Square One Bias in NLP:
Towards a Multi-Dimensional Exploration of the Research Manifold

Anonymous ACL submission

Abstract

The first NLP experiment many researchers performed in their careers likely involved training a standard architecture on labeled **English** data and optimizing for **accuracy**, without accounting for other dimensions such as **fairness**, **interpretability**, or computational **efficiency**. We show through surveys that this is indeed the case and refer to it as the *square one* experimental setup. NLP research often goes beyond the square one setup, e.g, focusing not only on accuracy, but also on fairness or interpretability, but typically *only* along a single dimension. Most work focused on multilinguality, for example, considers only accuracy; most work on fairness or interpretability considers only English; and so on. We show this through manual classification of recent NLP research papers and ACL Test-of-Time award recipients. Such one-dimensionality of most research means we are only exploring a fraction of the NLP research search space. We provide historical and recent examples of how the square one bias has led researchers to draw false conclusions or make unwise choices, point to promising yet unexplored directions on the research manifold, and make practical recommendations to enable more multi-dimensional research.

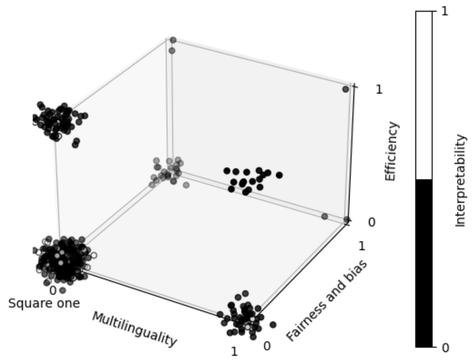

Figure 1: Visualization of contributions of ACL 2021 oral papers along 4 dimensions: multilinguality, fairness and bias, efficiency, and interpretability (indicated by color). Most work is clustered around the Square One or along a single dimension.

1 Introduction

Our categorization of objects, say screwdrivers or NLP experiments, is heavily biased by early prototypes (Sherman, 1985; Das-Smaal, 1990). If the first 10 screwdrivers we see, are red and for hexagon socket screws, this will bias what features we learn to associate with screwdrivers. Likewise, if the first 10 NLP experiments we see or conduct are in sentiment analysis, this will likely also bias how we think of NLP experiments in the future.

In this position paper, we postulate that we can meaningfully talk about *the* prototypical NLP experiment, and that *the existence of such an experimental prototype steers and biases the research dynamics in our community*. We will refer to this prototype as NLP's Square One—and to the bias that follows from it, as the Square One Bias. We argue this bias manifests in a particular way: Since research is a creative endeavor, and researchers aim to push the research horizon, *most research papers in NLP go beyond this prototype, but only along a single dimension at a time*. Such dimensions include multilinguality, efficiency, fairness, and interpretability, among others. The effect of the Square One Bias is to baseline novel research contributions, rewarding work that differs from the prototype in a concise, one-dimensional way.

We present several examples of this effect in practice. For instance, analyzing the contributions of ACL 2021 papers along 4 dimensions, we observe that most work is either clustered around the Square One or makes a contribution along a single dimension (see Figure 1). Multilingual work typically disregards efficiency, fairness, and interpretability. Work on efficient NLP typically only performs evaluations on English datasets, and disregards fairness and interpretability. Fairness and interpretability work is also mostly limited to English, and tends to disregard efficiency concerns.

We argue that the SQUARE ONE BIAS has several negative effects, most of which amount to the study of one of the above dimensions being biased by ignoring the others. Specifically, by focusing only on exploring the edges of the manifold, we are not able to identify the non-linear interactions between different research dimensions. We highlight several examples of such interactions in Section 3. Overall, we encourage a focus on combining multiple dimensions on the research manifold in future NLP research, and delve deeper into studying their (linear and non-linear) interactions.

Contributions. We first establish that we can meaningfully talk about the prototypical NLP experiment, through a series of surveys and annotation experiments. This prototype amounts to applying a standard architecture to an English dataset and optimizing for accuracy or F1. We discuss the impact of this prototype on our research community, and the bias it introduces. We then discuss the negative effects of this bias. We also list work that has taken steps to overcome the bias. Finally, we highlight blind spots and unexplored research directions and make practical recommendations, aiming to inspire the community towards conducting more 'multi-dimensional' research (see Figure 1).

2 Finding the Square One

In order to determine the existence and nature of a SQUARE ONE, we seek to identify commonalities between students' first exposure to NLP. In most cases, we expect such exposure to occur during an introductory NLP course.

