Word associations and the distance properties of context-aware word embeddings

Maria A. Rodriguez
University of Geneva
Maria.AnduezaRodriguez@unige.ch

Paola Merlo
University of Geneva
Paola.Merlo@unige.ch

Abstract

What do people know when they know the meaning of words? Word associations have been widely used to tap into lexical representations and their structure, as a way of probing semantic knowledge in humans. We investigate whether current word embedding spaces (contextualized and uncontextualized) can be considered good models of human lexical knowledge by studying whether they have comparable characteristics to human association spaces. We study the three properties of association rank, asymmetry of similarity and triangle inequality.

We find that word embeddings are good models of some word associations properties. They replicate well human associations between words, and, like humans, their context-aware variants show violations of the triangle inequality. While they do show asymmetry of similarities, their asymmetries do not map those of human association norms.

1 Introduction

What do people know when they know the meaning of words? Lexical semantic knowledge is rich and structured and comprises the knowledge of the relation between a word and its related concept, the relationships between concepts among themselves and between words themselves. Word associations —spontaneous elicitation of words by similarity, contrast or contiguity— have been widely used to tap into lexical representations and their structure, as a way of probing semantic knowledge in humans (De Deyne and Storms, 2014).

To process language in a way that mirrors human expectations and to develop usable technology, we need computational representations of the meaning of words that correspond to speakers' knowledge of words (De Deyne et al., 2016). Current word embeddings methods, without context or context-aware, represent lexical semantic knowledge as coordinates in a multi-dimensional space. Work in cognitive psychology, however, has argued that human lexical knowledge and in particular word associations are not well represented by geometrical models (Tversky, 1977; Gati and Tversky, 1982; Tversky and Hutchinson, 1986). More precisely, similarity judgments of associations in humans have been shown not to exhibit the properties of true distances. For example, associations are not symmetric: speakers will indicate that North Korea is more similar to China than China is to North Korea.

In an analysis inspired by Tversky (1977)'s critique of spatial measures of similarity in word associations, and by recent work on topic models and word embeddings (Griffiths et al., 2007; Nematzadeh et al., 2017), we compare current word embedding spaces to the large human norming study by Nelson et al. (2004) and investigate whether these word embedding spaces, especially those based on transformers, —the lexical representations used by almost all our current architectures— share some properties with human word associations, as human-like representations of lexical meaning.

2 Word Associations

The task of word free association consists in providing a *cue* word to a speaker and asking to produce, fast and without thinking, other words that come to mind, called the *target* words. These collections of free associations present interesting characteristics. First, they present a certain stability across speakers, so that it is possible to determine the most associated word on average, the second most associated and so forth. Second, they present interesting qualitative properties. They exhibit asymmetry of similarity judgments, as in the example on China and North Korea above. Finally, they also exhibit

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 376–385
Online, November 19-20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Exp.	Description	Word embedding	Hypothesis	Method
1	Find the top-k neighbors near the cue	BERT not in context, lemmatized and unlemmatized	Top-k near cue in BERT are associates in human associations.	kNN, median rank and P@K
2.1	Asymmetry of countries	BERT in context	Non-prominent countries are more similar to prominent countries than vice versa.	Cosine
2.2	Asymmetries as frequencies	Output of Exp.1	Frequent words are more often associates of less frequent words than vice versa.	Cues and targets from Nelson, cues and targets from Exp. 1; hits of Google's search engine.
2.3	Asymmetries of hypernyms and hyponyms	BERT in context	If salience = 'more general', hyponyms are more similar to hypernyms than the reverse. If salience = 'more specific', the effect is the opposite.	Nouns and verbs from Nelson; hypernyms/hyponyms from WordNet; cosine.
2.4	Asymmetry as neighbourhood density	BERT not in context unlemmatized; FastText (only for countries)	A semantically richer word elicits a greater number of close neighbours than a fainter word.	Extraction of asymmetric pairs from Nelson; kNN; cosine (threshold ≥ 0.2)
3	Violation of triangle inequality	BERT in context	BERT embedding space violates the triangle inequality.	Extraction of asymmetric triples; cosine, τ threshold.

