Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition

Shima Asaadi
Technische Universität Dresden, Germany
shima.asaadi@tu-dresden.de

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Svetlana Kiritchenko
National Research Council Canada
svetlana.kiritchenko@nrc-cnrc.gc.ca

Abstract

Bigrams (two-word sequences) hold a special place in semantic composition research since they are the smallest unit formed by composing words. A semantic relatedness dataset that includes bigrams will thus be useful in the development of automatic methods of semantic composition. However, existing relatedness datasets only include pairs of unigrams (single words). Further, existing datasets were created using rating scales and thus suffer from limitations such as inconsistent annotations and scale region bias. In this paper, we describe how we created a large, fine-grained, bigram relatedness dataset (*BiRD*), using a comparative annotation technique called *Best–Worst Scaling*. Each of BiRD's 3,345 English term pairs involves at least one bigram. We show that the relatedness scores obtained are highly reliable (split-half reliability $r = 0.937$). We analyze the data to obtain insights into bigram semantic relatedness. Finally, we present benchmark experiments on using the relatedness dataset as a testbed to evaluate simple unsupervised measures of semantic composition. BiRD is made freely available to foster further research on how meaning can be represented and how meaning can be composed.

1 Introduction

The term *semantic relatedness* refers to the extent to which two concepts are close in meaning. The ability to assess semantic relatedness is central to the use and understanding of language (Hutchison, 2003; Mohammad and Hirst, 2005; Huth et al., 2016). Manual ratings of semantic relatedness are useful for: (a) obtaining insights into how humans perceive and use language; and (b) developing and evaluating automatic natural language systems.

Existing datasets of semantic relatedness, such as the one by Finkelstein et al. (2002), only focus on pairs of unigrams (single words). However, the concept of semantic relatedness applies more generally to any unit of text. Work in semantic representation explores how best to represent the meanings of words, phrases, and sentences. Bigrams (two-word sequences) are especially important there since they are the smallest unit formed by composing words. Thus it would be useful to have large semantic relatedness datasets involving bigrams.

Existing datasets also suffer from shortcomings due to the annotation schemes employed. Except in the case of a few small but influential datasets, such as those by Miller and Charles (1991) and Rubenstein and Goodenough (1965), annotations were obtained using rating scales. (Annotators were asked to give scores for each pair; usually on a discrete 0 to 5 scale.) Rating scales suffer from significant known limitations, including: inconsistencies in annotations by different annotators, inconsistencies in annotations by the same annotator, scale region bias (annotators often have a bias towards a portion of the scale), and problems associated with a fixed granularity (Presser and Schuman, 1996).

Best–Worst Scaling (BWS) is an annotation scheme that addresses these limitations by employing comparative annotations (Louviere, 1991; Cohen, 2003; Louviere et al., 2015; Kiritchenko and Mohammad, 2017). Annotators are given n items at a time (an n-tuple, where $n > 1$ and commonly $n = 4$). They are asked which item is the *best* (highest in terms of the property of interest) and which is the *worst* (least in terms of the property of interest).[1] When

[1] At its limit, when $n = 2$, BWS becomes a *paired comparison* (Thurstone, 1927; David, 1963), but then a much larger set of tuples need to be annotated (closer to N^2).

working on 4-tuples, best–worst annotations are particularly efficient because each best and worst annotation will reveal the order of five of the six items (i.e., for a 4-tuple with items A, B, C, and D, if A is the best, and D is the worst, then A > B, A > C, A > D, B > D, and C > D). It has been empirically shown that annotating $2N$ 4-tuples is sufficient for obtaining reliable scores (where N is the number of items) (Louviere, 1991; Kiritchenko and Mohammad, 2016). Kiritchenko and Mohammad (2017) showed through empirical experiments that BWS produces more reliable and more discriminating scores than those obtained using rating scales.[2]

In this paper, we describe how we obtained fine-grained human ratings of semantic relatedness for English term pairs involving at least one bigram.[3] The other term in the pair is either another bigram or a unigram. We first selected a set of target bigrams AB (A represents the first word in the bigram and B represents the second word). For each AB, we created several pairs of the form AB–X, where X is a unigram or bigram. As X's we chose terms from a diverse set of language resources:

- terms that are transpose bigrams BA—where the first word is B and the second word is A (taken from occurrences in Wikipedia);
- terms that are related to AB by traditional semantic relations such as hypernymy, hyponymy, holonymy, meronymy, and synonymy (taken from WordNet); and
- terms that are co-aligned with AB in a parallel corpus (taken from a machine translation phrase table).

