to test on 2015, we would train using the training splits from 2012, 2013, and 2014). Each instance in this figure corresponds to a debater.

1. **Exponential growth:** the most recent debates are most indicative of skill, $\sum_{i=1}^{t-1} \frac{f(d_i^A)}{2^{t-i}}$. We take this to be the most intuitive choice, experimentally comparing against the alternatives below to confirm this intuition.
2. **Simple average:** each earlier debate's feature vector is weighted equally, $\frac{1}{t-1}\sum_{i=1}^{t-1} f(d_i^A)$.
3. **Exponential decay:** the first debates are most indicative of skill, $\sum_{i=1}^{t-1} \frac{f(d_i^A)}{2^i}$.
4. **Last only:** only the single most recent debate matters, $f(d_{t-1}^A)$ (an extreme version of "exponential growth").

In each variation of our method, some or all of the linguistic profile features are aggregated (using one of the four weighted averages), then applied to predict debate outcomes through logistic regression. We note that if our sole aim were to maximize predictive accuracy, we might explore much richer linguistic profiles, perhaps learning word embeddings for the task and combining them using neural networks and enabling interactions between a debate's two participants' profiles.[10] In this study, we seek to estimate skill but also to understand it, so our focus remains on linear models.

3.4 Experimental Setup

We validate our skill models with a binary classification task, specifically forecasting which of two debaters will win a debate (without looking at the content of the debate). Here we remove any debates where someone forfeits or there is no winner. We then considered debates

[10] Indeed, in preliminary experiments we *did* explore using a recurrent neural network instead of a fixed weighted average, but it did not show any benefit, perhaps owing to the relatively small size of our dataset.

where each debater has completed at least five of the remaining debates. As discussed in §3.2, the winner is taken to be the debater receiving the most "more convincing argument" votes from observers who did not initially agree with them. We create four training/evaluation splits of the data, using debates from 2013, 2014, 2015, and 2016 as evaluation sets (i.e., development and test) and debates prior to the evaluation year for training. Figure 3 shows the number of debates in each split. We note that our training sets are cumulative. For example, if we were to test on 2015 data, we would use the 2012, 2013, and 2014 training as training data.

Since we do not test on 2012 data or train on 2016 data due to the low number of debates before 2012 and after 2016, we treat the whole of 2012 as training data and 2016 as development and test. We report the accuracy for each run.

We compare several predictors:[11]

- **Full model**: our model with all features, as described in §3.1, and (except where otherwise stated) the exponential growth weighting. This model combines linguistic profiles from earlier debates with a conventional Elo score.
- **Full model with point difference**: our full model as described above, but we scale the Elo gain by the point difference as described in Equation 4.
- **Linguistic profile only**: our model with exponential growth weighting (except where otherwise stated), but ablating the Elo feature. This model is most similar to those found in prior literature (Zhang et al., 2016; Tan et al., 2018; Wang et al., 2017).
- **Elo**: the prediction is based solely on the Elo score calculated just before the debate. This is equivalent to ablating the linguistic profiles from our model.
- **Final Elo oracle**: the prediction is based solely on the two debaters' *final* Elo scores (i.e., using all debates from the past, present, and future).
- **Current debate text oracle**: a model that uses the linguistic profile derived just from the current debate. While this model is most similar to previous work, it is not a fair estimate of skill (since it ignores past performance). We therefore view it as another oracle.
- **Majority choice**: a baseline that always predicts that the contender will win.[12]

3.5 Results

In this section, we first show that the expertise of a debater can be better estimated with the linguistic profile, and then analyze the contribution of different components. We further examine the robustness of our results by controlling for additional variables.

[11] We use ℓ_2 regularization in our models with the linguistic profile features.

[12] In this dataset, contenders win nearly 59% of the time, a fact frequently discussed in the `Debate.org` community; see `http://ddo.wikia.com/wiki/Contender_Advantage`. The contender advantage is sometimes attribted to having the "final say," or to the fact that contenders choose the instigators they wish to debate.

3.5.1 Prediction Performance

We first present our results with what we consider our best models, i.e., our full model (with point differences), which consists of all features and uses the exponentially growing weight.

Importance of linguistic features. We see from Figure 4 that the full model outperforms the standard Elo baseline. The gap between the two models suggests that the addition of the linguistic profile contributes to the performance of the model and therefore plays a useful role in skill estimation. Moreover, the linguistic profile only model shows that the linguistic profile features are not only useful, but have at least as strong predictive power as Elo alone. By only using the linguistic features aggregated over the course of a debater's history, *without knowing winning records*, we can forecast at least as well as the Elo baseline.

