WordTies: Measuring Word Associations in Language Models via Constrained Sampling

Peiran Yao and **Tobias Renwick** and **Denilson Barbosa**
Department of Computing Science
University of Alberta
{peiran, renwick, denilson}@ualberta.ca

Abstract

Word associations are widely used in psychology to provide insights on how humans perceive and understand concepts. Comparing word associations in language models (LMs) to those generated by human subjects can serve as a proxy to uncover embedded lexical and commonsense knowledge in language models. While much helpful work has been done applying direct metrics, such as cosine similarity, to help understand latent spaces, these metrics are symmetric, while human word associativity is asymmetric. We propose WordTies, an algorithm based on constrained sampling from LMs, which allows an asymmetric measurement of associated words, given a cue word as the input. Comparing to existing methods, word associations found by this method share more overlap with associations provided by humans, and observe the asymmetric property of human associations. To examine possible reasons behind associations, we analyze the knowledge and reasoning behind the word pairings as they are linked to lexical and commonsense knowledge graphs. When the knowledge about the nature of the word pairings is combined with a probability that the LM has learned that information, we have a new way to examine what information is captured in LMs.

1 Introduction

What do you think of when you see a word? Word association is a task where a human participant is shown a *cue* word, and is asked to quickly list words (formally *responses*) that come to the mind without thinking (Nelson et al., 2004; De Deyne et al., 2019). These associations provide a way to measure human representations of semantic knowledge (Rodriguez and Merlo, 2020). Similarly, researchers have been mirroring the human word association task on pretrained language models (LMs), as a method for intrinsic evaluations of word embeddings (Thawani et al., 2019) and for measuring and mitigating social biases in language models (Kaneko and Bollegala, 2021; Bommasani et al., 2020). Word associations could be used as a proxy for measuring linguistic and commonsense knowledge in language models.

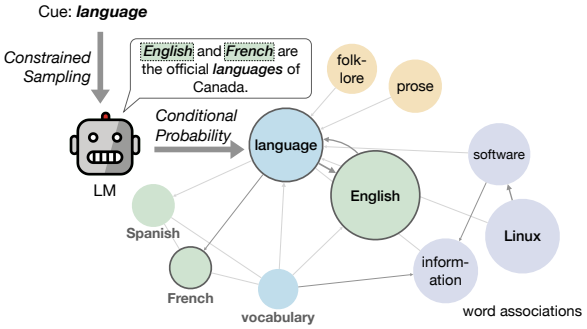

Figure 1: Overview of the workflow of our word association probing algorithm. The network plotted shows example word associations where the word *language* is a cue or response. The associations are probed from BERT (Devlin et al., 2019) using the proposed algorithm. The radius of a word circle represents the average frequency of the word being a response for one of the cues. The length of connections represents the relative associative strength between words. Note that the lengths might not be the same for the two directions between the same pair of words.

Existing approaches that probe word associations in language models (Rodriguez and Merlo, 2020; Kaneko and Bollegala, 2021; Bommasani et al., 2020; May et al., 2019) investigate the word embedding spaces of LMs. Word embeddings are contextualized in LMs, and they are converted to static embeddings for analyses with the help of external corpora or templates, which introduces confounding biases. In the meantime, associativity is often measured by the cosine similarity between the embeddings of the cue word and the response. A major problem here is that cosine similarity is symmetric, while human word associations are not (Rodriguez and Merlo, 2020).

Our code and data are available at https://github.com/U-Alberta/WordTies.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5988 - 5999

Instead of investigating embedding spaces, we propose to perform association rule mining on discrete word sequences sampled from LMs with constraints. To the best of our knowledge, this is the first application of association rule mining on the investigation of word associations in distributional semantic models. This novel approach more closely imitates human word association, and allows us to probe language models as a whole and without the use of external inputs. Our algorithm, named WordTies, samples sentences from language models with the constraint that the cue word must appear in the sentence, and uses the conditional probability that a word co-occurs with the cue word in the sample as the associativity score. The workflow of the WordTies algorithm is illustrated Figure 1. We validate our probing method by measuring the overlap between associations found in LMs by our algorithm and human associations, and testing if distance properties of human associations, like asymmetry, are preserved by our algorithm.

In another part of this work, we attempt to uncover what linguistic and commonsense knowledge and reasoning are involved in the word association process, for both humans and language models. In order to reach a reasonable cause for a given cue to response association, we link the two words simultaneously to a lexical knowledge graph (WordNet; Miller, 1995) and a commonsense knowledge graph (ASCENT++; Nguyen et al., 2021), which leads to new discoveries about word associations.

