Circles are like Ellipses, or Ellipses are like Circles? Measuring the Degree of Asymmetry of Static and Contextual Word Embeddings and the Implications to Representation Learning

Wei Zhang [1], **Murray Campbell** [1], **Yang Yu*** [2], **Sadhana Kumaravel** [1]
[1] IBM Research AI
[2] Google
zhangwei@us.ibm.com, mcam@us.ibm.com, yangyuai@google.com, sadhana.kumaravel1@ibm.com

Abstract

Human judgments of word similarity have been a popular method of evaluating the quality of word embedding. But it fails to measure the geometry properties such as asymmetry. For example, it is more natural to say "Ellipses are like Circles" than "Circles are like Ellipses". Such asymmetry has been observed from the word evocation experiment, where one word is used to recall another. This association data have been understudied for measuring embedding quality. In this paper, we use three well-known evocation datasets for the purpose and study both static embedding as well as contextual embedding, such as BERT. To fight for the dynamic nature of BERT embedding, we probe BERT's conditional probabilities as a language model, using a large number of Wikipedia contexts to derive a theoretically justifiable Bayesian asymmetry score. The result shows that the asymmetry judgment and similarity judgments disagree, and asymmetry judgment aligns with its strong performance on "extrinsic evaluations". This is the first time we can show contextual embeddings's strength on intrinsic evaluation, and the asymmetry judgment provides a new perspective to evaluate contextual embedding and new insights for representation learning.

Introduction

Popular static word representations such as word2vec (Mikolov et al. 2013) lie in Euclidean space and are evaluated against symmetric judgments. Such a measure does not expose the geometry of word relations, e.g., asymmetry. For example, "ellipses are like circles" is much more natural than "circles are like ellipses". An acceptable representation may exhibit such a property.

Tversky (1977) proposed a similarity measure that encodes asymmetry. It assumes each word is a feature set, and asymmetry manifests when the common features of two words take different proportions in their respective feature sets, i.e., a difference of the likelihoods $P(a|b)$ and $P(b|a)$ for a word pair (a,b). In this regard, the degree of correlation between asymmetry from humans and a word embedding may indicate the feature-encoding quality of the embedding.

Word evocation experiment devised by neurologist Sigmund Freud around the 1910s was to obtain such word directional relationship, where a word called cue is shown to a participant who is asked to "evoke" another word called target freely. [1] The experiment is usually conducted on many participants for many cue words. The data produced from the group of people exhibit a collective nature of word relatedness. The $P(a|b)$ and $P(b|a)$ can be obtained from such data to obtain an asymmetry ratio (Griffiths, Steyvers, and Tenenbaum 2007) that resonates with the theory of Tversky (1977) and Resnik (1995). Large scale evocation datasets had been created to study the psychological aspects of language. We are interested in three of them; the Edinburgh Association Thesaurus (Kiss et al. 1973), Florida Association Norms (Nelson, McEvoy, and Schreiber 2004) and Small World of Words (De Deyne et al. 2019) Those three datasets have thousands of cue words each and all publicly available. We use them to derive the human asymmetry judgments and see how well embedding-derived asymmetry measure aligns with this data.

Evocation data was rarely explored in the Computational Linguistics community, except that Griffiths, Steyvers, and Tenenbaum (2007) derived from the Florida Association Norms an asymmetry ratio for a pair of words to measure the directionality of word relations in topic models, and Nematzadeh, Meylan, and Griffiths (2017) used it for word embedding. In this paper, we conduct a larger scale study using three datasets, on both static embedding (word2vec) (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), fasttext (Mikolov et al. 2018)) and contextual embedding such as BERT (Devlin et al. 2018). We hope the study could help us better understand the geometry of word representations and inspire us to improve text representation learning.

To obtain $P(a|b)$ for static embedding, we leverage vector space geometry with projection and soft-max similar to (Nematzadeh, Meylan, and Griffiths 2017; Levy and Goldberg 2014; Arora et al. 2016); For contextual embedding such as BERT we can not use this method because the embedding varies by context. Thus, we use a Bayesian method to estimate word conditional distribution from thousands of contexts using BERT as a language model. In so doing, we can probe the word relatedness in the dynamic embedding space in a principled way.

*work done while with IBM

[1] A clip from A Dangerous Method describing Sigmund Freud's experiment https://www.youtube.com/watch?v=lblzHkoNn3Q

Comparing an asymmetry measure to the popular cosine measure, we observe that similarity judgment fails to correctly measure BERT's lexical semantic space, while asymmetry judgment shows an intuitive correlation with human data. In the final part of this paper, we briefly discuss the result and what it means to representation learning. *This paper makes the following contributions:*

1. An analysis of embedding asymmetry with evocation datasets and an asymmetry dataset to facilitate research;
2. An unbiased Bayesian estimation of word pair relatedness for contextual embedding, and a justifiable comparison of static and contextual embeddings on lexical semantics using asymmetry.

Related Work

Word embedding can be evaluated by either the symmetric "intrinsic" evaluation such as word pair similarity/relatedness (Agirre et al. 2009; Hill, Reichart, and Korhonen 2015) and analogy (Mikolov et al. 2013) or the "extrinsic" one of observing the performance on tasks such as text classification (Joulin et al. 2016), machine reading comprehension (Rajpurkar et al. 2016) or language understanding benchmarks (Wang et al. 2018, 2019).