The dataset includes 3,345 term pairs corresponding to 410 ABs. We refer to this dataset as the *Bigram Relatedness Dataset* (or, *BiRD*).

We use BWS to obtain semantic relatedness by: (1) creating items that are pairs of terms, and (2) prompting four items (pairs) at a time and asking annotators to mark the pair that is most related and the pair that is least related. Once the annotations are complete, we obtain real-valued scores of semantic relatedness for each pair using simple arithmetic on the counts of how often an item is chosen best and worst (Orme, 2009; Flynn and Marley, 2014). (Details in Section 3.) To evaluate the quality of BiRD we determine the consistency of the BWS annotations. A commonly used approach to determine consistency in dimensional annotations is to calculate split-half reliability (Cronbach, 1951). We show that our semantic relatedness annotations have a split-half reliability score of $r = 0.937$, indicating high reliability, that is, if the annotations were repeated then similar scores and rankings would be obtained. (Details in Section 4.)

We use BiRD to (a) obtain insights into bigram semantic relatedness, and (b) to evaluate automatic semantic composition methods.

Examining Bigram Semantic Relatedness: Since very little work exists on the semantic relatedness of bigrams, several research questions remain unanswered, including: What is the distribution and mean of the semantic relatedness between a bigram and its transpose?; What is the average semantic relatedness between a bigram and its hypernym?; Are co-aligned terms from a phrase table a good source of term pairs to be included in a semantic relatedness dataset (specifically, do they cover a wide range of semantic relatedness values)?; etc. In Section 5, we present an analysis of BiRD to obtain insights into these questions.

Evaluating Semantic Composition: A common approach to evaluate different methods of representing words via vectors is through their ability to rank pairs of words by closeness in meaning (Pennington et al., 2014; Levy and Goldberg, 2014; Faruqui and Dyer, 2014). BiRD allows for the evaluation of semantic composition methods through their ability to rank pairs involving bigrams, by semantic relatedness. In Section 6, we present benchmark experiments on using BiRD as a testbed to evaluate various common semantic composition methods using various pre-trained word representations. Specifically, we conduct experiments to gain insights into research questions such as: Which common mathematical operations for vector composition (e.g., vector addition, vector multiplication, etc.) capture the semantics of a bigram more accurately?; Which of the two words in a noun phrase bigram (the head noun or

[2]See Kiritchenko and Mohammad (2016, 2017) for further details on BWS and its use in NLP applications.

[3]In a separate project, the second author is developing a semantic relatedness dataset for unigrams using BWS (an order of magnitude larger than existing ones). Project page: http://saifmohammad.com/WebPages/Relatedness.html

the modifier) has greater influence on the semantics of the bigram?; etc.

Contributions: The contributions of this work can be summarized as follows:

- We obtain fine-grained human ratings of semantic relatedness for 3,345 term pairs, each of which includes at least one bigram. The other term in the pair is either another bigram or a unigram.
- We use the comparative annotation technique Best–Worst Scaling, which addresses the limitations of traditional rating scales. This is the first time BWS has been used to create a dataset for semantic relatedness. We show that the ratings obtained are highly reliable.
- We analyse BiRD to obtain insights into semantic relatedness when it involves bigrams. We also develop interactive visualizations that allow for easy exploration of the data. (Available on the project webpage.)
- We present benchmark experiments on using BiRD as a testbed to evaluate methods of semantic composition.

The Bigram Relatedness Dataset, visualizations of the data, and the annotation questionnaire are made freely available through the project's webpage.[4] We hope that the new dataset will foster further research on how meaning is composed in bigrams, on semantic representation in general, and on the understanding of bigram semantic relatedness.

The annotation task described in this paper was approved by the National Research Council Canada's Research Ethics Board (protocol number 2018-72). The board examines the proposed methods to ensure that they adhere to the required ethical standards. Special attention was paid to obtaining informed consent and protecting participant anonymity.

2 Background and Related Work

Semantic Relatedness and Semantic Similarity Closeness of meaning can be of two kinds: semantic similarity and semantic relatedness. Two terms are considered to be semantically similar if there is a taxonomic relationship between them such as hyponymy (hypernymy), or troponymy. Two terms are considered to be semantically related if there is any lexical semantic relation between them—taxonomic or non-taxonomic. Semantically similar items tend to share a number of properties. For example, *apples* and *bananas* (co-hyponyms of fruit) are both edible, they grow on trees, they have seeds, etc. On the other hand, semantically related concepts may not have many properties in common, but there exists some relationship between them which lends them the property of being semantically close. For example, *surgeon* and *scalpel* are semantically related as the former uses the latter for their work.