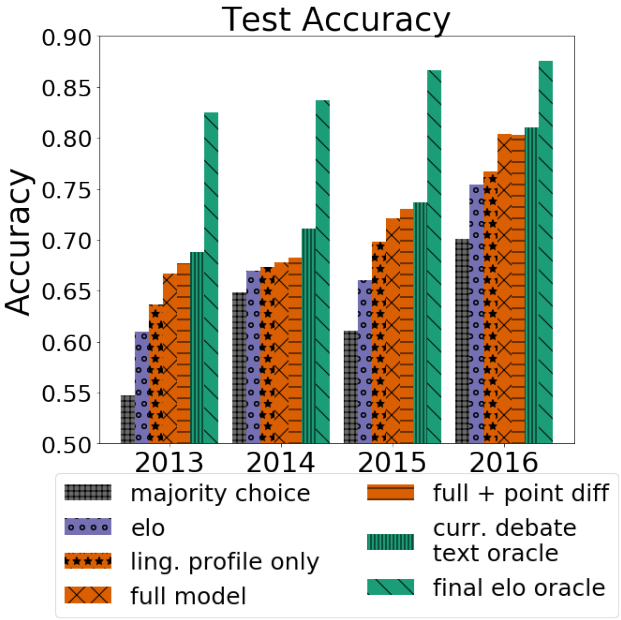

Figure 4: Our results for the prediction task. Our full model outperforms the Elo baseline and approaches the current debate text oracle.

Importance of multiple debates. We also note that our full model only performs slightly worse than the current debate text oracle despite the current debate text model directly observing the content of the debate. This result implies that using information from only previous debates has at least similar predictive strength to information from the debate at hand. Moreover, the large gap between the final Elo and current debate text oracles implies that a user's skill is evidenced by more than the content of a single debate. These results further demonstrate the importance of accounting for debaters' prior history.

Magnitude of victory might matter. Our full model with the point difference scaled Elo gain does roughly as well or slightly better than our normal full model. As the focus of this paper is on incorporating linguistic profiles, we use the **full model without the point difference scaling** for analyses in the rest of the paper.

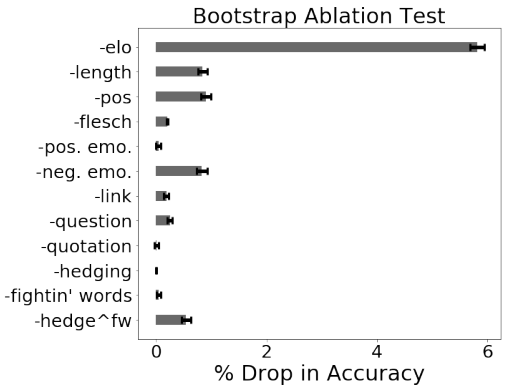

Figure 5: Our bootstrap on the feature ablations. We record the average drop in performance across 100 iterations and tested on the 2016 test set. Higher means a larger drop in performance.

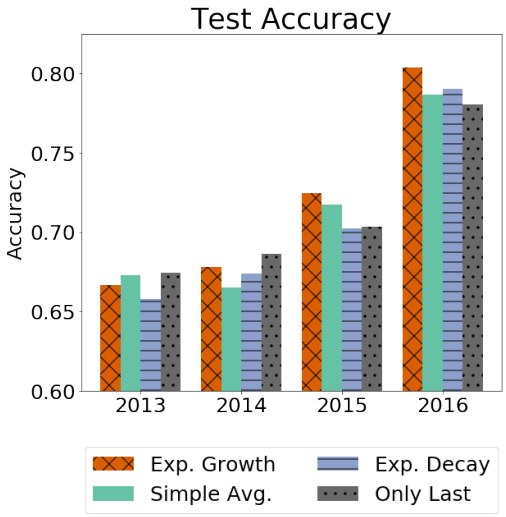

Figure 6: Comparison in performance for the four ways we aggregate features over time.

3.5.2 Feature Ablations

We inspect the contribution of each feature by removing each one from the model. Then, for each feature, we perform a bootstrap test over the last year of data (trained on 2012–2015 data; tested on 2016). At each iteration, we sample 1000 training examples to train on, but fix the test set across iterations. We then train our full model alongside several other models, each with a feature ablated, on the sample. We track the drop in performance between our full model and each of our other models. We record the average performance over 100 iterations for comparison.