On utterance-level, probing contextual embeddings was conducted mainly on BERT (Devlin et al. 2018), suggesting its strength in encoding syntactic information rather than semantics (Hewitt and Manning 2019; Reif et al. 2019; Tenney et al. 2019; Tenney, Das, and Pavlick 2019; Mickus et al. 2019), which is counter-intuitive given contextual representation's superior performance on external evaluation.

On the lexical level, it is yet unknown if previous observation also holds. Moreover, lexical-semantic evaluation is non-trivial for contextual embedding due to its dynamic nature. Nevertheless, some recent approach still tries to either extract a static embedding from BERT using PCA (Ethayarajh 2019; Coenen et al. 2019), or use context embeddings as is (Mickus et al. 2019) for lexical-semantic evaluation. Some instead use sentence templates (Petroni et al. 2019; Bouraoui, Camacho-Collados, and Schockaert 2020) or directly analyze contextual embedding for its different types of information other than lexical ones (Brunner et al. 2019; Clark et al. 2019; Coenen et al. 2019; Jawahar, Sagot, and Seddah 2019). So far, a theoretically justifiable method still is missing to mitigate the bias introduced in the assumption of above analysis methods. Thus, how contextual and static embedding compares on lexical semantics is still open.

An asymmetry ratio was used to study the characteristics of static embedding (Nematzadeh, Meylan, and Griffiths 2017) and topic models (Griffiths, Steyvers, and Tenenbaum 2007), and the study of asymmetry of contextual embedding is still lacking. Given the current status of research, it is natural to ask the following *research questions*:

- **RQ 1.** Which evocation dataset is the best candidate to obtain asymmetry ground-truth? And **RQ 1.1**: Does the asymmetry score from data align with intuition? **RQ 1.2**: Are evocation datasets correlated for the first place?
- **RQ 2.** How do static and context embeddings compare on asymmetry judgment? Furthermore, how do different context embeddings compare? Do larger models perform better?
- **RQ 3.** What are the main factors to estimating $P(b|a)$ with context embedding, and what are their effects?
- **RQ 4.** Does asymmetry judgment and similarity judgment agree on embedding quality? What does it imply?

We first establish the asymmetry measurement in Section and methods to estimate them from an evocation data or an embedding in Section followed by empirical results to answer those questions in Section .

The Asymmetry Measure

Log Asymmetry Ratio (LAR) of a word pair. To measure asymmetry of a word pair $(a; b)$ in some set of word pairs $\mathcal{S}$, we define two conditional likelihoods, $P_{\mathcal{E}}(b|a)$ and $P_{\mathcal{E}}(a|b)$, which can be obtained from a resource $\mathcal{E}$, either an evocation dataset or embedding. Under Tversky's (1977) assumption, if b relates to more concepts than a does, $P(b|a)$ will be greater than $P(a|b)$ resulting in asymmetry. A ratio $P_{\mathcal{E}}(b|a)/P_{\mathcal{E}}(a|b)$ (Griffiths, Steyvers, and Tenenbaum 2007) can be used to quantify such asymmetry. We further take the logarithm to obtain a *log asymmetry ratio* (LAR) so that the degree of asymmetry can naturally align with the sign of LAR (a ratio close to 0 suggests symmetry, negative/positive otherwise). Formally, the LAR of $(a; b)$ from resource $\mathcal{E}$ is

$$\text{LAR}_{\mathcal{E}}(a; b) = \log P_{\mathcal{E}}(b|a) - \log P_{\mathcal{E}}(a|b) \quad (1)$$

LAR is key to all the following metrics.

Aggregated LAR (ALAR) of a pair set. We care about the aggregated LAR on a word pair set $\mathcal{S}$, the expectation $\text{ALAR}_{\mathcal{E}}(\mathcal{S}) = \mathbf{E}_{(a;b)\in\mathcal{S}}[\text{LAR}_{\mathcal{E}}(a; b)]$, to quantify the overall asymmetry on $\mathcal{S}$ for $\mathcal{E}$. However, it is not sensible to evaluate ALAR on any set of $\mathcal{S}$, for two reasons: 1) because LAR(a;b) $= -$LAR$(b; a)$, randomly ordered word pairs will produce random ALAR values; 2) pairs of different relation types may have very different LAR signs to cancel each other out if aggregated. For example, "a is a part of b" suggest LAR$(a; b) > 0$, and "a has a b" for LAR$(a; b) < 0$. Thus, we evaluate ALAR on $\mathcal{S}(r)$, the relation-specific subset, as

$$\text{ALAR}_{\mathcal{E}}(\mathcal{S}(r)) = \mathbf{E}_{(a,b)\in\mathcal{S}(r)}[\text{LAR}_{\mathcal{E}}(a; b)] \quad (2)$$

where the order of (a, b) is determined by $(a, r, b) \in$ KG, the ConceptNet (Speer, Chin, and Havasi 2017). Note that when $\mathcal{E}$ is an evocation data, we can qualitatively examine if $\text{ALAR}_{\mathcal{E}}(\mathcal{S}(r))$ aligns with with human intuition as one (**RQ 1.2**) of the two metrics to determine the candidacy of an evocation data as ground truth; When $\mathcal{E}$ is any embedding, it measures the asymmetry of embedding space.