Figure 1: Summary of experiments.

violation of transitivity (called in the literature, violation of the triangle inequality). For example, *asteroid* is highly associated with *belt*, and *belt* is highly associated with *buckle*, but *asteroid* and *buckle* show little association.

Feature-based explanations of these properties makes use of the richness of the representation, in terms of number of features and the proportion of shared features between two representations (Tversky, 1977). For example, the asymmetry of similarity is explained by the assumption that only a few of the large number of features in speakers' mental representation of China are shared with North Korea, while the representation of North Korea involves a small number of features, many of which are shared with China. More recent approaches have proposed probabilistic representations in terms of topic models (Griffiths et al., 2007). Violation of the triangle inequality is explained in these models as an effect of the fact that topic models represent different uses of a word as different topics, and different topics do not necessarily preserve similarity.

We investigate whether current word embedding spaces can be considered good models of human lexical knowledge by studying whether they have comparable characteristics to human association spaces. We study the three properties of associations mentioned above: rank, asymmetry of similarity and violation of triangle inequality (lack of transitivity). More specifically, first, we verify that the notion of word association makes sense in word embedding spaces and compare if the target words that occur as preferentially associated with a cue word in a word embedding space correspond to the closest words in human association norms.

Then, we analyse asymmetry in similarity in several different ways: we look at its correlation with cue and target frequency, with neighbourhood densities and lexical entailment. Finally, we model violations of triangle inequality by looking at how similarity of words spreads across word embedding spaces. A summary of the experiments discussed in more detail in the following sections is given in Figure 1.

3 Data

To perform the studies indicated above, we use two sets of artificial word embedding data, and one set of human association norms.

3.1 Word embedding data

BERT Devlin et al. (2018) propose BERT, a transformer-based model that uses an attention mechanism to extract the context of words and subwords from text. The innovation of BERT is the application of a bi-directional training to the Transformer, achieving a better use of context from text than systems with unidirectional training. BERT is pretrained on the BookCorpus (800M words) and English Wikipedia (2500M words). The BookCorpus (Zhu et al., 2015) is a collection of 11,038 books available on the Web, from 16 different genres, taking into account only books with more than 20K words to avoid noise coming from shorter stories. In all experiments below, we use the Hugginface version of BERT[1], specifically the "bert-base-uncased" model that we expressly do not modify.

FastText The original FASTTEXT model is based on Wikipedia dumps[2] in nine differerent languages including English (Bojanowski et al., 2017). However, in this work, we used the pre-trained FASTTEXT embeddings provided by the official site of FASTTEXT, that we expressly do not modify.[3] The embeddings are trained on 600-billion tokens from CommonCrawl[4], resulting in two-million word vectors with subword information.

3.2 Human word association norms

Nelson et al. (2004) propose a large dataset of free association, rhyme and word fragment norms, elicited from more than 6000 participants. The participants were asked to write the first word that came to mind when presented a particular stimulus word. More than 750'000 free associations (called targets) from a total of 5019 stimulus words (called cues) were collected. The related quantitative information (such as number of participants and measures of association strength) were calculated.[5]

Other word association norms exist, such as the Small World of Words (De Deyne et al., 2019), but in this study, we wanted to be able to compare, at least indirectly, our work to previous work where the Nelson's norms were used (Griffiths et al., 2007; Nematzadeh et al., 2017).

4 Experiment 1: The notion of association

To simulate the process of production of associates of a free association task, we used the k-nearest neighbours algorithm (kNN) to find the top-k words that are near the cue. The intuition is that words near the cue word in the embedding space are probably associates. We use cosine similarity as the metric to find the nearest neighbours.

We compare if the target words that occur as preferentially associated with a cue word in a word embedding space correspond to the closest words in human association norms. For this comparison, we use two measures: median ranking and P@k, following Griffiths et al. (2007).

Rank	Unlemmatized Median Rank	Unlemmatized P@K (%)	Lemmatized Median Rank	Lemmatized P@K (%)
1	4	13.02	3	24.18
2	12	28.09	10	43.26
3	35	43.64	27	55.86
4	94	53.59	69	64.53
5	230	61.62	157.5	69.94

Table 1: Results of Experiment 1 for the unlemmatized and lemmatized data. For each rank in human associations (Rank column), we computed the median ranking on the basis of the BERT vector space. Furthermore, we computed the precision at K (P@K with K = 1, 2, 3, 4, 5 corresponding to the ranks) where we check if the first associate in the human associations appears in the top K associates of BERT.