Questionnaire for NLP Teachers. We therefore created a short questionnaire, which we sent to a geographically diverse set of teachers, including first authors from the last Teaching NLP workshop (Jurgens et al., 2021), asking about the first experiment that they presented in their NLP 101 course. We received 71 responses in total. Our first question was: *The last time you taught an introductory NLP course, what was the first task you introduced the students to, or that they had to implement a model for?* The relative majority of respondents (31.9%) said *sentiment analysis*, while 10.1% indicated topic classification.[1] More importantly, we also asked them about the language of the data used in the experiment, and what metric they optimized for. More than three quarters of respondents reported that they used *English* language training and evaluation data and more than three quarters of the respondents asked the students to optimize for *accuracy or F1*. The choice of using English language datasets is particularly interesting in contrast to the native languages of the teachers and their students: In around two thirds of the classes, most students shared an L1 language that was not English; and less than a quarter of the teachers were L1 English speakers themselves. In summary, the prototypical NLP 101 experiment, according to our survey, is on an **English classification** task with **accuracy or F1** as performance metric. None of the respondents reported to have optimized for fairness, interpretability or efficiency metrics.

Year	Book	Language	Task
1999	Manning and Schütze (1999)	English-French	Alignment
2009	Jurafsky and Martin (2009)	English	LM
2009	Bird et al. (2009)	English	Name cl.
2013	Søgaard (2013)	English	Doc.cl.
2019	Eisenstein (2019)	English	Doc.cl.

Table 1: First experiments in NLP textbooks. The objective across all books is optimizing for performance (AER, perplexity, or accuracy), rather than fairness, interpretability or efficiency.

Classification of NLP Textbooks. What, then, are the prototypical NLP experiments in undergraduate and graduate textbooks? We list five exemplary NLP textbooks, spanning 20 years, in Table 1. We observe that they, like the teachers in our survey, take the same point of departure: an English-language experiment in which we use supervised learning techniques to optimize for a standard performance metric, e.g., perplexity or error. We note an important difference, however: While the first four books largely ignore issues relating to fairness, interpretability, and efficiency, the most recent NLP textbook in Table 1 (Eisenstein, 2019) discusses efficiency (briefly) and fairness (more thoroughly).

ACL 2021 Oral Papers. We now seek to quantify whether the same bias exists in contemporary research. To this end, we annotate the 461 papers that were presented orally at ACL 2021, a representative cross-section of the 779 papers accepted to the main conference. We focus on 4 dimensions along which papers may differ from a prototypical

[1] The remaining responses included NER, language modeling, language identification, hate speech detection, etc.

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%

Table 2: The number of ACL 2021 oral papers (top row) and of papers in each area (bottom rows) as well as the fractions that only evaluate on English, only use accuracy / F1, make contributions along one of four dimensions, and make contributions along more than a single dimension (from left to right).

NLP experiment: multilinguality, fairness and bias, efficiency, and interpretability.[2] Compared to prior work that annotates the values of ML research papers (Birhane et al., 2021), we are not concerned with a paper's motivation but whether its *practical contributions* constitute a meaningful departure from SQUARE ONE. For each paper, we annotate whether it makes a contribution along each dimension[3] as well as the languages and metrics it employs for evaluation.

The general statistics from our classification of ACL 2021 papers are presented in Table 2. In addition, we highlight the statistics for the conference *areas* (tracks) corresponding to 3 of the 4 dimensions,[4] as well as for the top 5 areas with the most papers. We show statistics for the remaining areas in Appendix A.1. We additionally visualize their distribution in Figure 1. Overall, almost 70% of papers evaluate only on English, clearly highlighting a lack of language diversity in NLP (Bender, 2011; Joshi et al., 2020). Almost 40% of papers only evaluate using accuracy and/or F1, foregoing metrics that may shed light on other aspects of model behavior. Regarding work that moves from the SQUARE ONE, most papers make a contribution in terms of efficiency, followed by multilinguality. However, most papers that evaluate on multiple languages are part of the corresponding MT and Multilinguality track. Despite being an area receiving increasing attention (Blodgett et al., 2020), only 6.3% of papers evaluate the bias or fairness of a method. Overall, *only 6.1% of papers* make a contribution along two or more of these dimensions. Among these, joint contributions on both multilinguality and efficiency are the most common (see Figure 1). In fact, 22 of the 26 two-or-more-dimensional papers focus on efficiency, and 17 of these on the combination of multilinguality and efficiency. This means less than 1% of the ACL 2021 papers consider combinations of (two or more of) multilinguality, fairness and interpretability. We find this surprising, given these topics are considered among the most popular topics in the field.

Some areas have particularly concerning statistics. A large majority of research work in dialog (90.5%), summarization (91.7%), sentiment analysis (100%), and language grounding (100%) is done only on English; however, ways of expressing sentiment (Volkova et al., 2013; Yang and Eisenstein, 2017; Vilares et al., 2018) and visually grounded reasoning (Liu et al., 2021; Yin et al., 2021) do vary across languages and cultures. Systems in the top tracks tend to evaluate on efficiency, but in general do not consider fairness or interpretability of the proposed methods. Even the creation of new resources and evaluation sets (cf., Resource and Evaluation in Table 2) seems to be directed towards rewarding and enabling SQUARE ONE experiments; favoring English (77.1%), and with modest efforts on other dimensions. Notably, we only identified a single paper that considers three dimensions (Renduchintala et al., 2021). This paper considers gender bias (Fairness) in relation to speed-quality (Efficiency) trade-offs in multilingual machine translation (Multilinguality). Finally, we observe that best-paper award winning papers

[2] Other dimensions that could be considered in future work are robustness, multimodality, and privacy, among others.