2 Human Word Associations

Human word associations exhibit certain intriguing properties, such as stability, asymmetry and intransitivity (Rodriguez and Merlo, 2020). Stability is the property that different people usually come up with similar associations, which correlates with one definition of commonsense knowledge that they are shared among most human beings (Sap et al., 2020). This suggests that word associations could potentially be used as a signal for inferring commonsense knowledge. Secondly, some associations are not symmetric, as demonstrated by Rodriguez and Merlo's (2020) example that participants indicate that `North Korea` is more closely associated with `China` than vice versa. Finally, intransitivity means the associations do not follow the triangular inequality. For example, `iPhone` is associated with `apple` and `apple` is associated with `sour`, but `iPhone` is not associated with `sour`. These two geometric properties indicate that traditional tools for interpreting language models, such as vector norms for word embeddings, will not be sufficient to discover word associations as humans do.

It was previously shown that humans often associate words based on similarity, contrast, and contiguity (Thawani et al., 2019). We further investigated what specific types of semantic knowledge and reasoning, including lexical and commonsense knowledge and reasoning, are involved in human word associations, by breaking down the relations between the cue and response word pairs.

2.1 Association Norms

Collections of human word associations are called *word association norms*. We use the data from the *English Small World of Words* (SWOW; De Deyne et al., 2019) project as the word association norms. In SWOW, up to 100 responses were each collected for 12,292 cues, along with an association strength computed from the frequency with which a word appears as a top-3 response. This serves as the ground truth when evaluating word associations generated from language models.

Compared to other popular word association norms, for example the *University of South Florida norms* (USF; Nelson et al., 2004) and the *Edinburgh Associative Thesaurus* (EAT; Kiss et al., 1973), SWOW is more contemporary, heterogenous, and includes a much larger number of cues and responses (De Deyne et al., 2019). In the USF study, participants were instructed to list words that are "meaningfully related or strongly associated" to the cue word, while in both SWOW and our analogy for LMs, no such constraints are imposed.

2.2 Semantic Knowledge

The two research questions we would like to answer here are: what semantic knowledge do humans rely on to produce word associations, and what kind of reasoning is built on that knowledge for word associations?

We attempt to answer the two questions by finding a possible "*reasoning path*" for each of the cue-response pairs in the SWOW dataset. Based on the observations of human word associations discussed at the beginning of §2, such as stability and the reliance on similarity, contrast and contiguity, it is natural to assume that there exists a certain lexical relation between the pair, or they are related by some commonsense knowledge. For

Path	Interpretation	Frequency	Source
HasProp-HasProp^{-1}	Share the same property	87,415	ASCENT++
HasProp	*response* is a property of *cue*	20,202	ASCENT++
HasProp^{-1}-HasProp-HasProp^{-1}	-	19,830	ASCENT++
ReceivesAction-ReceivesAction^{-1}	Receives the same action	19,524	ASCENT++
$\varnothing$	Synonym	14,089	WordNet
Hypernymy-Hyponymy	In the same category	10,055	WordNet
Hypernymy	Hypernym	8,815	WordNet

Table 1: Most frequent reasoning paths for the cue-response pairs in the SWOW dataset, with a potential interpretation for the path. HasProperty is shortened as HasProp. $^{-1}$ denotes an inverse relation. For example, A HasProp^{-1} B means A is a property of B. "-" indicates that there is not a concise interpretation for the path.

example, we associate dark with light out of contrast (the *antonymy* lexical relation), and apple with sour, because by commonsense being sour is a property of apples. Therefore, the reasoning paths are determined by first linking the cue word and the response to nodes of two knowledge graphs respectively: a lexical knowledge graph and a commonsense knowledge graph. The shortest path between the two nodes is regarded as the reasoning path for the cue-response pair. When calculating the shortest paths, we treat the knowledge graphs as undirected graphs by adding inverse relations for all the edges.

Knowledge Graphs WordNet (Miller, 1995) is used as the lexical knowledge graph. It provides relations between senses of English words, such as *hypernymy / hyponymy* and *antonymy*. Specifically, the version we choose is English WordNet 2020 (McCrae et al., 2020), which is a fork of the original Princeton WordNet (Miller, 1995) that accommodates emerging phenomena in the English language, and is openly available.

For the commonsense knowledge graph, we use ASCENT++ (Nguyen et al., 2021), which contains over 2 million commonsense relationships for 10,000 concepts collected from a large web corpus. At the time of writing, this is the state-of-the-art commonsense knowledge graph in terms of precision and recall. Relations in ASCENT++ are related to properties of general concepts, such as *CapableOf* and *UsedFor*.