Method As human associations in Nelson are lemmatized, we tested both the lemmatized and unlemmatized versions on the "raw" BERT word embeddings obtained from the vocabulary of the model. [6]

The median rank is a measure of central tendency of the median rank in BERT for the n-ranked associate target in Nelson. For example, in Figure 2 for the unlemmatized data, we would have to take the median across ranks of the ranks given in BERT (1, 1, 2, 32, 22) for the first ranked target in Nelson.

This measure is calculated as follows. For each cue in Nelson, we extract the top-k nearest neighbours in the BERT space and rank the results by their cosine similarity in descending order. Then, for each same cue and each one of its targets in Nelson, we calculate the target's rank in BERT (see Figure 2). The median of these ranks is calculated for each of the Nelson's rank (see Table 1).

Using the same ranked lists, we calculate Precision at K (P@K), where $K = 1, 2, 3, 4, 5$. P@K tells us if the first associate in the human associa-

[1] https://github.com/huggingface/transformers

[2] https://dumps.wikimedia.org

[3] https://fasttext.cc/docs/en/english-vectors.html

[4] https://commoncrawl.org

[5] The words we use in this work are found in appendix A and B in Nelson et al. (2004). Appendix A presents the list of targets produced by each cue and Appendix B presents the list of cues that elicit a particular target.

[6] These raw vectors come with the pre-trained model of BERT. They are extracted from the vocabulary of the model, thus not in context. They can be downloaded from `https://github.com/ajitrajasekharan/bert_mask`. Their unlemmatized version uses the words and the embeddings as they come with the model. Their lemmatized words are derived using the WordNet lemmatizer from NLTK `http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.wordnet` and the word embeddings corresponding to a lemmatised word is the sum of all the word embeddings of the unlemmatised word forms.

	Unlemmatized				
	ABDOMEN	YELL	SAW	RISE	NECESSARY
Human Associations	**stomach (1)**	**shout (1)**	**see (2)**	lift (32)	important (22)
	belly (4)	scream (5)	hammer (207)	fall (37)	need (27)
	organ (3399)	whisper (27)	look (239)	stand (38)	must (263)
	body (4418)	loud (189)	cut (294)	wake (72)	money (11869)
	muscle (8368)	cheer (194)	tool (350)	shine (73)	object (13096)
BERT Predictions	**stomach**	**shout**	sees	rises	required
	abdominal	yells	**see**	risen	needed
	torso	yelled	seen	rising	essential
	belly	yelling	seeing	Rise	unnecessary
	groin	scream	Saw	rose	appropriate
	Lemmatized				
	ABDOMEN	YELL	SAW	RISE	NECESSARY
Human Associations	**stomach (1)**	**shout (1)**	**see (1)**	lift (23)	**need (3)**
	belly (4)	scream (3)	look (57)	lower (25)	important (20)
	organ (217)	whisper (15)	hammer (80)	fall (43)	must (221)
	body (250)	noise (39)	cut (221)	wake (52)	object (1867)
	sex (276)	anger (83)	tool (229)	stand (70)	money (9684)
BERT Predictions	**stomach**	**shout**	**see**	Rising	essential
	abdominal	yelled	Saw	arise	require
	torso	scream	noticed	raise	**need**
	belly	roar	felt	ascend	unnecessary
	groin	growl	heard	Rise	appropriate

Figure 2: Unlemmatized and lemmatized rankings in word associations from Nelson et al. (2004) and from BERT predictions, listed in descending order from the first to the fifth rank. Human associations are ranked by cosine similarity. In parentheses, the rankings of the human associations in the BERT space. See the text for how the examples were chosen.

tions appears in the top-k associates of BERT. For example, Figure 2 for the unlemmatized data shows a P@K=1 of 2 out of 5 and a P@K=3 of 3 out of 5. For the lemmatized data, we have a P@K=1 of 3 out of 5 and a P@K=3 of 4 out of 5.

Results The results are shown in Table 1 and examples are shown in Figure 2. The median BERT ranking for the first human associate in unlemmatized and lemmatized associations is respectively 4 and 3.