[3] For multilinguality, we consider papers that evaluate on 3 languages, or 4 languages if they focus on MT (as the standard MT experiment includes two languages). For fairness and bias, we consider papers that improve fairness in a specific setting or analyze the bias of a method, e.g. regarding gender. For efficiency, we consider papers that analyze memory, speed, or computational complexity. For interpretability, we consider papers that interpret or explain a model's predictions.

[4] Unlike EACL 2021, NAACL-HLT 2021 and EMNLP 2021, ACL 2021 had no area associated with efficiency. To compensate for this, we annotated the 20 oral papers of the "Efficient Models in NLP" track at EMNLP 2021 (see Appendix A.2).

Year	Paper	Language	Metric
1995	Grosz et al. (1995)	English	n/a
1995	Yarowsky (1995)	English	acc.
1996	Berger et al. (1996)	English	acc.
1996	Carletta (1996)	n/a	n/a
2010	Baroni and Lenci (2010)	English	acc.
2010	Turian et al. (2010)	English	F_1
2011	Taboada et al. (2011)	English	acc.
2011	Ott et al. (2011)	English	acc./F_1

Table 3: Test-of-Time Award 2021-22 papers

are *not* more likely to consider more than one of the four dimensions. Only 1 in 8 papers did; the best paper (Xu et al., 2021), like most two-dimensional ACL 2021 papers, considered multilinguality and efficiency.

Test-of-Time Award Recipients. Current papers provide us with a snapshot of *actual* current research practices, but the one-dimensionality of the best paper award winning papers at ACL 2021 suggest the SQUARE ONE BIAS also biases what we *value* in research, i.e., our perception of *ideal* research practices. This can also be seen in the papers that have received the ACL Test-of-Time Award in the last two years (Table 3). Seven in eight papers included empirical evaluations performed exclusively on English data. Six papers were exclusively concerned with optimizing for accuracy or F_1.

Blackbox NLP Papers. Finally, we check if more multi-dimensional papers were presented at a workshop devoted to one of the above dimensions. The rationale would be that if everyone at a workshop already explores one of these dimensions, maybe including another is a way to have an edge over other submissions. Unfortunately, this does not seem to be the case. We manually annotated the first 10 papers in the Blackbox NLP 2021 program[5] that were available as pre-prints at the time of submission.[6] Of the 10 papers, only one included more than one dimension (Abdullah et al., 2021). This number aligns well with the overall statistics of ACL 2021 (6.1%). All the other Blackbox NLP papers only considered interpretability for English.

3 Square One Bias: Examples

In the following, we highlight both historical and recent examples touching on different aspects of research in NLP that illustrate how the gravitational attraction of the SQUARE ONE has led researchers to draw false conclusions, unconsciously steer standard research practices, or make unwise choices.

Architectural Biases. One pervasive bias in our models regards **morphology**. Many of our models were not designed with morphology in mind, arguably because of the poor morphology of English. Traditional n-gram language models, for example, have been shown to perform much worse on languages with elaborate morphology due to data sparsity problems (Khudanpur, 2006; Bender, 2011; Gerz et al., 2018). Such models were nevertheless more commonly used than more linguistically informed alternatives such as factored language models (Bilmes and Kirchhoff, 2003) that represent words as sets of features. Word embeddings have been widely used, in part because pre-trained embeddings covered a large part of the English vocabulary. However, word embeddings are not useful for tasks that require access to morphemes, e.g., semantic tasks in morphologically rich languages (Avraham and Goldberg, 2017).

While studies have demonstrated the ability of word embeddings to capture linguistic information in English, it remains unclear whether they capture the information needed for processing morphologically rich languages (Tsarfaty et al., 2020). A bias towards morphologically rich languages is also apparent in our tokenization algorithms. Subword tokenization performs poorly on languages with reduplication (Vania and Lopez, 2017), while byte pair encoding does not align well with morphology (Bostrom and Durrett, 2020). Consequently, languages with productive morphological systems also are disadvantaged when shared 'language-universal' tokenizers are used in current large-scale multilingual language models (Ács, 2019; Rust et al., 2021) without any further vocabulary adaptation (Wang et al., 2020; Pfeiffer et al., 2021).

Another bias in our models relates to **word order**. In order for n-gram models to capture inter-word dependencies, words need to appear in the n-gram window. This will occur more frequently in languages with relatively fixed word order compared to languages with relatively free word order (Bender, 2011). Word embedding approaches such as skip-gram (Mikolov et al., 2013) adhere to the same window-based approach and thus have similar weaknesses for languages with relatively free word order. LSTMs are also sensitive to word order and perform worse on agreement prediction in

[5] https://blackboxnlp.github.io/

[6] These annotations are made publicly available along with the rest of the data collected for this paper at url.