Correlation on Asymmetry (CAM) of Resources. By using LAR defined in Eq. 1, for a resource $\mathcal{E}$ and a word pair set $\mathcal{S}$ we can define a word-pair-to-LAR map

$$\mathcal{M}(\mathcal{E}, \mathcal{S}) = \{(a; b) : \text{LAR}_{\mathcal{E}}(a; b)|(a; b) \in \mathcal{S}\} \quad (3)$$

and the Spearman Rank Correlation on asymmetry measure (CAM) between two resources $\mathcal{E}_i$ and $\mathcal{E}_j$ can be defined as

$$\text{CAM}(\mathcal{S}, \mathcal{E}_i, \mathcal{E}_j) = \text{Spearman}(\mathcal{M}(\mathcal{E}_i, \mathcal{S}), \mathcal{M}(\mathcal{E}_j, \mathcal{S})) \quad (4)$$

There are two important settings:

1. $\mathcal{E}_i$ is an evocation data and $\mathcal{E}_j$ is an embedding
2. $\mathcal{E}_i$ and $\mathcal{E}_j$ are two different evocation datasets

Setting one is to evaluate embedding using evocation data as the ground truth. Setting two is to measure the correlation between any pair of evocation data, which is the second metric (**RQ 1.1**) to validate the candidacy of an evocation dataset as asymmetry ground-truth: several studies indicate that the human scores consistently have very high correlations with each other (Miller and Charles 1991; Resnik 1995). Thus it is reasonable to hypothesize that useful evocation data should correlate stronger in general with other evocation data. We will discuss it more in experiments.

Next, we introduce how to obtain $P(b|a)$ for calculating LAR, ALAR, and CAM from different resources.

Method

$P_D(b|a)$ from Evocation Data

For evocation data D, $P_D(b|a)$ means when a is cue, how likely b can be evoked (Kiss et al. 1973). During the experiment, one has to go through a thought process to come up with b. This process is different from humans writing or speaking with natural languages, because the word associations are free from the basic demands of communication in natural language, making it an ideal tool to study internal representations of word meaning and language in general (De Deyne et al. 2019). Unlike similarity judgment where a rating (usually 1 to 10) is used to indicate the degree of similarity of (a, b), evocation data does not immediately produce such a rating, but a binary "judgment" indicating if b is a response of a or not. Yet, evocation data is collected from a group of participants (De Deyne et al. 2019; Nelson, McEvoy, and Schreiber 2004), and we could derive a count-based indicator to average the binary judgments to give a rating as is done for the similarity judgment averaging the scores of experts. The total number of responses usually normalizes such count, leading to a score called *Forward Association Strength* (FSG) (Nelson, McEvoy, and Schreiber 2004), a metric invented for psychology study. FSG score is essentially the $P_D(b|a)$,

$$P_D(b|a) = \frac{\text{Count}(b \text{ as a response}|a \text{ is cue})}{\text{Count}(a \text{ is cue})} \tag{5}$$

It is easy to confuse such evocation counts with the counts derived from texts. Again, the counts in evocation data are the aggregation of judgments (De Deyne et al. 2019) rather than co-occurrence from language usage which is subject to the demands of communication.

$P_B(b|a)$ from Contextual Embedding

It is easy to obtain $P(b|a)$ for static embedding by exploring geometric properties such as vector projection. But estimating it with contextual embedding is generally hard due to the embedding's dynamic nature invalidating the projection approach. Thus, to evaluate $P(b|a)$ within a contextual embedding space in an unbiased manner yet admitting its dynamic nature, we first find the contexts that a and b co-occur, and then use one word to predict the likelihood of another, admitting the existence of the context. Finally, to remove the bias introduced by context, we average the likelihood over many contexts to obtain an un-biased estimate of $P(b|a)$. This idea can be understood in a Bayesian perspective: we introduce a random variable $\mathbf{c}$ to denote a paragraph as context from corpus C, say, Wikipedia. Then we obtain the expectation of $P_B(b|a)$ over $\mathbf{c}$, using B, say BERT, as a language model as $P_B(b|a) = \mathbf{E}_{P(\mathbf{c}|a),\mathbf{c}\in C}[P_B(b|a,\mathbf{c})]$. The formation can be simplified: for $\mathbf{c}$ that does not contain a, $P(\mathbf{c}|a) = 0$, and for $\mathbf{c}$ that does not contain b, $P_B(b|a,\mathbf{c}) = 0$. Finally,

$$P_B(b|a) = \mathbf{E}_{P(\mathbf{c}|a),\mathbf{c}\in C(\{a,b\})}[P_B(b|\mathbf{c})] \tag{6}$$

where $C(x)$ indicates all the contexts in C that includes x, being either a single word or a word pair. Note that $P_B(b|a,\mathbf{c}) = P_B(b|\mathbf{c})$ if $a \in \mathbf{c}$, leading to Eq. 6. $P(\mathbf{c}|a)$ is estimated as $1/|C(a)|$ and $P_B(b|\mathbf{c})$ is estimated by masking b from the whole paragraph $\mathbf{c}$ and then getting the probability of b from the Soft-max output of a pre-trained BERT-like Masked language model B. When there are N words of b in $\mathbf{c}$, we only mask the one that is being predicted and repeat the prediction for each b. We append "[CLS]" to the beginning of a paragraph and "[SEP]"after each sentence. Word a can also appear $k > 1$ times in $\mathbf{c}$ and we regard $\mathbf{c}$ as k contexts for $C(a)$.

Using BERT as a language model, we can make an unbiased estimation of context-free word relatedness, if the contexts are sufficient and the distribution is not biased. But, like all unbiased estimators, Eq. 6 may suffer from high variance due to the complexity of context $\mathbf{c}$. We identify two factors of context that may relate to estimation quality (**RQ 3**): the number of contexts and the distance between words as a bias of context, which we discuss in the experiments.