For the unlemmatized version, the first associate in the human word associations is the word with the highest ranking in BERT in 13.02% of cases and in the top 5 ranks of BERT in 61.62% of the cases. For the lemmatized version, these values improve to, respectively, 24.18% of cases and 69.62% of the cases.

As can be seen in Table 2, the results in BERT are convincing. For the unlemmatized data, the first three columns show examples where BERT ranks the right associate at or near the top of its list. The last two columns are examples of not very good association rankings in BERT. Notice the third column, which shows the limitations of an unlemmatized approach as BERT does not correctly distinguish between forms of the same word. For the lemmatized data, we see an improvement in the prediction of BERT: the rankings of the human associations are in general lower than the rankings of the unlemmatized version. As a measure of indirect comparison, previous work (Griffiths et al., 2007) indicates that a topic model trained on the TASA corpus (Landauer and Dumais, 1997)[7], and compared to the same norms by Nelson's gives a median rank of 50.5 and predicts the first associate correctly in 10.24% of cases with an improvement of over 60% over a frequency baseline. They indicate that this improvement over the baseline results from having reduced dimensionality. Clearly, word embeddings benefit from similar properties, and to an even greater extent, being trained on much larger data sets and being based on non-linear dimensionality reductions.

5 Experiment 2: Asymmetries

Perhaps more interestingly than simple ranking measures, human word associations also exhibit peculiar qualitative properties that computational systems must also exhibit if they want to be considered human-like. These properties are especially interesting in a discussion of word embeddings as they seem to specifically defy a representation of

[7] This is a collection of reading materials spanning the school years from grade school to college.

the lexicon as a geometric space, in that they violate metric axioms, such as symmetry of distances and the triangle inequality.

Our intuition is, however, that context-aware word embeddings are no longer linear geometric representations in multi-dimensional space and that, as such, could mirror the geometrically-warped properties of human associations. For this reason, we use cosine, a symmetric similarity operator: any asymmetry found this way is to be ascribed to the context-aware vector and not to the similarity operator. Here we want to model the asymmetric association between cue and target words (Tversky, 1977). We want to model the intuition that these asymmetries stem from a richer and more specific representation for certain words (like China) and less specific or vaguer representation for less salient words (like North Korea).

5.1 Experiment 2.1: Asymmetry of countries

We start by testing the data discussed in Tversky (1977). It has been observed that in contexts that elicit similarity, the more prominent word is preferentially the second element in the similarity. So, for example, speakers prefer *North Korea is similar to China* to *China is similar to North Korea*.

Twenty-one pairs of country names served as stimuli. The pairs were constructed so that one element was more prominent (A) than the other (B) (e.g., China-North Korea, USA-Mexico, Belgium-Luxembourg). We used the pairs found in Tversky and Gati (1978), but also updated the list of countries. As Tversky's list was created in the 70s, some countries do not exist today or have changed their name (e.g., USSR is replaced by Russia, West Germany by Germany, Ceylon by Sri Lanka).[8]

Method Following Tversky's experimental procedure, we contextualised word embeddings of the country names by setting the names in three context sentences: "A is similar to B", "A is essentially B" and "A is roughly B". We indicate this context below as (A, B). As we wanted to test asymmetries, we also constructed these sentences in the opposite direction e.g. "B is similar to A". We indicate this context below as (B, A). In what follows, A refers to prominent countries and B to less prominent countries, so people prefer "B is similar to A".

On this basis, we used the sentences as input for BERT, we extracted the word vectors (in context) of each country name and we tested, for example, if $cos(A{=}China, B{=}NK) \geq cos(B{=}NK, A{=}China)$ or if $cos(B{=}NK, A{=}China) \geq cos(A{=}China, B{=}NK)$.

Context	$cos(B, A) \geq cos(A, B)$
A is similar to B	76.19%
A is essentially B	57.14%
A is roughly B	66.67%

Table 2: Results of Experiment 2.1: percent of times the cosine similarity in BERT is higher when the more prominent country is in second position ($cos(B, A)$), matching people's preferences, compared to when it is in first position ($cos(A, B)$).