We focus on semantic relatedness in this work, not only because it is the broader class subsuming semantic similarity, but also because many psychology and neuro-linguistic studies have demonstrated the importance of semantic relatedness. Notable among these are studies on semantic priming and fMRI studies that show that the human brain stores information in a thematic manner (based on relatedness) rather than based on similarity (Hutchison, 2003; Huth et al., 2016).

Word-Pair Datasets: Several semantic similarity and relatedness datasets involving unigram pairs (word pairs) exist. Rubenstein and Goodenough (1965) and Miller and Charles (1991) provided influential but small English word–pair datasets with fine–grained semantic similarity scores. More recent larger datasets including hundreds of pairs were provided by Finkelstein et al. (2002) (for relatedness) and Hill et al. (2015) (for similarity). Similar datasets exist in some other languages as well, such as the one by Gurevych (2006) and Panchenko et al. (2016) for relatedness. However, none of these datasets include items that are bigrams.

Bigram Semantic Similarity Datasets: Mitchell and Lapata (2010) created a semantic similarity dataset for 324 bigram pairs. The terms include adjective–noun, noun–noun, and verb–object bigrams. Annotators were asked to choose an integer between one and seven, indicating a coarse semantic similarity rating. Turney (2012) compiled a dataset of 2,180 bigram–unigram synonym pairs from WordNet synsets. (The bigrams are either noun–noun or adjective–noun phrases.) Other pairs were created taking bigrams and words that do not exist in the same synsets.

[4] http://saifmohammad.com/WebPages/BiRD.html

He thus created a dataset of synonyms and non-synonyms. In contrast to these datasets, BiRD has fine-grained relatedness scores.

Other Similarity Datasets: There exist datasets on the semantic similarity between sentences and between documents (Marelli et al., 2014; Agirre et al., 2014; Cera et al., 2017). Those are outside the scope of this work.

Other Natural Language Datasets Created Using BWS: BWS has been used for creating datasets for relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), word–sentiment intensity (Kiritchenko and Mohammad, 2016), word–emotion intensity (Mohammad, 2018b), and tweet–emotion intensity (Mohammad and Kiritchenko, 2018). The largest BWS dataset is the NRC Valence, Arousal, and Dominance Lexicon, which has valence, arousal, and dominance scores for over 20,000 English words (Mohammad, 2018a).

3 English Bigram Relatedness Dataset

We first describe how we selected the term pairs to include in the bigram relatedness dataset, followed by how they were annotated using BWS.

3.1 Term Pair Selection

Randomly selecting term pairs will result in most pairs being unrelated. This is sub-optimal in terms of the human annotation effort that is to follow. Further, since our goal is to create a gold standard relatedness dataset, we wanted it to include term pairs across the whole range of semantic relatedness: from maximally unrelated to maximally related. Thus, a key challenge in term-pair selection is obtaining pairs with a wide range of semantic relatedness scores, without knowing their true semantic relatedness in advance. In addition, we also wanted the dataset to satisfy the following criteria:

- For each target bigram AB we wanted to include several pairs of the form AB–X, where X is a unigram or bigram.
  Motivation: Applications of semantic relatedness, such as real-word spelling correction and textual entailment, often require judgments of the form '*is AB–X_1 more related or less related than AB–X_2*'.
- There should exist some pairs AB–X, such that X is BA and a common English bigram.
  Motivation: This is useful for testing sensitivity of semantic composition models to word order.
- The unigrams and bigrams should be commonly used English terms.
  Motivation: Data annotation of common terms is expected to be more reliable. Also, common terms are more likely to occur in application datasets.
- There should exist pairs that are taxonomically related (i.e., semantically similar), for example, hypernyms, hyponyms, holonyms, etc.; and there should exist pairs that are not taxonomically related but semantically related nonetheless.
  Motivation: This increases dataset diversity.
- We focus on noun phrases (adjective–noun and noun–noun bigrams).
  Motivation: Noun phrases are the most frequent phrases.

To pursue these criteria, we compiled a set of term pairs from three diverse sources (Wikipedia, WordNet, and a machine translation phrase table) as described below.