Breakdown of Knowledge Types If the reasoning path is shorter in the lexical graph, then the cue-response pair is assumed to be more likely to involve lexical knowledge. Otherwise it is assumed to be related to commonsense knowledge. Table 2 provides a breakdown of knowledge types involved in the SWOW dataset. The majority of pairs in the dataset can be linked to the two knowledge graphs, and the shortest reasoning paths are almost evenly split between lexical and commonsense knowledge. About 20% of the pairs in SWOW have no connection in either of the knowledge graphs (categorized as *Unknown* in Table 2).

Type	Count	Frequency
Lexical	346,690	36.1%
Commonsense	417,144	43.5%
Unknown	196,066	20.4%

Table 2: Number of cue-response pairs in the SWOW dataset with reasoning paths in the lexical and commonsense knowledge graphs.

Observations of Reasoning Paths The reasoning path provides an explanation of the reasoning process behind a word association. The most frequent reasoning paths are provided in Table 1.

The majority of responses can be reached within 3 hops from the cue word, as illustrated in Figure 2. We found that the length of reasoning paths only has a slightly negative correlation with the relative order with which a response comes up in SWOW (reflected by the association strengths in the SWOW dataset), with a Spearman correlation coefficient of -0.083 ($p < 0.01$).

3 The WordTies Algorithm

3.1 Word Association Mining

The proposed WordTies algorithm finds word associations in a language model by sampling discrete sentences from the language model, with the constraint that the sampled sentence must contain the given cue word (see §3.2 for details). It then applies association rule mining to the sampled sentences, and picks the words that most frequently appear in

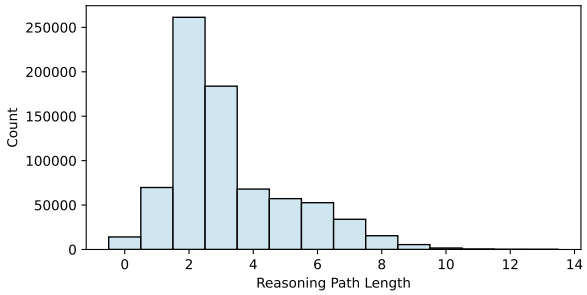

Figure 2: Distribution of reasoning path lengths in the SWOW dataset. The maximum path in SWOW is 14.

the sampled sentences as the response words[1]. Intuitively, the language model is asked to "write sentences" with the given cue word. The more likely that the LM uses a word to write such sentences, the higher chance that this word is associated with the cue word by the LM.

More formally, a language model, parametrized by Θ, is a probability distribution $P(\cdot;\Theta)$ that assigns a probability $P(\mathbf{x};\Theta)$ to any given word sequence $\mathbf{x} = x_1 x_2 \cdots x_n$. Such probability is commonly factorized by prefixes of the sequence, for example in this form:

$$P(\mathbf{x};\Theta) = P(x_1;\Theta) \cdot \prod_{i=2}^{n} P(x_i|x_{1:i-1};\Theta). \quad (1)$$

Each word pair w_1, w_2 is assigned a score $score(w_1 \rightarrow w_2)$ that indicates the associative strength with which the response word w_2 is associated with the cue word w_1. Suppose $\mathbf{x}$ is a random sequence drawn from the distribution defined by the LM, then we would like to use the following conditional probability as the score for word association:

$$\begin{aligned} &score(w_1 \rightarrow w_2) \\ \triangleq &P(\exists\, i, x_i = w_2 | \exists\, j, x_j = w_1) \end{aligned} \quad (2)$$

which is the conditional probability that given the cue word w_1 is in the sentence sampled from the LM, the response word w_2 is also in the sampled sentence.

In practice, the association score is calculated by estimating the expectation:

$$\begin{aligned} &score(w_1 \rightarrow w_2) \\ = &\mathop{\mathbb{E}}_{\mathbf{x}\sim P(\cdot;\Theta)} \frac{\mathbb{1}(\exists i, j\ x_i = w_1 \wedge x_j = w_2)}{\mathbb{1}(\exists i\ x_i = w_1)} \end{aligned} \quad (3)$$

[1]For obvious reasons, stop words are excluded.

which is done by sampling from the LM with the hard constraint that the cue word w_1 is in the sentence, and counting the words that co-occur with w_1 in the sampled sentences. It is computationally infeasible to estimate the score from unconstrained samples, i.e. sampling sentences directly from the LM and discarding the sentences without the appearance of the cue word. Word frequencies of common corpora, from which LMs are trained, follow Zipf's law and have a long-tail distribution (Zhao and Marcus, 2012), which means exponentially more samples are needed for rarer cue words.