Basque, which is both morphologically richer and has a relatively free word order (Ravfogel et al., 2018) compared to English (Linzen et al., 2016). They have also been shown to transfer worse to distant languages for dependency parsing compared to self-attention models (Ahmad et al., 2019). Such biases concerning word order are not only inherent in our models but also in our algorithms. A recent unsupervised parsing algorithm (Shen et al., 2018) has been shown to be biased towards right-branching structures and consequently performs better in right-branching languages like English (Dyer et al., 2019). While the recent generation of self-attention based architectures can be seen as inherently order-agnostic, recent methods focusing on making attention more efficient (Tay et al., 2020) introduce new biases into the models. Specifically, models that reduce the global attention to a local sliding window around the token (Liu et al., 2018; Child et al., 2019; Zaheer et al., 2020) may incur similar limitations as their n-gram and word embedding-based predecessors, performing worse on languages with relatively free word order.[7]

The singular focus on maximizing a performance metric such as accuracy introduces a bias towards models that are expressive enough to fit a given distribution well. Such models are typically **black-box** and learn highly non-linear relations that are generally not interpretable. Interpretability is generally studied in papers focusing exclusively on this topic; a recent example is BERTology (Rogers et al., 2020). Studies proposing more interpretable methods typically build on state-of-the-art methods (Weiss et al., 2018) and much work focuses on leveraging components such as attention for interpretability, which have not been designed with that goal in mind (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). As a result, researchers eschew directions focusing on models that are intrinsically more interpretable such as generalized additive models (Hastie and Tibshirani, 2017) and their extensions (Chang et al., 2021; Agarwal et al., 2021) but which have so far not been shown to match the performance of state-of-the-art methods.

As most datasets on which models are evaluated focus on sentences or short documents, state-of-the-art methods restrict their input size to around 512 tokens (Devlin et al., 2019) and leverage methods that are **inefficient** when scaling to longer documents. This has led to the emergence of a wide range of more efficient models (Tay et al., 2020), which, however, are rarely used as baseline methods in NLP. Similarly, the standard pretrain-fine-tune paradigm (Ruder et al., 2019) requires separate model copies to be stored for each task, and thus restricts work on multi-domain, multi-task, multi-lingual, multi-subpopulation methods that is enabled by more efficient and less resource-intensive (Schwartz et al., 2020) fine-tuning methods (Houlsby et al., 2019; Pfeiffer et al., 2020)

In sum, (what we typically consider as) standard baselines and state-of-the-art architectures favor languages with some characteristics over others and are optimized only for performance, which in turn propagates the SQUARE ONE BIAS: If researchers study aspects such as multilinguality, efficiency, fairness or interpretability, they are likely to do so *with and for commonly used architectures* (i.e., often termed 'standard architectures'), in order to reduce (too) many degrees of freedom in their empirical research. This is in many ways a sensible choice in order to maximize perceived relevance—and thereby, impact. However, as a result, multilinguality, efficiency, fairness, interpretability, and other research areas *inherit the same biases*, which typically slip under the radar.

Annotation Biases. Many NLP tasks can be cast differently and formulated in multiple ways, and differences may result in different annotation styles. Sentiment, for example, can be annotated at the document, sentence or word level (Socher et al., 2013). In machine comprehension, answers are sometimes assumed to be continuous, but Zhu et al. (2020) annotate discontinuous spans. In dependency parsing, different annotation guidelines can lead to very different downstream performance (Elming et al., 2013). How we annotate for a task may interact in complex ways with dimensions such as multilinguality, efficiency, fairness, and interpretability. The Universal Dependencies project (Nivre et al., 2020) is motivated by the observation that not all dependency formalisms are easily applicable to all languages. Aligning guidelines across languages has enabled researchers to ask interesting questions, but such attempts may limit the analysis of outlier languages (Croft et al., 2017).

Other examples of annotation guidelines interacting with the above dimensions exist: Slight nuances in how annotation guidelines are formulated can

[7] An older work of Khudanpur (2006) argues that free word order is less of a problem as local order within phrases is relatively stable. However, it remains to be seen to what degree this affects current models.

lead to severe model biases (Hansen and Søgaard, 2021a) and hurt model fairness. In interpretability, we can use feature attribution methods and word-level annotations to evaluate interpretability methods applied to sequence classifiers (Rei and Søgaard, 2018), but we cannot directly use feature attribution methods to obtain rationales for sequence labelers. Annotation biases can also stem from the characteristics of the annotators, including their domain experience (McAuley and Leskovec, 2013), demographics (Jørgensen and Søgaard, 2021), or educational level (Al Kuwatly et al., 2020).