Since BERT spaces take context into account, they should be able to detect the differences between the order of the words in the context and if they replicate human associations, we should find that $cos(B{=}NK, A{=}China) \geq cos(A{=}China, B{=}NK)$ more often than the reverse.

Results As showed by Table 2, the cosine similarity is higher when the less prominent country is in the first position. This results, then, confirms human judgements. But why is it so?

The explanation for the human result has been in terms of richness of representation and the relative proportion of common features and contrasting features, features unique to one of the two elements being compared (Tversky's contrast model). Griffiths et al. (2007) also show, however, that frequency is a strong predictor of salience, so that this effect could be simply due to frequency. We verify then how much frequency in general is related to salience.

5.2 Experiment 2.2: Asymmetries as frequencies

Frequency is a strong aspect of saliency and it could be that frequent words are more often evoked as associates of less frequent words than vice versa (Griffiths et al., 2007).

Method We use the hits of Google's search engine, because the British National Corpus[9], a properly balanced corpus of over 100 million tokens, did not contain all the countries we needed. We find that more prominent countries are more frequent than less prominent countries in 76.19% of the cases. The results are the same if we use the updated list of countries.

[8] The word lists are shown in the supplementary materials.

[9] http://www.natcorp.ox.ac.uk/

Rank	Avg Frequency Word Association	Avg Frequency BERT
1	1369	8716
2	1942	8432
3	2804	9691
4	2696	10619
5	2546	13136

Table 3: Results of Experiment 2.2 for unlemmatized associations. For each rank, we compute the average frequency of the targets in human associations and BERT predictions for a given cue.

Context	Mean cos(he,ho)	Mean cos(ho,he)	p-value
similar	0.452	0.443	< 0.0001
essentially	0.455	0.447	< 0.0001
roughly	0.450	0.443	< 0.0001

Table 4: Paired samples t-test on asymmetric verbs and nouns with He/Ho relation extracted from human word associations with BERT.

Given that we know that word embeddings encode information about the frequency of the underlying words (Schnabel et al., 2015), then the effect of country similarity could be an effect of frequency.

Could a frequency explanation, though, be extended to all words in BERT embeddings? We quantify the frequency of each cue ($freq_{cue}$) and target word ($freq_{target}$). As in Experiment 1, we conduct two tests for unlemmatized and lemmatized data. We used both unlemmatized and lemmatized lists of words in the British National Corpus to retrieve the frequency of words. Then, we extract from the corpus both the pairs of cues and targets from Nelson's human associations and the pairs from BERT, ordered by the rankings obtained in Experiment 1.

Results As shown in Table 3, the average frequency for unlemmatized (and lemmatized) associations in BERT is higher than in human word associations.

The human associations confirm what had already been found in Griffiths et al. (2007): cues tend to elicit targets with higher frequencies than themselves. Precisely, in the unlemmatized version, 62.95% of associations have a target with higher frequency. In contrast, if we compute the frequency of the targets found by BERT, we find that only 30% of associations have a target with higher frequency.

This indicates that, in general, if we were to find human-like asymmetric judgments of similarity in BERT spaces it would not be a frequency effect. The question remains, however, of how to operationalise the notion of salience.

5.3 Experiment 2.3: Asymmetry of hypernyms and hyponyms

Similarly to the experiment in section 5.1, we also tested the asymmetric associations where cues and targets are common nouns and verbs, and not proper nouns. Unlike proper nouns, such as country names, whose salience is related to the external world and the prominence of the referred country in it, for common nouns and verbs, the notion of salience also needs to be defined in linguistic terms, possibly as richness of representation.

We test here two possible interpretations that make opposite predictions in the case of common nouns and verbs: a rich representation can be interpreted as meaning 'more prototypical, more general', but also 'more specified'. These two operationalisations can be teased apart by looking at lexical entailment: the hypernym is more general and the hyponym is more specified. We conducted a test where the cue-target pairs were in a hyponym-hypernym relation, for short Ho/He relation.

Pairs of nouns and pairs of verbs were extracted from the Appendix A of Nelson et al. (2004), using only asymmetric pairs for a total of 2735 pairs. We further extract the pairs that are in a hyponym-hypernym relation using WordNet (Miller, 1995). If the target possesses some hyponyms in Wordnet, then we check that there is a hyponym that has the same category as the target in the list of hyponyms, i.e. if the target is a verb then the hyponym, in addition to being an existing association of the target, has to be the first verb in the list of hyponyms. We extracted a total of 79 pairs of verbs and 573 pairs of nouns.[10] From these pairs, we constructed three types of sentences, as in section 5.1.