$P_E(b|a)$ from Static Embedding

To answer **RQ 4**, we also calculate conditionals for static embedding. For word2vec or GloVe, each word can be regarded as a bag of features (Tversky 1977) and $P(b|a)$ can be obtained using a "normalized intersection" of the feature sets for a and b which corresponds to geometric projection in continuous embedding vector space:

$$proj(b|a) = \frac{emb(b) \cdot emb(a)}{||emb(a)||} \tag{7}$$

And we normalize them with Soft-max function to obtain P_E as

$$P_E(b|a) = \frac{exp(proj(b|a))}{\sum_x exp(proj(x|a)} \tag{8}$$

where the range of x is the range of the evocation dataset. If we compare Eq. 8 and 7 to the dot-product (the numerator in Eq. 7) that is used for similarity measurement (Levy and Goldberg 2014; Nematzadeh, Meylan, and Griffiths 2017; Arora et al. 2016), we can see dot-product only evaluates how much overlap two embedding vectors have in common regardless of its proportion in the entire meaning representation of a or b. In other words, it says "ellipses" are similar to "circles". But it fails to capture if there is more to "circles" that is different than "ellipses".

	EAT		FA		SWOW		Spearman's Correlation (**CAM**)		
r	count	**ALAR**	count	**ALAR**	count	**ALAR**	EAT-FA	SW-FA	SW-EAT
relatedTo	8296	4.50	5067	0.89	34061	4.83	0.59	0.68	0.64
antonym	1755	1.27	1516	0.38	3075	0.01	0.43	0.58	0.51
synonym	673	-15.80	385	-17.93	2590	-15.85	0.49	0.65	0.59
isA	379	43.56	342	31.59	1213	47.77	0.64	0.75	0.59
atLocation	455	17.48	356	9.59	1348	16.02	0.61	0.71	0.64
distinctFrom	297	-2.38	250	0.01	593	-1.07	0.32	0.57	0.43

Table 1: Pair count, ALAR($S(r)$) and the Spearman Correlation on Asymmetry Measure (CAM) between datasets. See Appendix for a complete list. P-value< 0.00001 for all CAM results

An Asymmetry-Inspired Embedding and Evaluation. Unlike static embedding that embeds all words into a single space, a word can have two embeddings. In word2vec learning (Mikolov et al. 2013), it corresponds to the word and context embedding learned jointly by the dot-product objective. Intuitively, such a dot-product could encode word relatedness of different relation types more easily than with a single embedding space. To verify this, we use the weight matrix in word2vec skip-gram as context embedding similar to (Torabi Asr, Zinkov, and Jones 2018) together with the word embedding to calculate $P_E(b|a)$. We denote it as **cxt**. With **cxt**, the asymmetry can be explicitly encoded: to obtain $P(b|a)$, we use word embedding for a and context embedding for b, and then apply Eq. 7 and 8.

Result and Analysis

To sum up, we first obtain $P(b|a)$ (Section) from evocation data, static and context embedding respectively, and then use them to calculate asymmetry measure LAR, ALAR and CAM (Section). Below we start to answer the plethora of research questions with empirical results.

RQ 1: Which evocation data is better to obtain asymmetry ground truth?

We answer it by examining two sub-questions: an evocation data's correlation to human intuition (**RQ 1.1**) and its correlation with other evocation data (**RQ 1.2**).

RQ 1.1: Correlation to Human Intuition To see if an evocation data conforms to intuition, we examine the ALAR for each relation type r separately, which requires grouping word pairs $\mathcal{S}$ to obtain $\mathcal{S}(r)$.

Unfortunately, evocation data does not come with relation annotations. Thus we use ConceptNet (Speer, Chin, and Havasi 2017) to automatically annotate word relations. Specifically, we obtain $S(r) = \{(a, b)\}$ where (a, b) is connected by r (we treat all relations directional) where a as head and b is tail. If (a, b) has multiple relations $\{r_i\}$, we add the pair to each $S(r_i)$. Pairs not found are annotated with the ConceptNet's *relatedTo* relation. Finally we calculate the $\text{ALAR}_{\mathcal{E}}(\mathcal{S}(r))$ using Eq. 2 for each relation r.

Table 1 shows a short list of relation types and their ALAR($\mathcal{S}(r)$). We can see that *isA*, *partOf*, *atLocation*, *hasContext* exhibit polarized ALAR; whereas *antonym* and *distinctFrom* are comparably neutral. These observations agree with intuition in general, except for *synonym* and *similarTo*, (e.g., "ellipses" and "circles") which ALARs show mild asymmetry. This may have exposed the uncertainty and bias of humans on the KG annotation of similarity judgments, especially when multiple relations can be used to label a pair, e.g. "circle" and "ellipse" can be annotated with either "isA" or "similarTo". However, such bias does not affect the correctness of asymmetry evaluation because the Spearman correlation of two resources is correctly defined no matter which set of pairs is used. The relation-specific ALAR is for qualitatively understanding the data. However, this interesting phenomenon may worth future studies.