From Figure 5, we find that removing Elo results in the most severe drop in performance (5.8%). Ablating part-of-speech, negative emotion, and length from our model had a moderate effect on performance. Surprisingly, we find that, although the H∧FW feature is the overlap between hedge cues and fightin' words, the latter two features individually contribute very little to the performance of our model compared to H∧FW.

3.5.3 Combining Prior Debate Features

As described in §3.3.4, we explore several ways of combining features over the past debates. By inspecting how these different aggregation functions might differ in performance, we hope to find out whether or not some debates are more important than others, if recency matters at all, and if some history is important at all. We do so by 1) giving the last debates more weight (**exponential growth**), 2) giving all weights equal weight (**simple average**), 3) giving the first debates more weight (**exponential decay**), and 4) giving all the weight to the last debate (**last debate only**).

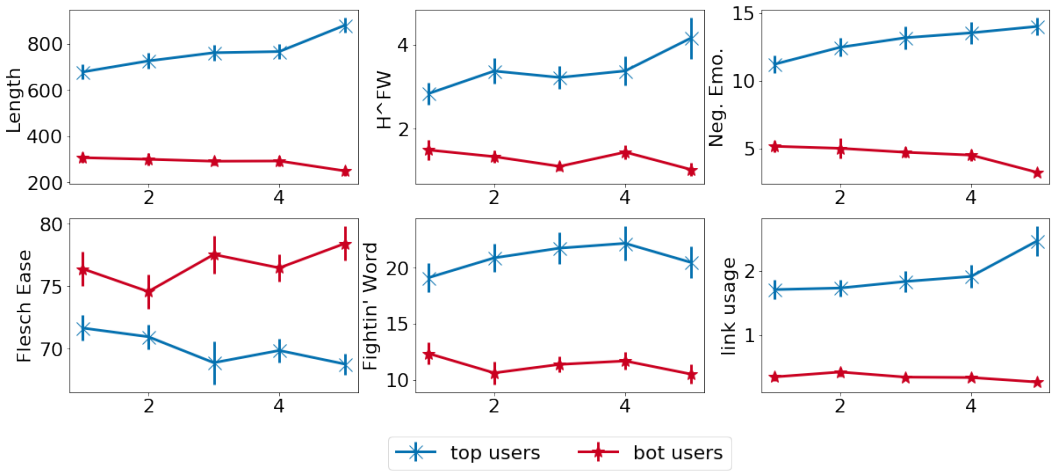

Figure 7: Feature measurements averaged across each history quintile. We see that there is a general trend for those who eventually become the best debaters to improve on these measures while the bottom users stagnate. The error bars represent the 95% confidence intervals.

More debates help. When using only the last debate's features and ignoring all previous debates, our performance is initially very good in the in years 2013 and 2014 (when our training sets are smaller and histories shorter). However, as debate histories become richer in the later years, last-only's performance drops in comparison to that of the growing weight aggregation. These results match our intuition: with more experienced users, considering more debates gives us a better gauge of how a debater performs and a debater's more recent debates give a better snapshot of the user's current skill level.

Later debates matter more. As hinted in our last-only results, giving the last debates more weight does best in all four years. In light of its performance compared to the simple mean and decaying weight settings, our results imply that not all debates contribute equally to a debater's skill under our model. Indeed, our results show that the most recent debates are also the most important for estimating a debater's skill rating.

3.6 Language Change over Time: Experts Improve and Dunces Stagnate

Our findings from §3.5.3 imply that users tend to experience some change over time and that these changes are helpful for forecasting who in a debate is more likely to win. In this section, we explore whether debaters' linguistic tendencies change over time by tracking their use of features over the course of their debate history.

We examine how language use changes over time for the best and worst users. To do so, we divide each user's debate history into quintiles. For each quintile, we average the same features we use in the linguistic profile that we use in our model (Table 5). We take the top 100 and bottom 100 users ranked by our model to see how the trajectories change over time.

Figure 7 shows that best and worst users have different linguistic preferences even from the beginning of their debating activities. The best debaters have a higher feature count in every case except for the Flesch reading ease score. However, the best users do seem to improve over time in length, H∧FW, use of emotional cues, and link use, which mostly correlates with our feature ablation in Figure 5 ($p < 0.05$ after Bonferroni correction). In contrast, the worst users do not seem to experience any significant change over time except for the length of their rounds and negative emotional cue use. In those cases, the worst users seem to worsen over time.

3.7 Related Work

In addition to the most relevant studies mentioned so far, our work is related to three broad areas: skill estimation, argumentation mining, and studies of online debates.