Depending on whether salience and richness of representation means 'more general' or 'more specific', we have two different expectations. If it means 'more general', then we should find the same effect as for proper nouns, where the preference is for the hyponym to be more similar to the hypernym than the reverse (e.g. *Dancing is similar to moving, A dog is essentially an animal*), an intuitive preference.

But an interpretation as 'more specified' yields a different prediction. In a He/Ho pair, the relation

[10] The word lists are shown in the supplementary materials.

is formalised by subsumption, so the features of the hypernym are all present in the representation of the hyponym. Hence, the number of matching features will be a greater proportion for the hypernym than for the hyponym and thus the expectation is reversed. We expect to prefer *Moving is similar to dancing, An animal is essentially a dog*, which does not seem natural.

The results in Table 4 show that, in general, the (He, Ho) pairs have a higher cosine similarity than (Ho, He) pairs. In all contexts, these results are confirmed by a paired samples t-test, where p-value<0.05. This confirms that BERT spaces encode salience in terms of richness of specifying features. But this is contrary to human intuition, at least to our intuition, which appears to prefer to identify salience with prototypicality and generality.

5.4 Experiment 2.4: Asymmetry as neighbourhood density

The intuition of vectorial spaces is that the meaning of a word is determined by the neighbouring words. If that is the case, then, a semantically richer word will elicit a greater number of close neighbours than a fainter word. We model this asymmetry by looking at the density of neighbouring words.

Method We extract words from human associations where we can find an asymmetric association: the cue word produces a certain target, but this specific target does not produce back the initial cue. There are 18’571 of these pairs of words. We compute the density around these words in the embedding space using kNN with cosine as the metric. We quantify the number of associates around a word given a threshold for the cosine similarity ($cosine \geq 0.2$). The threshold 0.2 was chosen to have enough data to analyse, a higher threshold would not provide samples with adequate numbers of responses.

Results In human associations, there is an imbalance between cue and target suggesting that targets are more salient, so that we expect them to have a denser neighbourhood in vectorial space. But in BERT’s vectors corresponding to human asymmetric associations (those where a cue elicits a target but not the reverse), the target is denser than the cue in only 26.58% of cases.[11]

We also tested the country data described in section 5.1 with the same procedure. The results show that the more prominent country has a higher density in 23.8% of cases. Recall that this result is not a frequency effect, as indicated by the country frequencies reported in experiment 2.2.[12]

In conclusion, context-aware vectorial spaces do not encode asymmetries in similarities analogously to human associations. In human associations, targets are more frequent, and more general, while in vectorial spaces only targets as proper nouns are more frequent, but both common nouns and verbs as targets are less frequent, more specific and have a sparser neighbourhood than their cues.

6 Experiment 3: Violations of triangle inequality

Human word associations violate the triangle inequality, also called transitivity here. It is easy to find sets of words that have this property. For example, *asteroid* is highly associated with *belt*, and *belt* is highly associated with *buckle*, but *asteroid* and *buckle* have little association.

The triangle inequality restricts the possible relationships between three words in embedding spaces: If w_1 and w_2 are highly associated and w_2 and w_3 are highly associated, then we expect w_1 and w_3 to be highly associated.

However, as already motivated in section 5, our intuition is that context-aware word embeddings are no longer following the rules of linear geometric representations. For this reason, we use the cosine that, in addition to being symmetric, does not violate the triangle inequality. If any violation of the triangle inequality is found, it is attributable to the context-aware vector and not to the similarity operator.

Method We extracted a subset of the triples (w_1, w_2, w_3) from Appendix A in Nelson et al. (2004)’s norms for a total of 12664 triples, where both (w_1, w_2) and (w_2, w_3) are pairs of asymmetric nouns that share the same word w_2. We call these “pivot triples”. Of these 12664 pivot triples only 263 are transitive, that is, they do not violate the triangle inequality and (w_1, w_3) exists in the

[11] As a control, we also extract bi-directional associations: the cue word produce a certain target and this specific target produces back the initial cue. For example, given the cue *ball*, one of the produced associates is the target *baseball*. Conversely, *baseball* elicits *ball* as target. For these 6232 pairs, in 50.09% of cases, the target is denser than the cue.