Wikipedia: We chose to collect our target bigrams from the English Wikipedia dump (2018).[5] The corpus was tagged with parts of speech (POS) using the NLTK toolbox.[6] For each of the adjective–noun and noun–noun bigrams AB in the corpus, we checked to see if the bigram BA (its transpose) also exists in the corpus. We will refer to such pairs of bigrams as *transpose bigrams*. Only those transpose bigrams (AB and BA) were selected that were both noun phrases and where both AB and BA occur in the corpus with frequencies greater than a pre-chosen threshold t (we chose $t = 30$). For a pair of transpose bigrams, the bigram with the higher frequency was chosen as AB and the bigram with the lower frequency was chosen as the corresponding BA. The above process resulted in 4,095 transpose pairs (AB–BA).

WordNet: Among the 4,095 ABs, 330 exist in WordNet version 3.0 (Fellbaum, 1998).[7] For each of these, we selected (when available) synonyms (at most five), a hypernym, a hyponym, a holonym, and a meronym from WordNet.

[5] https://dumps.wikimedia.org/
[6] https://www.nltk.org/
[7] https://wordnet.princeton.edu/download/current-version

Translation Phrase Table: Word-aligned parallel corpora map words in text of one language to those in text of another language. Often this can lead to more than one word/phrase in one language being mapped to a common word/phrase in the other language. We will refer to such terms as being *co-aligned*. Due to the nature of languages and the various forms that the same text can be translated to, co-aligned terms tend to include not just synonyms but also other semantically related terms, and sometimes even unrelated terms. Thus, we hypothesize that it is beneficial to include pairs of co-aligned terms in a semantic relatedness dataset as they pertain to varying degrees of semantic relatedness.

We used an English–French phrase table from the Portage Machine Translation Toolkit (Larkin et al., 2010) to determine additional pairs AB–X.[8] Specifically, for each AB–F entry in the phrase table (where F is a French term) we keep the five most frequent English unigrams and the five most frequent English bigrams (other than AB) that are also aligned to F. Among the 4,095 ABs, 454 occurred in the phrase table. This resulted in 3,255 AB–X pairs in total (1,897 where X is a unigram, and 1,358 where X is a bigram).

Finally, we chose to filter the term pairs, keeping only those ABs that occurred in at least three unique pairs. (So for a given AB, apart from the AB–BA entry, there should be at least two other entries of the form AB–X, generated using WordNet or the phrase table.) We also manually examined the remaining entries and removed those with obscure terms. The final master term pairs list consists of 3,345 AB–X pairs in total (1,718 where X is a unigram, and 1,627 where X is a bigram), corresponding to 410 ABs. Thus on average, each AB occurred in about 8 distinct pairs. This is yet another aspect that makes BiRD unique, as existing datasets were not designed to include terms in multiple pairs. Table 1 shows the number of adjective–noun pairs, the number of noun–noun pairs, and the total number of pairs in BiRD. (We grouped the hypernym and hyponym pairs into a common class, which we will refer to as the *is-a* pairs. Similarly we group the meronym and holonym pairs into a common class, which we will refer to as the *part-whole* pairs.)

[8]French was chosen as it is close to English and there exist English–French parallel corpora of sufficient size.

Source	# a–n	# n–n	# both
Wikipedia_transpose	80	330	410
WordNet_synonym	18	70	88
WordNet_is-a	49	220	269
WordNet_part-whole	7	30	37
PhraseTable_co-aligned	440	2,101	2,541
All	**594**	**2,751**	**3,345**

Table 1: Number of pairs from different sources.

3.2 Annotating For Semantic Relatedness

As mentioned in the introduction, we use the comparative annotation method Best–Worst Scaling (BWS) to obtain the annotations. From the list of $N = 3,345$ term pairs, we generated $2N = 6,690$ distinct 4-tuples (each 4-tuple is a set of four term pairs) such that each term pair appears in roughly equal distinct tuples, and no term pair appears more than once in a tuple.[9] (Recall that past research has shown that generating 2N 4-tuples in this manner is sufficient for obtaining fairly reliable scores (Louviere, 1991; Kiritchenko and Mohammad, 2017; Mohammad, 2018a).) The annotators were presented with one tuple at a time and were asked to specify which of the four pairs is most close in meaning (or most related) and which term is the least close (or least related).

Detailed annotation instructions (with examples of appropriate and inappropriate responses) were provided. Notably, we made it clear that if terms in the pair have several meanings, then the annotators should consider the meanings that are closest to each other. We also asked the annotators to be mindful of word order (i.e., the meaning of a bigram AB may be different from the meaning of its transpose BA).

We set up the annotation task on the crowdsourcing platform, Figure Eight.[10] We did not collect personally identifiable information from the annotators. The compensation that the annotators would receive was clearly stated. We selected a pool of annotators fluent in English and with a history of high-quality annotations. Annotators were told that they could annotate as many instances as they wished. As mentioned in the Introduction, prior to the annotation, the planned procedure was approved by the National Research Council Canada's Research Ethics Board (protocol number 2018-72).