For each cue word, we pick the words with the highest association scores as the response words, while filtering out stop words. In practice, we use the spaCy (Honnibal et al., 2020) tokenizer from its `en_core_web_sm` model to tokenize the sampled sentences, and only keep words that exist in WordNet to reduce noise. Readers can refer to Table 6-8 in the appendix for samples of the mined word associations. In the terms of association rule mining literature (e.g. Piatetsky-Shapiro's (1991)), the association score we define is the *confidence*, and the filtering of stop words is equivalent to setting a threshold on the *lift*.

3.2 Constrained Sampling

From Masked LMs Masked LMs (Devlin et al., 2019; Liu et al., 2019), or MLMs in short, are not trained with the traditional language modeling objective of minimizing the negative log likelihood of training sequences. Instead, they are auto-encoder based de-noising models trained to predict what the masked tokens should be as a distribution $P_{MLM}(x_m|\mathbf{x}_{\setminus m})$ over the vocabulary, given an input sequence $\mathbf{x}_{\setminus m}$ where the token at position m is replaced with a mask. Wang and Cho (2019) proved mathematically that a masked LM trained with this different objective still conforms to the definition of a language model described in §3.1, in the sense that it provides a probability for each sequence as a Markov random field. In the Markov random field defined by a masked LM, tokens of a sequence form a fully-connected graph, and the probability of a sequence is the normalized potential of that graph (the largest clique):

$$P_{MLM}(\mathbf{x};\Theta) = \frac{1}{Z} \prod_{i=1}^{n} P_{MLM}(x_i|\mathbf{x}_{\setminus i}). \quad (4)$$

Although the exact value of the normalizing factor Z cannot be tractably computed, it is still possible to sample from the distribution with Markov

chain Monte Carlo methods. For example, Wang and Cho (2019) provide a Gibbs sampling algorithm for masked LMs. Starting from a randomly initialized sequence, at each step we choose a random position i, sample a token from the distribution $P_{MLM}(x_i|\mathbf{x}_{\setminus i})$, and replace the token at position i with the sampled token. We modified it to impose the hard constraint that the cue word is in the sequence while sampling by keeping certain tokens fixed as the cue, as shown in Algorithm 1.

Algorithm 1 Sampling from a masked LM with a hard constraint that the cue word must be in the sequence. L_{min} and L_{max} control the length of the sampled sequence, and S is the number of MCMC steps.

sample $L \sim Uniform(\{L_{min} \cdots L_{max}\})$
sample $pos \sim Uniform(\{0 \cdots L\})$
$\mathbf{s} \leftarrow \underbrace{\texttt{[MASK]} \cdots \texttt{[MASK]}}_{L}$
$s_{pos} \leftarrow cue$
for $step \in \{1 \cdots S\}$ **do** ▷ Gibbs sampling modified from Wang and Cho's (2019).
  sample $i \sim Uniform(\{0 \cdots L\} \setminus \{pos\})$
  $s_i \leftarrow \texttt{[MASK]}$
  sample $w \sim P_{MLM}(s_i|\mathbf{s}; \Theta)$
  $s_i \leftarrow w$
end for
return $\mathbf{s}$

From Causal LMs Causal LMs factor the probability of a sequence in an autoregressive way as described in Eq. 1. Usually, sampling or decoding from a causal LM is also done in an autoregressive fashion, for example generating one token at a time from left to right. However, the conditional probability $P(\mathbf{x}|c)$ of a sequence c with the constraint c will no longer have the nice linear structure, and this poses a major obstacle for sampling.

Recent practice utilizes the fact that $P(\mathbf{x}|c) \propto P(\mathbf{x}; \Theta) \cdot P(c|\mathbf{x})$ where $P(c|\mathbf{x})$ is a differentiable classifier for the constraint, and samples from the unnormalized distribution defined by the product of the two distribution functions with variations of Hamiltonian Monte Carlo (Neal, 2011), such as Langevin Monte Carlo (Kumar et al., 2022; Qin et al., 2022). As a Markov Chain Monte Carlo process, randomly initialized text sample is updated with enough steps by gradient descent with added Gaussian noise.

In our case, we could define the constraint classifier $P(c|\mathbf{x})$ to be based on the distance (measured in embedding or simplex space) between the cue word and a token in the sequence, as suggested by Kumar et al. (2022). This Langevin Dynamics-based method provides a theoretically plausible way to apply WordTies to causal LMs such as GPT-2 (Radford et al., 2019). However, we are yet unable to produce good samples with the hyper-parameters provided and some tuning from Kumar et al.'s (2022) algorithm. We leave it as future work to continue on this direction.

3.3 Evaluation

We evaluate the performance of WordTies as a word association mining algorithm, by calculating the alignment with human associations and the precision of finding asymmetric associations, and comparing to methods from previous work.