Skill estimation. Ranking player strength has been studied extensively for sports and for online matchmaking in games such as *Halo* (Herbrich et al., 2007). The Bradley-Terry models (of which Elo is an example) serve as a basis for much of the research in learning from pairwise comparisons (Bradley and Terry, 1952; Elo, 1978). Another rating system used for online matchmaking is Microsoft's Trueskill™ rating system (Herbrich et al., 2007), which assumes performance is normally distributed. Neural networks have recently been explored (Chen and Joachims, 2016; Menke and Martinez, 2008; Delalleau et al., 2012), incorporating player or other contextual game features from previous games at the cost of interpretability of those features.

Argumentation and persuasion. Past studies have noted the persuasiveness of stylistic effects such as phrasing or linguistic accommodation. For example, Danescu-Niculescu-Mizil et al. (2012) showed that, in a pool of people vying to become an administrator of a website, those who were promoted tended to coordinate more than those who were not. Similarly, other works define and discuss power relations over discussion threads such as emails (Prabhakaran and Rambow, 2014, 2013). Additionally, Tan et al. (2018) explored how debate quotes are selected by news media. They found that linguistic and interactive factors of an utterance are predictive of whether or not it would be quoted. Prabhakaran et al. (2014) also studied political debates and found that a debater's tendency to switch topics correlates with their public perception. Argumentation has also been studied extensively in student persuasive essays and web discourse (Persing and Ng, 2015; Ong et al., 2014; Song et al., 2014; Stab and Gurevych, 2014; Habernal and Gurevych, 2017; Lippi and Torroni, 2016). Most relevant to our work on how users improve over time, Zhang et al. (2017) study how one document may improve over time through annotated revisions. While our work examines users' linguistic change across multiple debates, they focus on how a user improves a single document over multiple revisions.

Online debates. There has also been recent work in characterizing specific arguments in online settings in contrast to our focus on the debaters themselves. For example, Somasundaran and Wiebe (2009), Walker et al. (2012), Qiu et al. (2015), and Sridhar et al. (2015) built systems for identifying the *stances* users take in online debate forums. Lukin et al. (2017) studied how persuasiveness of arguments depends on personality factors of the audience.

Other researchers have focused on annotation tasks. For example, Park and Cardie (2014) annotated online user comments to identify and classify different propositions. Hidey et al. (2017) annotated comments from the *changemyview* subreddit, a community where participants ask the community to change a view they hold. Likewise, Anand et al. (2011) annotated online blogs with a classification of persuasive tactics. Inspired by Aristotle's three modes of persuasion (ethos, pathos, and logos), their work annotates claims and premises within the comments. Habernal and Gurevych (2016) used crowdsourcing to study what makes an argument more convincing. They paired two similar arguments and asked annotators to indicate the more convincing one. This framework allowed them to study the flaws in the less convincing arguments. The annotations they produced offer a rich understanding of arguments which, though costly, can be useful as future work.

4 Proposed Work

In ongoing work, we investigate the effects of temporal misalignment, when an NLP model is trained on data from one time period, but deployed on data from another. This work is inspired from three observations:

1. in §3, we concluded that a debater's use of language can evolve over time.
2. in §2, we found that, while syntactically correct, generations by SciGen often failed to explain concurrent work well. That is, SciGen struggled due to training data missing certain types of demonstrations.
3. Both §2 and §3 investigate phenomena specific to scientific articles and debaters respectively.

Changes in the ways language is used over time are widely attested (Labov, 1994; Altmann et al., 2009; Eisenstein et al., 2014). Consequently, NLP systems deployed in the present are evaluated on types of data not present during training time. How these changes will affect NLP systems, and in particular their long-term performance, is not as well understood.

In today's pretraining-finetuning paradigm, this misalignment can affect a pretrained language model—a situation that has received recent attention (Jaidka et al., 2018; Lazaridou et al., 2021; Peters et al., 2018; Raffel et al., 2020; Röttger and Pierrehumbert, 2021)—or the finetuned task model, or both. We suspect that the effects of temporal misalignment will vary depending on the genre or domain of the task's text, the nature of that task or application, and the specific time periods.

Task Curation Different domains may be subject to temporal misalignment at differing rates. For example, social media platforms like Twitter have been known to experience frequent fluctuation in language over time with phenomena such as the introduction or diffusion of new words (Eisenstein et al., 2014; Tamburrini et al., 2015; Wang and Goutte, 2017). In contrast, food reviews may experience less change.