Annotation biases form an integral part of the SQUARE ONE BIAS: In NLP experiments, we commonly rely on the same pools of annotators, e.g., computer science students, professional linguists, or MTurk contributors. Sometimes these biases percolate through reuse of resources, e.g., through human or machine translation into new languages. Examples of such recycled resources include Conneau et al. (2018) and Kassner et al. (2021), among others. Even when such translation-based resources resonate with syntax and semantics of the target language, and are fluent and natural, they still suffer from *translation artefacts*: they are often target-language surface realizations of source-language-based conceptual thinking. As a consequence, evaluations of cross-lingual transfer models on such data typically overestimate their performance as properties such as word order and even the choice of lexical units are inherently biased by the source language (Vanmassenhove et al., 2021). Put simply, the choice of the data creation protocol, e.g., translation-based versus data collection directly in the target language (Clark et al., 2020) can yield profound differences in model performance for some groups, or may have serious impact on the interpretability or computational efficiency (e.g., sample efficiency) of our models.

Selection Biases. For many years, the English Penn Treebank (Marcus et al., 1994) was an integral part of the SQUARE ONE of NLP. This corpus consists entirely of newswire, i.e., articles and editorials from the Wall Street Journal, and arguably amplified the (existing) bias toward news articles. Since news articles tend to reflect a particular set of linguistic conventions, have a certain length, and are written by certain demographics, the bias toward news articles had an impact on the linguistic phenomena studied in NLP (Judge et al., 2006), led to under-representation of challenges with handling longer documents (Beltagy et al., 2021), and had impact on early papers in fairness (Hovy and Søgaard, 2015). Note how such a bias may interact in non-linear ways with efficiency, i.e., efficient methods for shorter documents need not be efficient for longer ones, or fairness, i.e., what mitigates gender biases in news articles need not mitigate gender biases in product reviews.

Protocol Biases. In the prototypical NLP experiment, the dataset is in the English language. As a consequence, it is also standard protocol in multilingual NLP to use English as a source language in zero-shot cross-lingual transfer (Hu et al., 2020). In practice, there are generally better source languages than English (Lin et al., 2019; Turc et al., 2021), and results are heavily biased by the common choice of English. For instance, effectiveness and efficiency of few-shot learning can be impacted by the choice of the source language (Pfeiffer et al., 2021; Zhao et al., 2021). English also dominates language pairs in machine translation, leading to lower performance for non-English translation directions (Fan et al., 2020), which are particularly important in multilingual societies. Again, such biases may interact in non-trivial ways with dimensions explored in NLP research: It is not inconceivable that there is an algorithm A that is more fair, interpretable or efficient than algorithm B on, say, English-to-Czech transfer or translation, but not on German-to-Czech or French-to-Czech.

Organizational Biases. The above architectural, annotation, selection and protocol biases follow from the SQUARE ONE BIAS, but they also conserve the SQUARE ONE. If our go-to architectures, resources, and experimental setups are tailored to some languages over others, some objectives over others, and some research paradigms over others, it is considerably more work to explore new sets of languages, new objectives, or new protocols. The organizational biases we discuss below may also reinforce the SQUARE ONE BIAS.

The organization of our conferences and reviewing processes perpetuates certain biases. In particular, both during reviewing and for later presentation at conferences, papers are organized in areas. Upon submission, a paper is assigned to a single area. Reviewers are recruited for their expertise in a specific area, which they are associated with. Such a reviewing system incentivizes papers that make contributions to the chosen area, in order to appeal to the reviewers of this area and implicitly penalizes papers that make contributions along multiple dimensions, as reviewers unfamiliar with the related areas may not appreciate their inter-disciplinary or inter-areal magnitude or value. Even new initiatives that seek to improve reviewing such as ARR[8] adhere to this area structure[9] and thus further the SQUARE ONE BIAS. A reviewing system that allows papers to be associated with multiple dimensions of research and that assigns reviewers with *complementary* expertise—similar to TACL[10]—would ameliorate this situation. Once a paper is accepted, presentations at conferences are organized by areas, limiting audiences in most cases to members of said area and thereby reducing the cross-pollination of ideas.[11]

Unexplored Areas of the Research Manifold. The discussed biases, which seem to originate from a SQUARE ONE BIAS, leave areas of the research manifold unexplored. Character-based language models are often reported to perform well for morphologically rich languages or on non-canonical text (Ma et al., 2020), but little is known about their fairness properties, and attribution-based interpretability methods have not been developed for such models. Annotation biases that stem from annotator demographics have been studied for English POS tagging (Hovy and Søgaard, 2015) or English summarization (Jørgensen and Søgaard, 2021), for example, but there has been very little research on such biases for other languages. While linguistic differences among genders is shared among some languages, genders differ in very different ways between other languages, e.g., Spanish and Swedish (Johannsen et al., 2015). We discuss important unexplored areas of the research manifold in §5, but first we briefly survey existing, multi-dimensional work, i.e., the counter-examples to our claim that NLP research is biased to one-dimensional extensions of the square one.