3.3.1 Setting

Dataset We execute the experiments on a subset of 3,000 cues in SWOW. The subset of cues is chosen by uniformly sampling without replacement from the set of cues, and is available in the supplement materials. For the filtering of responses, we use English WordNet 2020 (McCrae et al., 2020) and the stop word list from NLTK (Bird et al., 2009).

Pre-trained Models The LMs we use for evaluation are BERT (`base-uncased`; Devlin et al., 2019), RoBERTa (`base`; Liu et al., 2019), and DistilBERT (`base-uncased`; Sanh et al., 2019), all implemented by Wolf et al. (2020).

Hyper-parameters In Algorithm 1, the range of sequence length is set to $L_{min} = 5$ and $L_{max} = 16$, and the number of MCMC steps $S = 100$ as suggested by Wang and Cho (2019).

3.3.2 Baselines

Contextualized2Static Bommasani et al. (2020) evaluated a scheme for averaging contextualized embeddings of a word in various contexts to a static embedding. The obtained static embeddings were then used by Kaneko and Bollegala (2021) to find associated words via cosine similarity. We replicate the static embeddings in Bommasani et al.'s (2020) work by using the best hyper-parameters they found and WikiText-103 (Merity et al., 2017). For every word in the vocabulary of the WikiText-103 corpus, we sample at most 1,000 context sentences containing that word, and average the embeddings from

Model	Method	Precision@k						Spearman's ρ
		1	3	5	10	15	30	
BERT	C2S	0.171	0.215	0.219	0.196	0.148	0.132	0.098
	Vocab	0.247	0.250	0.222	0.158	0.119	0.073	0.063
	WordTies	**0.368**	**0.352**	**0.327**	**0.281**	**0.250**	**0.195**	**0.213**
RoBERTa	C2S	0.149	0.132	0.119	0.093	0.076	0.053	-0.086
	Vocab	0.158	0.139	0.117	0.094	0.081	0.063	0.051
	WordTies	**0.255**	**0.320**	**0.212**	**0.181**	**0.161**	**0.127**	**0.163**
DistilBERT	C2S	0.177	0.222	0.197	**0.200**	**0.191**	**0.152**	0.091
	Vocab	0.254	**0.256**	0.207	0.167	0.132	0.085	0.050
	WordTies	**0.263**	0.245	**0.223**	0.189	0.168	0.133	**0.151**
	Corpus	0.543	0.452	0.399	0.325	0.287	0.224	0.228

Table 3: Evaluation results for the alignment between human word associations and LM word associations. C2S is short for the Contextualized2Static baseline, Vocab is short for the Vocab Embedding baseline, and Corpus is short for the corpus-only baseline. All reported Spearman's ρs are statistically significant ($p < 0.01$). The best results for each metric and model combination are marked in bold.

the first layer of the model for each subtoken of the word and each context.

Vocab Embedding In Rodriguez and Merlo's (2020) recent analysis of word associations in LMs, the authors directly measured the cosine similarity between embeddings in the vocabulary layer without contextualization.

Corpus Only We directly apply the same algorithm and score (2) as in WordTies to the same corpora, English Wikipedia and BookCorpus (Zhu et al., 2015), that were used to train BERT.

3.3.3 Statistical Tests

Since the WordTies algorithm involves sampling, we introduce statistical tests to make sure that an irrelevant word will not be chosen as a response by chance. Words are sampled from the multinomial distribution defined in Eq. 2, and to say that a response is not a noisy word that ends up in the top 50 most probable words by chance, the following null hypothesis needs to be rejected: there exist at least $N - 50$ words whose probability as defined in Eq. 2 is significantly lower than the chosen word where N is the size of the vocabulary. And for each pair of words, we test the null hypothesis that the probability of the first word is significantly higher than the the second by a binomial test. In our experiments, most of the words in the top-10 response list are statistically significant ($p < 0.1$). Responses that passed the tests are highlighted in Table 6-8 in the appendix. Such tests provide a guideline for choosing the number of samples to generate.

3.3.4 Alignment

We measure how good the word associations produced from LMs by the algorithms align with human associations. The alignment is measured by both precision@k, which reflects the overlap, and Spearman's correlation coefficient (ρ) between the scores in the algorithms and the strengths in SWOW, which provides an indication of whether LM and human produce word associations in the same order. The results are shown in Table 3. Our method achieves much better precision@k than the baselines on both BERT and RoBERTa, and results at the same level for DistilBERT. It also achieves higher ρ on all three models. This means associations obtained with WordTies share more similarity with human associations in terms of both word choices and strengths.