RQ 1.2: Correlation to each other The observation that good human data have high correlations with each other (Miller and Charles 1991; Resnik 1995) provides us a principle to understand the quality of the three evocation datasets by examining how they correlate. Our tool is CAM of Eq. 4 defined on the common set of word pairs, the intersection $\mathcal{S}(r) = \mathcal{S}_{\text{EAT}}(r) \cap \mathcal{S}_{\text{FA}}(r) \cap \mathcal{S}_{\text{SWOW}}(r)$ where each set on RHS is the collected pairs for r in a dataset. The number of common pairs for r is about 90% of the smallest dataset for each r in general. Then, the CAM is obtained by Eq. 3 and 4, e.g. $\text{CAM}_{\mathcal{S}(r)}$(SWOW, FA). The calculated CAM is shown in Table 1, and a longer list is in Table **??** in Appendix. In general, SWOW and FA show stronger ties, probably because they are more recent and closer in time; EAT correlates less due to language drift.

Answering RQ 1: Which data to use? From Table 1, we favor SWOW because 1) in the study of **RQ 1.1** we see SWOW aligns with human intuition as well as, if not better than, the other two datasets, e.g., it made almost symmetric ALAR estimation on pair-abundant relations such as *antonym*; 2) According to the answer to **RQ 1.2**, in general, SWOW correlates to all other datasets the best, e.g., on the most pair-abundant relation *relatedTo*, SWOW has the top two CAM scores, 0.68 and 0.64 to other datasets; 3) it is the largest and the most recent dataset. Thus we mainly use SWOW for later discussions.

RQ 2: Asymmetry of Embedding

Setup. We compare embeddings using CAM in Eq. 4 and set $\mathcal{E}_i$ to an embedding E and $\mathcal{E}_j$ to evocation data D using LAR obtained according to Section . For context embeddings, we leverage the masked language models obtained from Huggingface Toolkit (Wolf et al. 2019) (See Appendix for a full list of models), and for static embeddings we both

	EAT (12K pairs)				**FA** (8K pairs)				**SWOW** (30K pairs)			
	w2v/cxt	glv	fxt	bert/bertl	w2v/cxt	glv	fxt	bert/bertl	w2v/cxt	glv	fxt	bert/bertl
relatedTo	.08/.25	.37	.20	.48/**.55**	.17/.30	.41	.27	.44/**.50**	.06/.28	.42	.14	.43/**.50**
antonym	.05/.15	.25	.07	.31/**.38**	.16/.23	.31	.21	.30/**.38**	.04/.15	.33	.09	.33/**.41**
synonym	-.21/.14	.43	-.03	.52/**.59**	.19/.37	**.44**	.40	.33/.41	.00/.29	.43	.16	.38/**.47**
isA	.06/.33	.45	.27	.50/**.58**	.23/.41	.44	.37	.43/**.50**	.03/.34	.39	.16	.50/**.57**
atLocation	.08/.28	.44	.29	.45/**.52**	.22/.36	**.47**	.31	.33/.44	.08/.35	.47	.21	.44/**.52**
distinctFrom	-.12/-.20	.05	-.09	.17/**.28**	.11/.17	.35	.17	.34/**.42**	.03/.19	.40	.16	.38/**.49**
SA	.05/.22	.34	.16	.45/**.52**	.17/.29	.39	.26	.38/**.46**	.05/.27	.41	.15	.42/**.49**
SR	-.02/.15	.30	.08	.40/**.47**	.17/.27	.36	.25	.35/**.43**	.02/.24	.38	.16	.40/**.48**

Table 2: Spearman Correlation on Asymmetry Measure (CAM) between embedding LAR and data LAR. Acronyms: w2v (word2vec), glv (GloVe), fxt (fasttext), bert (`BERT-base`), bertl (`BERT-large`). SA (Weight-Averaged Spearman, where weights are calculated as $|S(r)|/|S|$); SR (SA excluding *relateTo* relation). P-value$<$0.0001 for BERT and GloVe in general. Using Eq. 9, where V is the intersection of above embeddings, we collect 12K pairs for EAT data, 8K for FA, and 30K for SWOW.

obtain pre-trained embeddings (for GloVe and fasttext) and train embeddings ourselves (w2v and cxt) using Wikipedia corpus (October 2019 dump, details are in Appendix). An issue that hampers fair comparison of embeddings is their vocabularies differ. To create a common set of word pairs, we take the intersection V of the vocabularies of all embeddings that are to be compared and the evocation dataset, and for Eq. 4 we obtain $\mathcal{S}(r)$ as

$$\{(a,b)|(a,b) \in D \wedge r(a,b) \in \text{KG} \wedge a \in V \wedge b \in V\} \quad (9)$$

where KG is ConceptNet. which means any word in the set of pairs has to be in-vocabulary for any embedding.

We have two settings: 1) comparing static and contextual embeddings (BERT as a representative), wherein applying Eq. 9 leads to 12K, 8K, and 30K pairs on EAT, FA, and SWOW dataset in Table 2. 2) comparing contextual embeddings, which leads to 7.3K SWOW pairs with asymmetry scores that will be made public.