[9]If 2N 4-tuples are generated from N items, and each item is to occur in an equal number of tuples, then each item will occur in eight tuples.

[10]https://www.figure-eight.com/

# Term Pairs	# Tuples	# Annotations per Tuple	# Annotations	# Annotators	SHR
3,345	6,690	8 (for most tuples), >8 (for some)	57,482	427	0.9374

Table 2: BiRD annotation statistics. SHR = split-half reliability (as measured by Pearson correlation).

About 2% of the data was annotated beforehand by the authors. These questions are referred to as gold questions. Figure Eight interspersed the gold questions with the other questions. If a crowd worker answered a gold question incorrectly, then they were immediately notified. This served as an additional way to guide the annotators. If an annotator's accuracy on the gold questions fell below 70%, then they were refused further annotation, and all of their annotations were discarded. This served as a mechanism to avoid malicious annotations.

In the task settings for Figure Eight, we specified that we needed annotations from eight people for each 4-tuple.[11] In all, 57,482 pairs of best and worst responses were obtained from 427 annotators.[12]

Annotation Aggregation: The final semantic relatedness scores were calculated from the BWS responses using a simple counting procedure (Orme, 2009; Flynn and Marley, 2014): For each term pair, the semantic relatedness score is the proportion of times the term pair was chosen as the best minus the proportion of times the term pair was chosen as the worst.[13] The scores were linearly transformed to the interval: 0 (lowest semantic relatedness) to 1 (highest semantic relatedness). We refer to the final list of 3,345 English term pairs along with their scores for semantic relatedness as the *Bigram Relatedness Dataset (BiRD)*. Table 2 summarizes key annotation statistics.

4 Reliability of Data Annotations

A commonly used measure of quality in dimensional annotation tasks is the reproducibility of the final scores—the extent to which repeated independent manual annotations produce similar results. To assess this reproducibility, we calculate average *split-half reliability (SHR)* (Cronbach, 1951) as follows: The annotations for each 4-tuple are randomly split into two halves. One set is put in bin 1 and another set in bin 2. Next, two sets of semantic relatedness scores are produced independently from the two bins, 1 and 2, respectively. Then the Pearson correlation between the two sets of scores is calculated. If the annotations are of good quality, then the correlation between the two sets of relatedness scores will be high (closer to 1).[14] This process is repeated 100 times, and the correlations are averaged. The last column in Table 2 shows the result. An SHR of $r = 0.9374$ indicates high reliability.

5 Studying Bigram Semantic Relatedness

Since very little prior work exists on the semantic relatedness of bigrams, several research questions remain unanswered, including:

- If both AB and BA are common English bigrams, then what is the average semantic relatedness between AB and BA?
- What is the range of semantic relatedness between a bigram and its hypernym or hyponym? What is the average semantic relatedness of such pairs? How do these averages and standard deviations vary with respect to the different semantic relations?
- What is the distribution of semantic relatedness values for co-aligned terms?

We now present analyses of the relatedness dataset to obtain insights into these questions.

Figure 1 shows example adjective–noun and noun–noun entries from BiRD. Observe that for the term *ageing population*, the most related term is *ageing society*—a co-aligned term in the phrase table. (Other co-aligned terms have lower relatedness scores.) The transpose bigram *population ageing* is also marked as highly related. WordNet does not provide a synonym for *ageing population*. For the term *adult female*, the WordNet synonym and the transposed bigram (BA) are marked as being most related. Note that the WordNet-provided hyponym *amazon* is marked as less related (probably because that sense of *amazon* is rare). BiRD can be examined

[11] Note that since each term pair occurs in eight different 4-tuples, it is involved in $8 \times 8 = 64$ best–worst judgments.

[12] Gold questions were annotated more than eight times.

[13] More complex optimization algorithms exist, such as those described in (Hollis, 2018); however, our past experiments showed that the simple counting procedure obtained the most reliable results.

[14] Scores close to 0 indicate no correlation.