3.3.5 Asymmetry

We test if the association scores produced by WordTies can be used to find asymmetries in word associations, an important feature of human word associations that previous methods fail to accommodate. The level of asymmetry is measured by the ratio between scores of both directions of association:

$$asymmetry(w_1, w_2) = \frac{\max\left(score(w_1 \to w_2), score(w_2 \to w_1)\right)}{\min\left(score(w_1 \to w_2), score(w_2 \to w_1)\right)} \quad (5)$$

We evaluate the precision by whether the found asymmetric pair has the correct direction as in the SWOW dataset. It is meaningless to measure recall,

Model	Precision	Spearman's ρ
BERT	98.5%	0.138
RoBERTa	99.6%	0.755
DistilBERT	97.8%	0.166

Table 4: Precision and Spearman's ρ of WordTies for finding asymmetric association pairs. All ρs are statistically significant ($p < 0.01$).

because virtually every pair of words is asymmetric in human associations. For the same reason, precision is only calculated on the overlap between SWOW and the output of WordTies. Additionally, we measure the Spearman's ρ of the asymmetric measure between WordTies and human word associations to see if LM and human perceive similar level of asymmetry.

See Table 4 for the results. The baseline methods are unable to find asymmetric word associations because cosine similarity is symmetric, and therefore they are not listed for comparison. Meanwhile, WordTies is able to find asymmetric word associations that have the same direction as in SWOW, and there is a positive correlation for the level of asymmetry.

3.3.6 Discussion

Running Time On average it takes around 12s to generate 1,000 samples from BERT or RoBERTa with Algorithm 1, and 6s for DistilBERT. Time is measured on a single NVIDIA A40 GPU with a batch size of 2048. In our experiments we generated at least 3,000 samples per cue. For other models, the statistical tests described in §3.3.3 provide a framework for estimating the number of samples needed and hence the running time.

Comparison of Methods We have already discussed how the symmetrical nature of cosine similarity used in previous methods do not fit well with word association. Adding to that, we suspect there are 2 other reasons behind the inferior performance of baseline methods: First, previous methods try to obtain a unified embedding for each word from contextualized models by either averaging embeddings in different contexts, or simply using the layers before contextualization. Such conversions defeat the purpose of building contextualized models and incur information loss. For example, contextualized BERT embeddings for a polysemous word are distinct enough for accurate word sense disambiguation (Hadiwinoto et al., 2019) while averaging eliminates the distinctions. Second, embeddings from a contextualized model must be computed with a context sentence, which is sampled from an external corpus in the Contextualized2Static method. The choice of corpus or context affects embeddings, which introduces confounding biases to word association measurements. Conversely, a "pseudo corpus" is generated from the LM in WordTies, similar to a *training data extraction attack* (Carlini et al., 2021) on the LM. No external factors are involved so it is certain that we are only examining the LM itself. When we apply the same score as in WordTies to the real corpus used to train the LM, we observe an overlap with human word associations that is larger than any of the LMs evaluated. This observation hints that, no matter a LM can overcome reporting bias (Shwartz and Choi, 2020) and extrapolate beyond the corpus or not, it still has a gap to reach the upper bound of word associations.

Comparison of Models BERT was trained on English Wikipedia and BookCorpus (Zhu et al., 2015), and achieves the best overlap with human word associations. RoBERTa is a replica of BERT with more carefully selected hyper-parameters and a larger training corpus, which additionally incorporates news, stories, and web content. However, with better training settings it performs worse than BERT on the word association task. In this sense, world knowledge in RoBERTa is not as similar to that of humans, and we suspect it is because of the relevance and quality of the additional training corpus. The sampling process in WordTies is a reflection of the corpus (Carlini et al., 2021), and we observed more URLs and email addresses in the samples from RoBERTa, which are irrelevant to the knowledge involved in word association. DistilBERT is a smaller model trained on the same corpus as BERT and with BERT as the teacher. Embedding-based baselines perform on par with sampling-based method for DistilBERT, and we conjecture the reason to be that DistilBERT is not as good an MLM in the first place. Sanh et al. (2019) only reported that the model perform equally well as BERT on downstream tasks but not the MLM objective, and few studies used DistilBERT in MLM-based zero-shot tasks.

4 Language Model Word Associations

WordTies, as a more suitable method for probing word associations in LMs, enables us to scrutinize

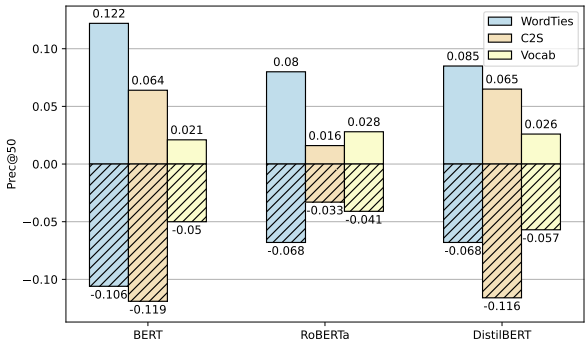

Figure 3: Precision@k for cue-response pairs involving **commonsense (upper)** and **lexical (lower, hatched)** knowledge. In each group of bars, the bars from the left to the right are for WordTies, Contextualized2Static and Vocab Embedding respectively.