[8] aclrollingreview.org/

[9] www.2022.aclweb.org/callpapers

[10] transacl.org/index.php/tacl

[11] Another previously pervasive organizational bias, which is now fortunately being institutionally mitigated within the *ACL community through dedicated mentoring programs and improved reviewing guidelines, concerned penalizing research papers for their non-native writing style, where it was frequently suggested to the authors whose native language is not English to 'have their paper proofread by a native speaker'. As one hidden consequence, this attitude might have set a higher bar for the native speakers of minor and endangered languages working on such languages to put their research problems in the spotlight, that way also implicitly hindering more work of the entire community on these languages.

4 Counter-Examples

Most of the exceptions to our thesis about the 'one-dimensionality' of NLP research, in our classification of ACL 2021 Oral Papers, came from studies of **efficiency in a multilingual context**. Another example of this is Ahia et al. (2021), who show that for low-resource languages, weight pruning hurts performance on tail phenomena, but improves robustness to out-of-distribution shifts—this is not observed in the SQUARE ONE (high-resource) regime. There are also studies of **fairness in a multilingual context**. Huang et al. (2020), for example, show significant differences in social bias for multilingual hate speech systems across different languages. Zhao et al. (2020) study gender bias in multilingual word embeddings and cross-lingual transfer. González et al. (2020) also study gender bias, but by relying on reflexive pronominal constructions that do not exist in the English language; this is a good example of research that would not have been possible taking SQUARE ONE as our point of departure. Dayanik and Padó (2021) study adversarial debiasing in the context of a multilingual corpus and show some mitigation methods are more effective for some languages rather than others. Nozza (2021) studies multilingual toxicity classification and finds that models misinterpret non-hateful language-specific taboo interjections as hate speech in some languages. There has been much less work on other combinations of these dimensions, e.g., **fairness and efficiency**, but Hansen and Søgaard (2021b) show that weight pruning has disparate effects on performance across demographics; the min-max difference in group disparities is negatively correlated with model size. Renduchintala et al. (2021) also observe that techniques to make inference more efficient, e.g., greedy search, quantization, or shallow decoder models, have a small impact on performance, but dramatically amplify gender bias. In a rare study of **fairness and interpretability**, Vig et al. (2020) propose a methodology to interpret which parts of a model are causally implicated in its behavior. They apply this methodology to analyze gender bias in pre-trained Transformers, finding that gender bias effects are sparse and concentrated in small parts of the network.

5 Blind Spots

We identified several under-explored areas on the research manifold. The common theme is a lack of studies of how dimensions such as multilinguality, fairness, efficiency, and interpretability interact. We now summarize some open problems that we believe are particularly important to address: (i) While recent work has begun to study the trade-off between **efficiency and fairness**, this interaction remains largely unexplored, especially outside of the empirical risk minimization regime; (ii) **fairness and interpretability** interact in potentially many ways, i.e., interpretability techniques may affect the fairness of the underlying models (Agarwal, 2021), but rationales may also, for example, be biased toward certain demographics in how they are presented (Feng and Boyd-Graber, 2018; González et al., 2021); (iii) finally, **multilinguality and interpretability** seem heavily underexplored. While there exists resources for English for evaluating interpretability methods against gold-standard human annotations, there are, to the best of our knowledge, no such resources for other languages.[12]

6 Discussion

Is SQUARE ONE BIAS not the Flipside of Scientific Protocol? One potential argument *for* a community-wide SQUARE ONE BIAS is that when studying the impact of some technique t, say a novel regularization term, we want to compare some system with and without t, i.e., control for all other factors. To maximize impact and ease workload, it makes sense at first sight to stick to a system and experimental protocol that is familiar or well-studied. Always returning to the SQUARE ONE is a way to control for all other factors and relating new findings to known territory. The reason why this is *only seemingly a good idea*, however, is that the factors we study in NLP research, may be non-linearly related. The fact that t makes for a positive net contribution under one set of circumstances, does not imply that it would do so under different circumstances. This is illustrated most clearly by the research surveyed in §3. Ideally, we thus want to study the impact of t under as many circumstances as possible, but in the absence of resources to do so, it is a better (collective) search strategy to apply t to a *random* set of circumstances (within the space of relevant circumstances, of course).

Should Each Paper Aim to Cover All Dimensions? We believe that a researcher should aspire to cover as many dimensions as possible with their research. While this may not be possible in every instance due to various factors (lack of data, time, standardization, tooling, etc), considering the dimensions encourages us to think more holistically about our research and its final impact. It may also accelerate progress as follow-up work will already be able to build on the insights of multi-dimensional analyses of new methods. It will also promote the cross-pollination of ideas, which will no longer be confined to their own sub-areas. At the same time, multi-dimensional research requires researchers to become experts in multiple areas.