Comparing Static and Contextual embedding In Table 2, we compare BERT (base and large) with static embeddings on three different evocation datasets with **CAM** (Relation-specific ALAR can provide us a qualitative understanding how embedding performs on each relation in general but it does not affect the correctness of CAM). GLoVe is the most competitive among static embeddings because it takes into account the ratio $P(x|a_1)/P(x|a_2)$ that may help learn $P(b|a)$, which can lead to better accuracy on some relations. BERT, especially `BERT-large`, has a stronger correlation with the three datasets than any other static embedding, which aligns with the empirical evidence from external evaluation benchmarks such as SQUAD (Rajpurkar et al. 2016) and GLUE (Wang et al. 2018). It may be the first time we can show context embedding outperforms static embedding on intrinsic evaluation. Moreover, by comparing **CAM** on a per-relation basis, we see BERT performs competitively on LAR-polarized, asymmetric relations such as "relatedTo", "isA" and "atLocation", while not so much on symmetric ones. Also, the context embedding (cxt) consistently outperforms word2vec on almost all relation types. Combining these observations, we think that the dot-product of two embedding spaces can encode rich information than a single embedding space can. BERT does it with a key-query-value self-attention mechanism, being one reason for it to perform well on the asymmetry judgment. Also, it is not surprising that `BERT-large` outperforms `BERT-base`, suggesting larger models can indeed help better "memorize" word semantics, which we also show for other models soon later.

What about LAR Directions? An embedding could have a high correlation (CAM) with data but totally wrong on asymmetry directions (LAR). Thus in Fig. 1, we compare embeddings' ALAR to the ALAR of data (SWOW). We took log over the ALAR(r) while retaining the sign to smooth out the numbers. `BERT-base` produces small ALAR values, which we scale by x1000 before log to make the figure easy to read. `BERT-base` and GLoVe are two strong embeddings that show better directional correlation with SWOW. Note that word pairs under *hasContext* aligns with SWOW data generally well, but words with relations *hasProperty*, *capableOf*, *hasA* is hard for all text embeddings. These findings may suggest a non-negligible gap (Spearman's Correlation for BERT-SWOW and GloVe-SWOW has P-value<0.0001) between text-based word embeddings and the human-generated evocation data, regardless of how embeddings are trained. It may be either because texts do not entail relations of those pairs or because relations are too hard for current embedding techniques to discover, which requires further investigation.

Comparing Contextual Embeddings. By applying Eq. 9, we use all candidate contextual embeddings' vocabulary to obtain V and use SWOW as D, resulting in 7.3K pairs in total, for which in Table 3 we show the comparison of embeddings on CAM. In general, we see that larger models indeed show a higher correlation with SWOW data, which suggests models with larger capacity can help encode lexical semantics better in general. For example, CAM of BERT (Devlin et al. 2018), roBERTa (Liu et al. 2019) and ALBERT (Lan et al. 2019) grow with the number of parameters, yet ELECTRA (Clark et al. 2020) does not show the same trend. One reason for the abnormality may be that ELECTRA uses generated synthetic text for training, which may

	# Params (M)	relatedTo	antonym	synonym	isA	atLocation	distinctFrom
BERT							
-base	110	.33	.23	.32	.36	.44	.23
-large	340	.41	.26	.38	.45	.49	.28
roBERTa							
-base	125	.41	.27	.38	.45	.48	.27
-large	355	.42	.27	.38	.46	.49	.29
ALBERT							
-base	11	.36	.24	.37	.42	.41	.25
-large	17	.38	.25	.38	.44	.42	.24
-xl	58	.39	.26	.37	.45	.43	.24
-xxl	223	.39	.25	.37	.43	.44	.26
ELECTRA							
-base	110	.39	.24	.38	.45	.44	.24
-large	335	.36	.26	.35	.34	.42	.29

Table 3: Spearman Correlation on Asymmetry Measure (CAM) between SWOW and contextual embedding for 7.3K SWOW pairs obtained with Eq. 9. Counts: 5409, 910, 217, 211, 262, 206 from left to right.

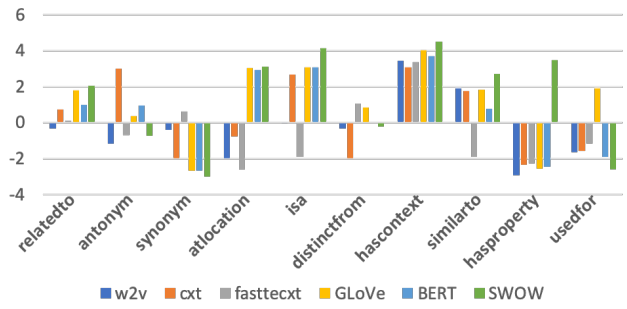

Figure 1: LAR Prediction Comparison. Same setting as Table 2. BERT refers to BERT-base

result in data drift that is exacerbated by using larger models. While ELECTRA and ALBERT have shown their advantage over BERT in many external evaluation benchmarks such as GLUE, SUPERGLUE, and SQuAD (Rajpurkar et al. 2016; Wang et al. 2018, 2019), they do not improve significantly over asymmetry judgment compared to BERT or roBERTa. It is reasonable to doubt if the improvements of ELECTRA or ALBERT over BERT come from better model tuning or better semantic representation, and asymmetry judgment may shed some light on the answer to this question. Also, RoBERTa outperform BERT may be because RoBERTa improves BERT's optimization, and in turn it confirms that optimization matter on semantic encoding. Thus, exploring better optimization techniques may still be valuable.