Model	Spearman's ρ
BERT	-0.248
RoBERTa	-0.243
DistilBERT	-0.239

Table 5: Correlation between precision@50 and reasoning path length for different models. All ρs are statistically significant ($p < 0.01$).

the properties of those associations.

Semantic Knowledge We observed that LMs are slightly better at associating words by commonsense knowledge than lexically, judging by the precision@k for cue-response pairs broken down by the type of knowledge (Figure 3). This is consistent with the finding that humans use commonsense knowledge more for associations (Table 2).

Reasoning Path Length LMs' ability to find human-like associations is negatively associated with the length of the reasoning path to the response. In other words, the more hops to get from the cue to the response in the KGs, the harder for LMs to associate the cue with the response. See Table 5 for the correlation coefficients. Meanwhile, longer reasoning paths only slightly degrade human association strength (§2.2).

5 Related Work

The study of Rodriguez and Merlo (2020) is mostly similar to ours, where they concluded that properties of human word associations, discovered in the 1970s (Tversky, 1977; Tversky and Gati, 1978; Tversky and Hutchinson, 1986), still hold in language models. They probed associations by ranking words by the cosine similarity of embeddings in the vocabulary layer, and measured asymmetry by handcrafted templates. Evert and Lapesa (2021) also tested word associations with word embeddings, but they held the same view as us that it is self-contradictory to obtain decontextualized embeddings from contextualized LM, and therefore did not extend their study on LMs. Measuring and mitigating social biases in pre-trained LMs, often formulated as measuring associations to a certain set of words, is a more popular task. Associations to the words related to social aspects are often measured by the cosine similarity of embeddings aggregated from context sentences (May et al., 2019; Bommasani et al., 2020; Kaneko and Bollegala, 2021). As we have been arguing, cosine similarity is not compatible with the asymmetry of word associations, while our algorithm takes asymmetry into consideration. In some work, biases are also measured via constrained generation, where the constraints (often prompts or templates) are collected from the web (Dhamala et al., 2021) or by crowdsourcing (Nangia et al., 2020). In comparison, our method relies on no external resources, and no confounder is introduced consequently.

Constrained text generation is used to evaluate the commonsense reasoning ability of LMs through other tasks. CommonGen (Lin et al., 2020) is a task where, instead of only one *cue* word as in our study, multiple words pertaining to commonsense concepts are required to be present in the generated text, as a way to measure how well LMs can link concepts together with commonsense knowledge. In abductive commonsense reasoning (Bhagavatula et al., 2020), LMs are used to complete text when the beginning and ending are given, to test their ability to reason about pre- and post-conditions.

It is considered non-trivial to impose constraints on left-to-right generations for causal LMs. Mostly, recent work (Qin et al., 2022; Dathathri et al., 2020) focus on constrained (also known as controlled) decoding, a related problem of finding the sequence that maximizes the likelihood, by modifying the original distribution. Prior to the Langevin Dynamics algorithm by Qin et al. (2022) and Kumar et al. (2022), Miao et al. (2019) proposed CGMH, a constrained sampling algorithm in discrete space based on Metropolis-Hasting sampling, but it uses a bidirectional causal LM to reduce computation (i.e. the LM also predicts the previous word based on suffixes). More recent causal LMs, such as GPT (Radford et al., 2019; Brown et al., 2020), are uni-

directional, and it is therefore not very meaningful to apply CGHM in our work.

In a broader context, it has been an interesting idea to try to explain the behavior of neural networks by optimizing over the input. Sampling from a language model, as in our WordTies algorithm and in Carlini et al.'s (2021), can be seen as optimizing over the discrete input text sequence to minimize the negative log-likelihood with noise, and it provides a way to uncover how LMs associate words or properties of the training corpus. Bäuerle and Wexler (2020) optimized the activation of certain neurons in BERT over the input sequence, as an attempt to find the responsibilities of individual neurons, and Goh et al. (2021) applied similar thoughts on vision-language models.

6 Conclusion

In this study, we verified the proposition that examining discrete sequence samples from LMs is a better approach than inspecting embedding spaces for word associations. We also explored properties related to semantic knowledge and reasoning in both human and LM word associations. These revealed the high potential of using word associations as a proxy for probing, and as a signal for finetuning language models.