[12] For this experiment, we used FASTTEXT, as countries cannot be used in context with this model and not all the countries were included in the BERT “raw” word embeddings.

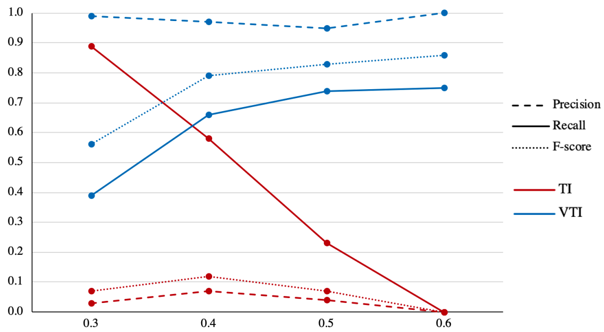

Figure 3: Precision, recall and F-score for transitive and intransitive triples. TI indicates tuples that respect the triangle inequality, and VTI refers to tuples that violate the triangle inequality. The x-axis indicates the different values of τ and the y-axis the score. Calculations are done considering Nelson's norms as the gold and BERT word embeddings as the system output.

human associations as a cue-target pair.[13] To calculate the BERT word embeddings, we contextualise each pair in the three different contexts already shown for previous experiments. To determine if the two pairs of words violate the triangle inequality in BERT space, we follow a procedure similar to Griffiths et al. (2007)'s. We set a threshold τ and extract those (w_1, w_2) and (w_2, w_3) whose cosine is greater than τ. Then, for each of these pivot triples, we quantify how many (w_1, w_3) pairs exist in the associations as a cue-target pair. If the pair exists and (w_1, w_3) is greater than τ, then BERT's embeddings also show transitivity; if the pair exists, but (w_1, w_3) is less than τ, then (w_1, w_3) in BERT does not reflect the transitivity found in humans; if the pair does not exist and (w_1, w_3) is lesser than τ, then there is a lack of transitivity both in human associations and in BERT embeddings.

Results Figure 3 shows some comparisons, in terms of precision and recall, of the BERT spaces against the gold human associations for different values of τ. The results show that the BERT space has many more triples for which there is triangle inequality (transitive triples), especially for low values of τ. This is expected, as BERT is trained on a much larger vocabulary space. The really interesting and relevant results are the measures of recall of the violations of triangle inequality (recall of VTI, blue solid line in the figure). As the results show, for increasing values of τ and more and more stringent definitions of similarity, the agreement of BERT with the human norms on violations of triangle inequality is high, and this despite a clear tendency to overestimate triangle inequalities (transitivity), as the TI values show.

7 Discussion

We have found converging evidence for BERT being like human word associations in ranks of associations, quantitatively and qualitatively. In studying whether BERT similarity spaces are asymmetric, we find converging evidence to human experiments, using country names, both in the fact that the notion of salience influences the calculation of similarity and also in the fact that frequency correlates with the preferential direction of similarity. However, using a larger test set, we do not find convergence with frequency, or generality (hyperym/hyponym pairs of verbs and nouns) or neighbourhood density. Finally, human word associations violate the triangle inequality. So do BERT embedding spaces, for reasonably high values of the similarity measure.

In conclusion, we confirm the properties of rank association and triangle inequality and also the influence of frequency for certain kinds of associations. We find, instead, that the property of asymmetric similarity does not appear to conform to the operationalisations we tested.

8 Related work

Recent interest in vectorial representations of words derives from the realisation that the meaning of words is much better represented when the rich networks of similarities and dissimilarities among words are taken into account. But the realisation that such notions are central to our word representations, that they can be estimated from a corpus and that such representations can be very technologically apt is not new.