RQ 3: Two Factors of the Bayesian Estimation

Since we evaluate $P(b|a)$ with contexts, the quality and distribution of it matter the most for the LAR, ALAR, and CAM. Comparing BERT-base and SWOW as an example, we study two quantities: 1) The number of contexts of a word pair because more contexts suggest the more accurate the estimation may be; and 2) the distance of word pairs in context, because the closer two words are, the more likely they relate. For 1), we group word pairs into bins of size 200 by the number of contexts collected for each pair and use average LAR directional accuracy (Appendix **??**) in each bin as a tool to study the factors' impact. The three upper figures in Figure 2 suggest a mild trend where pairs with more contexts have higher direction accuracy, which discontinues beyond 5000 contexts. We hypothesis that the discontinuation is due to the pairs grouped into >5000 bin may contain "systematical bias", such as topics, where a word depends more on the topic than the other word, which pushes asymmetry prediction towards random. Techniques of diversifying the context may help alleviate the problem, an extensive study too complicated for this paper. For 2), we group word pairs by character distance into size-200 bins. The three bottom ones in Figure 2 show word distance correlates weakly to direction accuracy due to BERT's ability to model long-distance word dependency.

RQ 4: Similarity v.s. Asymmetry Judgment

In comparison to asymmetry judgment, we would like to see if similarity judgment can say otherwise about embeddings. We compare all embeddings on popular symmetric similarity/relatedness datasets. We take the dot-product score of two word embeddings for static embeddings and calculate the Spearman's correlation on the scores. For contextual embeddings, we use the geometric mean of $P(a|b)$ and $P(b|a)$ as similarity score (see Appendix for justification) to be comparable to dot-product. Table 4 shows that although this approach helps BERT performs better on 2 out of 3 datasets than PCA on the contextual embedding approach (Ethayarajh 2019), the result on other contextual embeddings looks extremely arbitrary compared to static embeddings. Similarity judgment, in general, fails to uncover contextual embedding's ability on lexical semantics: it focuses on similarity rather than the difference that BERT seems to be good at, which can also be supported by contextual embeddings being superior on WS353 REL than SIM. Similarity judgment tells us that contextual embedding does not correctly encode semantic features and static embeddings, but it can beat them reasonably well on asymmetry judgment, suggesting otherwise. Are they conflicting with each other? Let us look into it now.

Discussion and Conclusion

The rise of Transformers has aroused much speculation on how the model works on a wide variety of NLP tasks. One lesson we learned is that learning from large corpora how to match contexts is very important, and many tasks require the ability. From the intrinsic evaluation perspective, the asymmetry judgment and similarity judgment also support this. BERT can encode rich features that help relatedness modeling, but it fails frustratingly on similarity judgments that suggest otherwise.

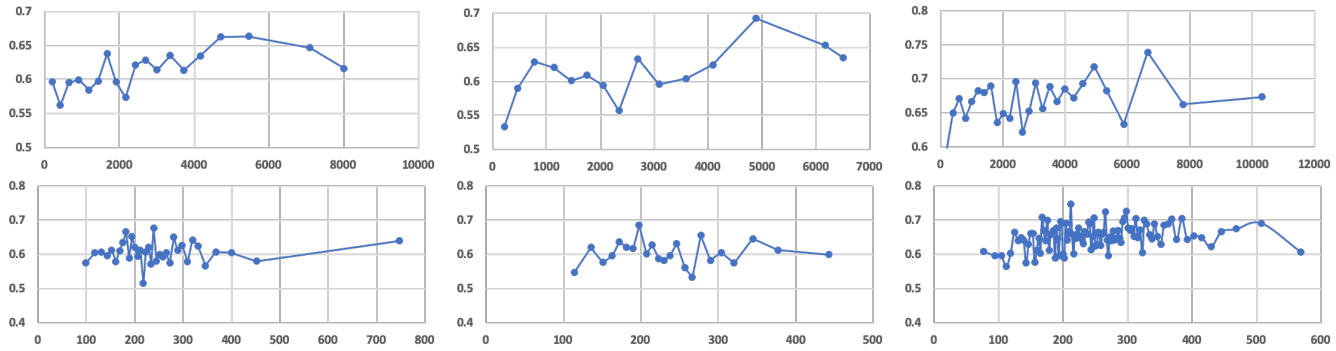

Figure 2: BERT directional accuracy (y-axis) v.s. context frequency (x-axis for upper 3 figures) and average character distance (x-axis lower 3 figures) for EAT (left), FA (middle) and SWOW (right)

	w2v	cxt	glv	fxt	brt	brtl	rbt	rbtl	abt	abtl	abtxl	abtxxl	elt	eltl	KEl1	KEl12
MEN	.74	.76	.74	**.79**	**.26**	.07	.11	.11	-.04	-.16	.00	.01	.07	.08	.20	.08
SIMLEX	.33	34	.37	**.45**	**.36**	.16	.18	.34	.16	.18	.19	.19	.12	.12	.32	.23
WS353	.70	.71	.64	**.73**	.28	.28	.26	.09	.06	.02	.01	-.01	.17	-.03	**.39**	.33
–SIM	.76	.76	.70	**.81**	**.12**	.08	.07	-.02	-.05	-.08	-.01	.16	.01	-.14	-	-
–REL	.64	.67	.60	**.68**	.45	.48	.47	**.49**	.23	.12	.11	.26	.34	.09	-	-

Table 4: Spearman Correlation between model scores and oracle word similarity datasets MEN (Bruni, Tran, and Baroni 2014), SIMLEX999 (SIMLEX) (Hill, Reichart, and Korhonen 2015), WordSim353 (Finkelstein et al. 2002) similar (SIM) and relatedness (REL) subsets. KEl1 and KEl2 are two results listed in (Ethayarajh 2019) which extract static embeddings from BERT-base model. l1 and l12 are principle components from first and last layer BERT embedding therein.