Limitations

We have yet to apply the WordTies algorithm to popular LMs such as GPT-2 that are causal, despite having provided a theoretically sound method to do so in §3.2. Due to the constraint of computation resources, we only evaluated our algorithm on the base version of popular pre-trained LMs. Models with a larger number of parameters, such as `bert-large-cased` and `roberta-large`, are yet to be evaluated. For the same reason, we were only able to run the experiments on a subset of SWOW. Our method is notably slower than simply running k-nearest-neighbor search on embedding spaces, although the running time is still acceptable and we have a method for estimating the running time required (§3.3.6). Potential downstream use cases of word associations, such as measuring social biases in language models, are not evaluated in this paper.

Ethics Statement

As discussed in §5, measuring and mitigating social biases have been a prominent and motivating application of word associations. The algorithm we proposed contributes a practical way to measure associations to words related to social aspects (such as profession, gender, race, and other aspects) in language models with higher precisions and fewer confounders. These associations, in addition to being a measure of biases, could potentially serve as a signal for fine-tuning LMs, and lead to language models with less biases.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), 5010405. This work is also supported in part by a gift from Scotiabank. Robot icon in Figure 1 was designed by OpenMoji, under CC BY-SA 4.0 license.

References

Alex Bäuerle and James Wexler. 2020. What does BERT dream of? In *3rd Workshop on Visualization for AI Explainability (VISxAI) at IEEE VIS, Online, October 26, 2020*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine

Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 862–872. ACM.

Stefan Evert and Gabriella Lapesa. 2021. FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 588–595, Online. Association for Computational Linguistics.

Gabriel Goh, Chelsea Voss, Daniela Amodei, Shan Carter, Michael Petrov, Justin Jay Wang, Nick Cammarata, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *OpenAI blog*.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Mattew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1256–1266. Association for Computational Linguistics.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of English and its computer analysis. *The computer and literary studies*, 153.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Constrained sampling from language models via langevin dynamics in embedding spaces. *CoRR*, abs/2205.12558.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1823–1840. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH: constrained sentence generation by metropolis-hastings sampling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6834–6842. AAAI Press.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1953–1967. Association for Computational Linguistics.

Radford M. Neal. 2011. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman and Hall/CRC, Boca Raton, FL, USA.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2021. Refined commonsense knowledge from large-scale web contents. *CoRR*, abs/2112.04596.

Gregory Piatetsky-Shapiro. 1991. Discovery, analysis, and presentation of strong rules. In Gregory Piatetsky-Shapiro and William J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. COLD decoding: Energy-based constrained text generation with langevin dynamics. *CoRR*, abs/2202.11705.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maria A. Rodriguez and Paola Merlo. 2020. Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 376–385, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6863–6870. International Committee on Computational Linguistics.

Avijit Thawani, Biplav Srivastava, and Anil Singh. 2019. SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 43–51, Minneapolis, USA. Association for Computational Linguistics.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327–352.

Amos Tversky and Itamar Gati. 1978. Studies of similarity. *Cognition and categorization*, pages 79–98.

Amos Tversky and J Hutchinson. 1986. Nearest neighbor analysis of psychological spaces. *Psychological review*, 93(1):3–22.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Qiuye Zhao and Mitch Marcus. 2012. Long-tail distributions and unsupervised learning of morphology. In *Proceedings of COLING 2012*, pages 3121–3136, Mumbai, India. The COLING 2012 Organizing Committee.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Example Associations

The following tables provide examples of word associations found by WordTies. Human associations from SWOW are also included for reference. Words that did not pass the statistical tests are in italics.

Model	Top-10 Responses
BERT	web, search, site, page, www, online, news, map, available, internet
RoBERTa	android, search, maps, play, apple, moon, chrome, amazon, *web*, *windows*, copy, *mobile*
DistilBERT	web, search, page, maps, database, map, site, website, chrome, index
Human	search, internet, find, search engine, corporation, maps, web, email, engine, evil

Table 6: Example associations with **Google** as the cue.

Model	Top-10 Responses
BERT	university, professors, associate, assistant, former, college, history, english, science, david, senior, physics
RoBERTa	student, university, prof, ex, assistant, co, school, college, education, senior
DistilBERT	university, associate, assistant, emeritus, college, former, visiting, institute, phd, *senior*
Human	teacher, university, college, doctor, school, smart, lecturer, mentor, educated, tweed

Table 7: Example associations with **professor** as the cue.

Model	Top-10 Responses
BERT	data, computes, let, web, site, sofa, *net*, *call*, system, *power*
RoBERTa	comp, python, using, make, java, import, test, class, function, version
DistilBERT	picture, pictures, data, function, *simulator*, code, using, *math*, *map*, *string*
Human	calculate, computer, add, figure, understand, think, math, data, does not, figure out

Table 8: Example associations with **compute** as the cue.