Church and Hanks (1990) brought to the attention of the computational linguistic community a notion of word association as the information-theoretic measure of mutual information estimated on large corpora. It was shown that if word order was taken into account the measure could be asymmetric. Mutual information and other word association measures have been intensely studied to describe multi-word expressions or collocations, a stumbling block for many NLP applications (for a recent survey, see Constant et al. (2017)). Levy and Goldberg (2014) show that word embeddings are closely related to information-theoretic notions

[13] Samples of the word lists are shown in the supplementary materials.

of mutual information, although denser and better performing in many tasks.

Our work builds on previous work in the non-associationist tradition: Word associations are a reflex of underlying properties and representations of words and their meaning (Clark, 1970; Tversky, 1977; Griffiths et al., 2007; De Deyne et al., 2016) and not the reverse. Tversky and colleagues's body of work is centred in a *contrast model*, a model of objects, concepts and words based on features and not on a position in a multi-dimensional space. In this model, the similarity between objects increases with shared common features and decreases with distinctive features (Tversky, 1977; Tversky and Gati, 1978). From this point of view, computation of similarity is based on set-theoretic operations, rather than the computation of metric distances (Tversky, 1977). In a probabilistic topic model, Griffiths et al. (2007), words are a set of probability distributions on topics so that words that have a high probability under the same topic will be highly predictive of one another. The representations induced by the topic model and their correspondence to human memory are compared to a spatial representation model (Latent Semantic Analysis, LSA) (Landauer and Dumais, 1997) and found to better reflect human association norms.

Current vector space semantic representations can be seen as inheriting from both the feature-based tradition and the similarity space tradition, exemplified by LSA. While initial similarity-space proposals like LSA represent words as atomic and occupying a single point in space, current geometric approaches represent words as vectors. The approaches proposed by Mikolov et al. (2013a,b); Bojanowski et al. (2017); Devlin et al. (2018) are based on a distributed representation of words, and aim to produce vectors that represent a word, or the substrings that compose a word, with information about its surroundings, so that word vectors that share the same meaning tend to be close. A recent comparison of word embeddings, Word2vec and Glove, to Nelson's norms indicates that vectorial representations that do not take context into account, unlike BERT, still are unable to capture the triangle inequality (Nematzadeh et al., 2017).

Investigations of word associations also belongs to the growing literature of evaluating vector spaces for natural language applications. Word associations are an interesting, intrinsic way to evaluate vector spaces (Vulić et al., 2017; Thawani et al., 2019), and have revealed important properties of these spaces, from gender stereotypes and demographic variation (Du et al., 2019; Garimella et al., 2017) to their usefulness in the detection of puns (Sevgili et al., 2017), among many others.

9 Conclusions

The work described in this paper starts from the assumption that word associations are the expression of underlying meaning properties of words. It confirms that context-aware word embeddings exhibit some properties of human association norms, despite being a vectorial representation of words in space. Future work needs to clarify the underlying mechanisms that give rise to these properties, extend the study to new languages, leveraging also newer association norms (De Deyne et al., 2019). It will also extend the investigation of word associations to other properties, such as the minimal contrast rule —associations tend to establish a minimal contrast— or the marking rule —marked cues elicit unmarked targets more often than the reverse (Clark, 1970), and model documented differences between adults and children.

Acknowledgments

We are very grateful to Suzanne Stevenson and the anonymous reviewers for useful suggestions.

References

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

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Herbert H Clark. 1970. Word associations and linguistic theory. *New horizons in linguistics*, 1:271–286.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

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.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870, Osaka, Japan. The COLING 2016 Organizing Committee.

Simon De Deyne and Gert Storms. 2014. Word associations. In John R. Taylor, editor, *The Oxford Handbook of the Word*. Oxford University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171—4186, Minneapolis, Minnesota, USA.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring Human Gender Stereotypes with Word Association Test. 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 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

Itamar Gati and Amos Tversky. 1982. Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):325–340.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211–244.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

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

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.

Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. 2017. Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 859–854, London, UK.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Özge Sevgili, Nima Ghotbi, and Selma Tekir. 2017. N-hance at SemEval-2017 task 7: A Computational Approach using Word Association for Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 436–439, Vancouver, Canada. Association for 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. Hillsdale, Erlbaum*, pages 79–98.

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

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by Association: A Systematic Study of Quantitative Word Association Evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175, Valencia, Spain. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich 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 *Proceedings of the IEEE international conference on computer vision*, pages 19–27.