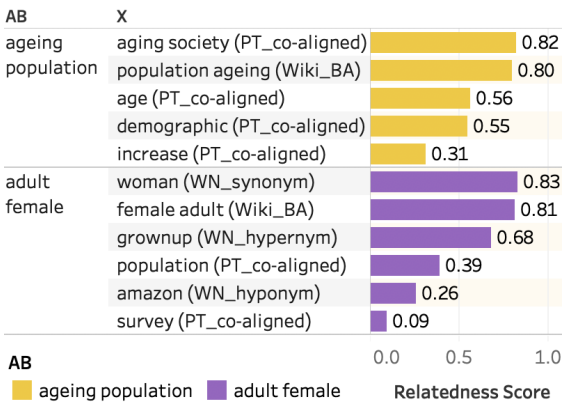

Figure 1: Example entries from BiRD.

for each individual relation and sorted by relatedness scores to determine other example pairs that seemingly should be closely related, but are not highly semantically related in the perception of the average English speaker. These include pairs such as *subject area–discipline* (WordNet synonym) and *frying pan–spider* (WordNet hyponym). The AB–BA pairs with low relatedness, such as *law school–school law, home run–run home,* and *traffic light–light traffic* are especially useful in testing whether measures of semantic composition generate suitably different representations for the terms in such pairs.

Table 3 shows the average semantic relatedness scores as well as standard deviations for the term pairs from various sources.[15] Observe that, on average, the AB–BA pairs and the AB–WordNet synonym pairs are found to be the most related. On average, the AB–WordNet part-whole pairs and the AB–phrase table co-aligned pairs have the lowest semantic relatedness scores. The high average relatedness and low standard deviation (σ) for the transpose bigrams, indicate that these pairs tend to be closely related to each other. The standard deviation is markedly higher for the other sources of word pairs. Manual examination of such pairs (especially those involving WordNet synonyms) revealed that this is often because one of the terms might be related to the other in a rare sense (such as in the *amazon* example). The high standard deviations for hypernyms, hyponyms, meronyms, and holonyms, indicate that pairs connected by this relation in WordNet can still exhibit a wide range of semantic relatedness.

The standard deviations also indicate that 95% of the co-aligned pairs have semantic relatedness between 0.09 and 0.83 (a wide interval). Manual examination revealed that the lowest score pairs were unrelated and the highest score terms were often synonymous. Thus co-aligned pairs from phrase tables are indeed a good source of term pairs for a semantic relatedness dataset, since they include pairs with a wide variety of relatedness values.

[15]The scores for just the adjective–noun pairs and just the noun—noun pairs are similar.

Source	avg. rel.	σ
Wikipedia_transpose	0.669	0.118
WordNet_synonym	0.640	0.194
WordNet_is-a	0.550	0.177
WordNet_part-whole	0.453	0.193
PhraseTable_co-aligned	0.463	0.189

Table 3: Average relatedness and standard deviation (σ) scores for term pairs from the various sources.

6 Evaluating Methods of Semantic Composition on BiRD

A popular approach to represent word meaning in natural language systems is through vectors that capture the contexts in which the word occurs. An area of active research is how these word vectors can be composed to create representations for larger units of text such as phrases and sentences (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Socher et al., 2012; Tai et al., 2015). Even though there is a large body of work on how to represent the meanings of sentences (Le and Mikolov, 2014; Kiros et al., 2015; Lin et al., 2017), there is relatively less work on how best to compose the meanings of two words to represent the meaning of a bigram. One reason for this is a lack of suitable evaluation resources. A common approach to evaluate representations of unigrams is through their ability to rank pairs of words by closeness in meaning (Pennington et al., 2014; Levy and Goldberg, 2014; Faruqui and Dyer, 2014). BiRD allows for the evaluation of semantic composition methods through their ability to rank pairs involving bigrams, by semantic relatedness.

Here, we present benchmark experiments on commonly used semantic composition methods by measuring their ability to rank the term pairs in BiRD by relatedness scores. The underlying assumption is that the more accurately a method of semantic composition can determine the representation of a bigram, the more accurately systems can determine the relatedness of that bigram with other terms.

We focus on unsupervised approaches as we wanted to identify how well basic composition operations perform. The applicability of BiRD is much broader though, and it can be used: (1) for evaluating the large number of proposed supervised methods of semantic composition; (2) for evaluating the large number of measures of semantic relatedness; (3) to study the mechanisms underpinning semantic composition; etc. We leave those for future work.

We test three vector space models to obtain word representations: GloVe (Pennington et al., 2014), fastText (Grave et al., 2018), and a traditional model based on matrix factorization of a word–context co-occurrence matrix (Turney et al., 2011). We test four mathematical composition operations: (1) vector addition, (2) element-wise vector multiplication, (3) tensor product with circular convolution (Widdows, 2008), and (4) dilation (Mitchell and Lapata, 2010). In adjective–noun and noun–noun bigrams, the second word usually plays a role of a head noun, and the first word is a modifier. We test the performance of two baseline methods that do not employ vector composition: one that represents a bigram with the vector for the first word and one that represents a bigram with the vector for the second word.