Practical Recommendations. What can we do to incentivize and facilitate multi-dimensional research? **i)** Currently, most NLP models are evaluated by one or two performance metrics, but we believe dimensions such as fairness, efficiency, and interpretability need to become integral criteria for model evaluation, in line with recent proposals of more user-centric leaderboards (Ethayarajh and Jurafsky, 2020; Ma et al., 2021). This requires new tools, e.g., to evaluate environmental impact (Henderson et al., 2020), as well as new benchmarks, e.g., to evaluate fairness (Koh et al., 2021). **ii)** We believe separate conference tracks (areas) lead to unfortunate silo effects and inhibit multi-dimensional research. Rather, we imagine conference submissions could provide a checklist with dimensions along which they make contributions, similar to reproducibility checklist. Reviewers can be assigned based on their expertise corresponding to different dimensions. **iii)** Finally, we recommend awareness of research prototypes and encourage reviewers and chairs to prioritize research that departs from prototypes in *multiple* dimensions, in order to explore new areas of the research manifold.

7 Conclusion

We identified the prototypical NLP experiment through surveys and annotation experiments. We highlighted the associated SQUARE ONE BIAS, which encourages research to go beyond the prototype in a single dimension. We discussed the problems resulting from this bias, by studying the area statistics of a recent NLP conference as well as by discussing historic and recent examples. We finally pointed to under-explored research directions and made practical recommendations to inspire more multi-dimensional research in NLP.

[12] We again note that there are other possible dimensions, not studied in this work, that can expose more blind spots: e.g., **fairness and multi-modality**, **multilinguality and privacy**.

References

Badr Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2021. How familiar does that sound? Cross-lingual representational similarity analysis of acoustic word embeddings. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 407–419.

Judit Ács. 2019. Exploring BERT's Vocabulary. *Blog Post*.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. 2021. Neural additive models: Interpretable machine learning with neural nets. In *Proceedings of NeurIPS 2021*.

Sushant Agarwal. 2021. Trade-offs between fairness and interpretability in machine learning. In *Proceedings of the IJCAI 2021 Workshop on AI for Social Good*.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of NAACL-HLT 2019*, pages 2440–2452.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. In *Proceedings of EACL 2017*, pages 422–426.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Iz Beltagy, Arman Cohan, Hannaneh Hajishirzi, Sewon Min, and Matthew E. Peters. 2021. Beyond paragraphs: NLP for long sequences. In *Proceedings of NAACL-HLT 2021: Tutorials*, pages 20–24.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Jeff Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 4–6.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *CoRR*, abs/2106.15590.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *CoRR*, abs/2005.14050.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. 2021. How interpretable and trustworthy are GAMs? In *Proceedings of KDD 2021*, pages 95–105.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse Transformers. *CoRR*, abs/1904.10509.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets universal dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63–75.

Edith A. Das-Smaal. 1990. Biases in categorization. volume 68 of *Advances in Psychology*, pages 349–386. North-Holland.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In

Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A critical analysis of biased parsers in unsupervised parsing. *CoRR*, abs/1909.09428.

Jacob Eisenstein. 2019. *Introduction to Natural Language Processing*. Adaptive Computation and Machine Learning series. MIT Press.

Jakob Elming, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Hector Martinez Alonso, and Anders Søgaard. 2013. Down-stream effects of tree-to-dependency conversions. In *Proceedings of NAACL-HLT 2013*, pages 617–626.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of EMNLP 2020*, pages 4846–4853.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. *arXiv preprint arXiv:2010.11125*.

Shi Feng and Jordan L. Boyd-Graber. 2018. What can AI do for me: Evaluating machine learning interpretations in cooperative play. *CoRR*, abs/1810.09648.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP 2018*, pages 316–327.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of EMNLP 2020*, pages 2637–2648.

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Victor Petrén Bach Hansen and Anders Søgaard. 2021a. Guideline bias in Wizard-of-Oz dialogues. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 8–14.

Victor Petrén Bach Hansen and Anders Søgaard. 2021b. Is the lottery fair? evaluating winning tickets across demographics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3214–3224.

Trevor J. Hastie and Robert J. Tibshirani. 2017. *Generalized additive models*. Routledge.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML 2019*, pages 2790–2799.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of ACL-IJCNLP 2015*, pages 483–488.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multitask Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of LREC 2020*, pages 1440–1448.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL 2015*, pages 103–112.

Anna Jørgensen and Anders Søgaard. 2021. Evaluation of summarization systems across gender, age, and race. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 51–56.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of ACL 2020*.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. QuestionBank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 497–504.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

David Jurgens, Varada Kolhatkar, Lucy Li, Margot Mieskes, and Ted Pedersen, editors. 2021. *Proceedings of the Fifth Workshop on Teaching NLP*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of EACL 2021*, pages 3250–3258.