We should not take this contradiction of similarity and asymmetry judgment on BERT slightly. If correct, our analysis shows BERT can not encode the meaning of words as well as static embedding can, but it learns contextual matching so well that it supports the modeling of $P(b|a)$ to exhibit a correct asymmetry ratio. Does it make sense at all? It is not clear if BERT learns word meaning well because similarity judgment does not provide supporting evidence. It probably is hard, if not impossible, to extract a stable word meaning representation out of the rest of the information that the dynamic embedding can encode for context matching. Even if we can, the evaluation may not be sound since they probably are correlated in the BERT's decision making.

But, do we even care about if Transformers encode meanings? Is it OK to encode an adequate amount and let context matching do the rest? On the one hand, feeding meanings to Transformers in the form of external hand-crafted knowledge has not been as successful as we had hoped, yet the work is still going on under this philosophy (Liu et al. 2020); On the other hand, we continue to relentlessly pursue larger models such as the unbelievably colossal GPT-3 (Brown et al. 2020) with 175-billion-parameters, and Table 3 shows that naively scaling up model size does not guarantee significantly better word relatedness modeling. It may be high time that we stop and think how far we should chug along the path of Transformers with bells and whistles, and if it can lead us from 0.99 to 1. Learning representation by capturing the world's complexity through automatic discovery [2] may still be the way we can follow. However, we may need either a very different family of models that encode meaning and context matching together harmoniously or a new objective very different from predicting a masked word or the next sentence. We do not know if BERT will look like it, or it will look like BERT. It must be out there waiting for us to discover.

Future Work

There are still many questions left: 1) How does lexical semantics (through asymmetry judgment) encoded in contextual embeddings change before and after transfer learning (task-specific fine-tuning) or multi-task learning, and how do we measure them? It can guide us on when to stop pre-training, how much to fine-tune or what task/data to learn from; 2) Can you explicitly encode such asymmetry during model training? For example, can you regularize the network using asymmetry data? How does it affect the accuracy of the model? 3) Can we create new architectures based on the asymmetry insight? 4) How can multi-modal embeddings be better at asymmetry judgment?

Acknowledgements

We thank anonymous reviewers for their feedbacks, and the supports from our families during this special time to make this work happen.

[2] http://www.incompleteideas.net/IncIdeas/BitterLesson.html

References

Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Paşca, M.; and Soroa, A. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27. Boulder, Colorado: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N09-1003.

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2016. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics* 4: 385–399. doi:10.1162/tacl_a_00106. URL https://www.aclweb.org/anthology/Q16-1028.

Bouraoui, Z.; Camacho-Collados, J.; and Schockaert, S. 2020. Inducing Relational Knowledge from BERT. *AAAI* .

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners .

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49: 1–47.

Brunner, G.; Liu, Y.; Pascual, D.; Richter, O.; and Wattenhofer, R. 2019. On the validity of self-attention as explanation in transformer models. *arXiv preprint arXiv:1908.04211* .

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *BlackboxNLP* .

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ICLR 2020* .

Coenen, A.; Reif, E.; Yuan, A.; Kim, B.; Pearce, A.; Viégas, F.; and Wattenberg, M. 2019. Visualizing and Measuring the Geometry of BERT. *arXiv preprint arXiv:1906.02715* .

De Deyne, S.; Navarro, D. J.; Perfors, A.; Brysbaert, M.; and Storms, G. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior research methods* 51(3): 987–1006.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Ethayarajh, K. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *ArXiv* abs/1909.00512.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems* 20(1): 116–131.

Griffiths, T. L.; Steyvers, M.; and Tenenbaum, J. B. 2007. Topics in semantic representation. *Psychological review* 114(2): 211.

Hewitt, J.; and Manning, C. D. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics* 41(4): 665–695.

Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* .

Kiss, G. R.; Armstrong, C.; Milroy, R.; and Piper, J. 1973. An associative thesaurus of English and its computer analysis. *The computer and literary studies* 153–165.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* .

Levy, O.; and Goldberg, Y. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *NIPS*.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020. K-bert: Enabling language representation with knowledge graph. *AAAI* .

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Mickus, T.; Paperno, D.; Constant, M.; and van Deemeter, K. 2019. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model.

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, G. A.; and Charles, W. G. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes* 6(1): 1–28.

Nelson, D. L.; McEvoy, C. L.; and Schreiber, T. A. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3): 402–407.

Nematzadeh, A.; Meylan, S. C.; and Griffiths, T. L. 2017. Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Reif, E.; Yuan, A.; Wattenberg, M.; Viegas, F. B.; Coenen, A.; Pearce, A.; and Kim, B. 2019. Visualizing and Measuring the Geometry of BERT. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8592–8600. Curran Associates, Inc. URL http://papers.nips.cc/paper/9065-visualizing-and-measuring-the-geometry-of-bert.pdf.

Resnik, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, 448–453. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8, 978-1-558-60363-9. URL http://dl.acm.org/citation.cfm?id=1625855.1625914.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tenney, I.; Das, D.; and Pavlick, E. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950* .

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Van Durme, B.; Bowman, S. R.; Das, D.; et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316* .

Torabi Asr, F.; Zinkov, R.; and Jones, M. 2018. Querying Word Embeddings for Similarity and Relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 675–684. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1062. URL https://www.aclweb.org/anthology/N18-1062.

Tversky, A. 1977. Features of similarity. *Psychological review* 84(4): 327.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *NIPS* .

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *BlackBoxNLP* .

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771.