Word representations: We use GloVe word embeddings pre-trained on 840B-token CommonCrawl corpus[16] and fastText word embeddings pre-trained on Common Crawl and Wikipedia using CBOW.[17] For the traditional model, we use the exact word–context co-occurrence matrix described in Turney et al. (2011).[18] They created the matrix from a corpus of 5×10^{10} tokens gathered from university websites. The rows correspond to terms (single words from WordNet) and the columns correspond to contexts (single words from WordNet appearing to the left or to the right of the term). Each cell of the matrix is the positive pointwise mutual information between the term and the context. The matrix is decomposed to $\mathbf{U_d \Sigma_d V_d^\top}$ (d denotes dimensionality) via truncated singular value decomposition. Word vectors are obtained from the matrix $\mathbf{U_d \Sigma_d^p}$, where rows correspond to the d-dimensional word vectors and p is the weight factor for singular values in $\mathbf{\Sigma_d}$. We set parameter p to 0.5, and the dimensionality of word vectors to $d = 300$ for all three vector space models.

Unsupervised Compositional Models: For a bigram $w_1 w_2$, let $u \in \mathbb{R}^{1\times d}$ and $v \in \mathbb{R}^{1\times d}$ denote the vectors for words w_1 and w_2, respectively. Each of the methods below applies a different composition function f on the word vectors u and v to obtain the vector representation p for the bigram $w_1 w_2$: $p = f(u, v)$:

- *Addition* (Salton and McGill, 1986): add the two word vectors ($p = u + v$).
- *Multiplication* (Mitchell and Lapata, 2010): element-wise multiplication of the two vectors ($p = u \odot v$, where $p_i = u_i \cdot v_i$).
- *Tensor product with convolution* (Widdows, 2008): outer product of two vectors resulting in matrix Q ($q_{ij} = u_i v_j$). Then, circular convolution is applied to map Q to vector p. This is equivalent to: $p_i = \sum_j u_j \cdot v_{i-j}$.
- *Dilation* (Mitchell and Lapata, 2010): decompose v to parallel and orthogonal components to u, and then stretch the parallel component along u ($p_i = v_i \sum_j u_j u_j + (\lambda - 1) u_i \sum_j u_j v_j$, where λ is the dilation factor). We set $\lambda = 2$.

For the two baseline experiments that do not employ vector composition, *head only*: $p = v$ and *modifier only*: $p = u$.

Semantic Relatedness: The relatedness score for a term pair AB–X in the Bigram Relatedness Dataset (BiRD) is computed by taking the cosine between the vectors representing AB and X, where X can be a unigram or a bigram.

Evaluation: As evaluation metric, we use the Pearson correlation of the relatedness scores predicted by a method with the gold relatedness scores in BiRD. Some words in BiRD do not occur in some of the corpora used to create the word vectors. Thus we conduct experiments on a subset of BiRD (3,159 pairs) for which word vectors exist for all models under consideration. To determine if the differences between the correlation scores are statistically significant, we perform Steiger's Z significance test (Steiger, 1980).

[16] https://nlp.stanford.edu/projects/glove/
[17] https://fasttext.cc/docs/en/crawl-vectors.html
[18] We thank Peter Turney for providing the data.

Method	GloVe	fastText	Matrix Factor.
Baselines			
head only	0.342	0.403	0.339
modifier only	0.438	0.495	0.425
Composition methods			
addition	**0.564**	**0.601**	**0.582**
multiplication	0.182	0.328	0.244
tensor product	0.374	0.382	0.451
dilation	0.523	0.495	0.496

Table 4: Pearson correlations of model predictions with BiRD relatedness ratings. Highest scores are in bold.

Results: Table 4 shows the results. Observe that among the three methods of word vector representations, the best results are obtained using fastText (word–context matrix factorization model being a close second). Among the methods of semantic composition, the additive models perform best (for all three ways of representing word vectors). The scores are statistically significantly higher than those of the second best (dilation). The element-wise vector multiplication and tensor product with convolution perform poorly (even worse than the baseline methods). These results differ substantially from the observations by Mitchell and Lapata (2010). In particular, in their work the multiplication model showed the best results, markedly outperforming the addition model. Our results are consistent with the findings of Turney (2012), where too the addition model performed better than the multiplication model. It should be noted though that unlike BiRD which has scores for semantic relatedness, the Mitchell and Lapata (2010) and Turney (2012) datasets have scores for semantic similarity. Further work is required to determine whether certain composition models are better suited for estimating one or the other.