Sanjeev P Khudanpur. 2006. Multilingual language modeling. *Multilingual Speech Processing*, page 169.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of ICML 2021*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of ACL 2019*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of EMNLP 2021*.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Łukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In *Proceedings of ICLR 2018*.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *CoRR*, abs/2106.06052.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop*.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of WWW 2013*, page 897–908.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NeurIPS 2013*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of LREC 2020*, pages 4034–4043.

Debora Nozza. 2021. Exposing the Limits of Zero-shot Cross-lingual Hate Speech Detection. In *Proceedings of ACL 2021*, pages 907–914.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of ACL 2011*, pages 309–319.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of EMNLP 2020*, pages 7654–7673.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of EMNLP 2021*.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of NAACL-HLT 2018*, pages 293–302.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of ACL-IJCNLP 2021*, pages 99–109.

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

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of ACL-IJCNLP 2021*, pages 3118–3135.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of ACL 2019*, pages 2931–2951.

Yikang Shen, Zhouhan Lin, Chin-wei Huang, and Aaron Courville. 2018. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In *Proceedings of ICLR 2018*.

Tracy Sherman. 1985. Categorization skills in infants. *Child Development*, 56(6):1561–73.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP 2013*, pages 1631–1642.

Anders Søgaard. 2013. *Semi-supervised learning and domain adaptation for NLP*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. *CoRR*, abs/2009.06732.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of ACL 2020*, pages 7396–7408.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the Primacy of English in Zero-shot Cross-lingual Transfer. *arXiv preprint arXiv:2106.16171*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of ACL 2017*, pages 2016–2027.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the EACL 2021*, pages 2203–2213.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of NeurIPS 2020*.

David Vilares, Haiyun Peng, Ranjan Satapathy, and Erik Cambria. 2018. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1292–1298.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP 2013*, pages 1815–1827.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of EMNLP 2020*, pages 2649–2656.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples. In *Proceedings of ICML 2018*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 11–20.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of ACL-IJCNLP 2021*, pages 7361–7373.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL 1995*, pages 189–196.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of EMNLP 2021*, pages 2115–2129.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Proceedings of NeurIPS 2020*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of ACL 2020*, pages 2896–2907.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of ACL-IJNCLP 2021*, pages 5751–5767.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Appendix

A.1 Analysis of remaining areas at ACL 2021

We provide statistics for the remaining areas at ACL 2021 in Table 4.

A.2 Analysis of Efficiency area at EMNLP 2021

We annotated the 20 papers presented orally at EMNLP 2021 in the "Efficient Models in NLP" area. Among the presented papers, 19/20 are monolingual and 17 focus only on English. Among the other two, one focuses on Indonesian and one on Chinese. The last paper focuses on MT with multiple languages. Papers mainly evaluate using accuracy and/or F1 and many papers evaluate on GLUE. There is a single two-dimensional paper according to our criteria (the paper focusing on MT, which makes contributions on multilinguality and efficiency) while two other papers can be considered two-dimensional but cover dimensions that we do not annotate, i.e. privacy and robustness respectively. This analysis corroborates our findings that research papers depart from SQUARE ONE in such dedicated conference areas/tracks, but largely only across a single dimension.

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
Question Answering	24	95.8%	41.7%	4.2%	4.2%	8.3%	4.2%	0.0%
Sentence-level Semantics	23	87.0%	56.5%	8.7%	0.0%	4.3%	17.4%	4.3%
Computational Social Science	18	77.8%	66.7%	0.0%	22.2%	0.0%	16.7%	0.0%
Language Generation	18	83.3%	0.0%	11.1%	5.6%	11.1%	11.1%	5.6%
Sentiment Analysis	18	100.0%	72.2%	0.0%	0.0%	11.1%	11.1%	0.0%
Summarization	12	91.7%	0.0%	0.0%	8.3%	0.0%	8.3%	0.0%
Semantics: Lexical Semantics	12	58.3%	41.7%	25.0%	0.0%	16.7%	0.0%	8.3%
Information Retrieval	12	91.7%	8.3%	0.0%	0.0%	0.0%	0.0%	8.3%
Language Grounding to Vision	11	100.0%	18.2%	0.0%	0.0%	9.1%	27.3%	0.0%
Syntax	10	40.0%	20.0%	30.0%	0.0%	20.0%	10.0%	20.0%
Best Paper Session	8	50.0%	50.0%	12.5%	0.0%	25.0%	25.0%	12.5%
Speech and Multimodality	6	66.7%	33.3%	16.7%	0.0%	0.0%	0.0%	0.0%
Phonology and Morphology	6	33.3%	33.3%	33.3%	0.0%	0.0%	16.7%	16.7%
Linguistic Theories	6	100.0%	16.7%	0.0%	0.0%	16.7%	33.3%	0.0%
Theme	5	20.0%	40.0%	20.0%	20.0%	20.0%	20.0%	20.0%

Table 4: The number of papers in the remaining areas as well as the fractions that only evaluate on English, only use accuracy / F1, make contributions along one of four dimensions, and make contributions along more than a single dimension (from left to right).