To facilitate in-depth study of temporal misalignment, we consider eight tasks, ranging from summarization to entity typing, a subproblem of entity recognition (Grishman and Borthwick, 1999). Each of these tasks have datasets that span at least five years. Notably, these task datasets span four different domains: social media, scientific articles, news, and reviews.

Effects on Finetuning We find that the amount of performance degradation can vary by task; in some cases the degradation can be severe (as much as a 40-point drop in F_1 score over 5 years). The level of degradation appears greater than reported in prior studies (Röttger and Pierrehumbert, 2021). Notably, even two tasks within the same domain can deteriorate at very different rates. The high levels of variation suggests that temporal misalignment also affects performance through the labeled datasets.

Effects on Pretraining Temporal misalignment seems to have an effect on language modeling perplexity on all 4 domains. Indeed, continued pretraining on a language model with temporally aligned data helped alleviate performance degradation due to temporal misalignment. In contrast, unsupervised adaptation did not mitigate the effects of temporal misalignment in the downstream task. This finding underscores the importance of temporally-aligned labeled data.

Calibrating Staling Models Our ongoing work suggests that temporally-aligned labeled data is essential as the gap between training and testing time periods grow larger. One method of tackling this problem could be to calibrate our models – that is, to have our models provide a confidence alongside the prediction.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of ICWSM*.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, 4(11):e7678.

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. 2011. Believe me? we can do this! annotating persuasive acts in blog text. In *Proceedings of the Workshops at AAAI*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Ivy Cao, Zizhou Liu, Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2021. Quantifying the effects of COVID-19 on restaurant reviews. In *Proceedings of the International Workshop on Natural Language Processing for Social Media*, Online. Association for Computational Linguistics.

Shuo Chen and Thorsten Joachims. 2016. Predicting matchups and preferences in context. In *Proceedings of SIGKDD*.

Arman Cohan and Nazli Goharian. 2017. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *SIGIR*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*.

Olivier Delalleau, Emile Contal, Eric Thibodeau-Laufer, Raul Chandias Ferrari, Yoshua Bengio, and Frank Zhang. 2012. Beyond skill rating: Advanced matchmaking in ghost recon online. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(3):167–177.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of NAACL-HLT*.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *Proceedings of WWW*.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11).

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Ralph Grishman and Andrew Borthwick. 1999. A maximum entropy approach to named entity recognition.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of EMNLP*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

David A Hanauer, Yang Liu, Qiaozhu Mei, Frank J Manion, Ulysses J Balis, and Kai Zheng. 2012. Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *Proceedings of AMIA*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueskillTM: a Bayesian skill rating system. In *Proceedings of NIPS*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the Workshop on Argument Mining*.

Ari Holtzman, Jan Buys, M. Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Ken Hyland. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *ACL*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*.

J Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *CNTECHTRA Research Branch Report 8-75*.

W. Labov. 1994. Principles of linguistic change.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, et al. 2021. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*.

Stephanie M. Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of EACL*.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. In *ACL*, Online. Association for Computational Linguistics.

Joshua E. Menke and Tony R. Martinez. 2008. A Bradley–Terry artificial neural network model for individual ratings in group competitions. *Neural Computing and Applications*, 17(2):175–186.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the Workshop on Argumentation Mining*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the Workshop on Argumentation Mining*.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of ACL*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of EMNLP*.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of EMNLP*.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of IJCNLP*.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of ACL*.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Coling 2008*. Coling 2008 Organizing Committee.

Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of SDM*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners. *OpenAI Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019b. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:1–67.

Shruti Rijhwani and Daniel Preoţiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *ACL*.

Paul Röttger and Janet B Pierrehumbert. 2021. Temporal adaptation of bert and performance on downstream document classification: Insights from social media. *arXiv preprint arXiv:2104.08116*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *ACL*, pages 255–269, Online.

Nate Silver. 2015. How our NFL predictions work. `https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/`. Accessed: 2019-02-30.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL-IJCNLP*.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the Workshop on Argumentation Mining*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*.

Nadine Tamburrini, Marco Cinnirella, Vincent AA Jansen, and John Bryden. 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84–89.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled. In *Proceedings of ACL*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Chenhao Tan, Hao Peng, and Noah A. Smith. 2018. You are no Jack Kennedy: On media selection of highlights from presidential debates. In *Proceedings of WWW*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of NAACL-HLT*.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Yunli Wang and Cyril Goutte. 2017. Detecting changes in twitter streams using temporal clusters of hashtags. In *Proceedings of the Events and Stories in the News Workshop*.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *NeurIPS*.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of ACL*.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of NAACL-HLT*.

Michael J.Q. Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. *EMNLP*.