Surprisingly, the baseline model that uses the vector for the modifier word obtains better results than the one that uses the vector for the head noun. (The difference is statistically significant.) To better understand this, we compute relatedness correlations using the *weighted addition* of the two word vectors ($p = \alpha u + (1 - \alpha)v$), where α is a parameter that we vary between 0 and 1, in steps of 0.1. Figure 2 shows the results. Observe that giving more weight (but not too much weight) to the modifier word than the head word is beneficial. $\alpha = 0.7$ and $\alpha = 0.8$ produce the highest correlations. These results raise further questions under what conditions is the role of the modifier particularly prominent, and why. We leave that for future work.

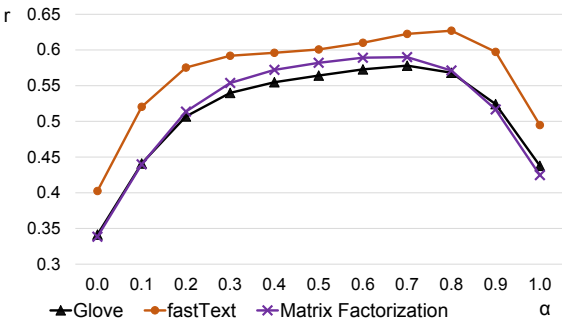

Figure 2: Pearson correlation coefficient (r) of the model predictions using weighted addition with BiRD relatedness ratings. α is varied from 0 to 1 in steps of 0.1. $\alpha = 0.7$ and $\alpha = 0.8$ produce the highest scores.

7 Conclusions

We created a dataset with fine-grained human ratings of semantic relatedness for term pairs involving bigrams. We used the comparative annotation technique Best–Worst Scaling, which addresses the limitations of traditional rating scales. We showed that the ratings obtained are highly reliable (high SHR, $r = 0.937$). We analyzed the dataset to obtain insights into the distributions of semantic relatedness values for pairs associated through various relations such as WordNet assigned lexical semantic relations, transposed bigrams, and co-aligned terms in a parallel corpus. We show that co-aligned terms can be related to varying degrees (from unrelated to synonymous), thereby making them a useful source of term pairs to include in relatedness datasets. Finally, we presented benchmark experiments on using BiRD as a testbed to evaluate various unsupervised methods of semantic composition. We found that the additive models performed best and that giving more weight to the modifier word can improve results further. We make BiRD freely available to foster further research. In the short term, it will be interesting to explore the use of supervised semantic composition methods, including resources and models such as BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018), to determine bigram relatedness.

Acknowledgments

We thank Peter Turney, Michel Simard, and Tara Small for helpful discussions. This work is partially supported by the German Research Foundation (DFG) within the Research Training Group QuantLA (GRK 1763).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193.

Daniel Cera, Mona Diab, Eneko Agirrec, Inigo Lopez-Gazpioc, Lucia Speciad, and Basque Country Donostia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, pages 1–14.

Steven H. Cohen. 2003. Maximum difference scaling: Improved measures of importance and preference for segmentation. Sawtooth Software, Inc.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Herbert Aron David. 1963. *The method of paired comparisons*. Hafner Publishing Company, New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.

T. N. Flynn and A. A. J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, pages 178–201. Edward Elgar Publishing.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Iryna Gurevych. 2006. Thinking beyond the nouns-computing semantic relatedness across parts of speech. Technical report, Darmstadt University of Technology, Germany, Department of Computer Science, Telecooperation.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Geoff Hollis. 2018. Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, 50(2):711–729.

Keith A Hutchison. 2003. Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic Bulletin & Review*, 10(4):785–813.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Atlanta, GA, USA.

David Jurgens, Saif M. Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval)*, pages 356–364, Montréal, Canada.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pages 3294–3302.

Samuel Larkin, Boxing Chen, George Foster, Uli Germann, Eric Joanis, J. Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage System at WMT 2010. In *Proceedings of the 5th Workshop on Statistical Machine Translation (WMT-2010)*, pages 127–132.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*, pages 1188–1196.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pages 2177–2185.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the International Conference on Learning Representations*.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. *arXiv:1203.1858. 2005*.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB.

Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia Loukachevitch, and Chris Biemann. 2016. Human and machine judgements for Russian semantic relatedness. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*, pages 221–235.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Stanley Presser and Howard Schuman. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Gerard Salton and Michael J. McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*, volume 26.