CARD-660: Cambridge Rare Word Dataset – a Reliable Benchmark for Infrequent Word Representation Models

Mohammad Taher Pilehvar **Dimitri Kartsaklis** **Victor Prokhorov** **Nigel Collier**
Language Technology Lab, Department of Theoretical and Applied Linguistics
University of Cambridge, United Kingdom
{mp792,dk426,vp361,nhc30}@cam.ac.uk

Abstract

Rare word representation has recently enjoyed a surge of interest, owing to the crucial role that effective handling of infrequent words can play in accurate semantic understanding. However, there is a paucity of reliable benchmarks for evaluation and comparison of these techniques. We show in this paper that the only existing benchmark (the Stanford Rare Word dataset) suffers from low-confidence annotations and limited vocabulary; hence, it does not constitute a solid comparison framework. In order to fill this evaluation gap, we propose CAmbridge Rare word Dataset (CARD-660), an expert-annotated word similarity dataset which provides a highly reliable, yet challenging, benchmark for rare word representation techniques. Through a set of experiments we show that even the best mainstream word embeddings, with millions of words in their vocabularies, are unable to achieve performances higher than 0.43 (Pearson correlation) on the dataset, compared to a human-level upperbound of 0.90. We release the dataset and the annotation materials at https://pilehvar.github.io/card-660/.

1 Introduction

Words in a corpus of natural language utterances approximately follow a Zipfian distribution with their majority, in the "long tail" of frequency distribution, occurring rarely. The prominent distributional approach to semantic representation relies on enormous occurrences for each individual word; therefore, it falls short of learning accurate representations for rare words in the long tail. Moreover, it is unreasonable to expect that all words in the vocabulary of a language are observed in a text corpus, even if it is massive in size. Out-of-vocabulary (OOV) words pose one of the major ongoing challenges for word embedding techniques. Given that effective handling of rare and OOV words is crucial to accurate natural language understanding, several studies have focused on the topic during the past few years, resulting in a wide range of techniques.

However, despite the popularity of rare and subword semantic representation, the field of research has suffered from the lack of high quality generic evaluation benchmarks. A task-based evaluation, i.e., one which directly verifies the impact of representation models in a downstream NLP system, despite being very important, does not provide a solid base for comparing different models, given that small variations in the architecture, parameter setting, or initialisation can lead to performance differences. Moreover, such an evaluation would reflect the "suitability" of representations for that specific configuration and for that particular task, and might not be conclusive for other settings.

As far as generic evaluation is concerned, existing benchmarks generally target frequent words. An exception is the Stanford Rare Word (RW) Similarity dataset (Luong et al., 2013) which has been the standard evaluation benchmark for rare word representation techniques for the past few years. In Section 2.1, we will provide an in-depth analysis of RW and highlight that crowdsourcing the annotations, with no rigorous checkpoints, has compromised the reliability of the dataset. This is mainly reflected by the low inter-annotator agreement (IAA), a performance ceiling which is easily surpassed by many existing models.

To overcome this barrier and to fill the gap for a reliable benchmark for the evaluation of subword and rare word representation techniques, we introduce a new dataset, called CARD-660: Cambridge Rare Word Dataset. Compared to existing benchmarks, CARD-660 provides multiple advantages: (1) thanks to a manual curation by experts, we report IAA of around 0.90 (see Table 3) which is substantially higher than those for exist-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1391–1401
Brussels, Belgium, October 31 - November 4, 2018.

ing datasets; (2) word pairs are selected manually from a wide range of domains and, unlike existing datasets, are not bound to a specific resource; (3) word pairs in the dataset are balanced across the similarity scale; and (4) the huge gap between state of the art and IAA (more than 0.50 in terms of Spearman correlation) promises a challenging dataset with lots of potential for future research.

The paper is structured as follows. The following Section covers the related work, highlighting some of the issues with the RW dataset. Section 3 details the construction procedure for CARD-660. In Section 4, we analyse the dataset from different aspects, showing how it improves existing benchmarks. Section 5 reports our evaluation of mainstream word embeddings and recent word representation techniques on the dataset. Finally, concluding remarks are mentioned in Section 6.

2 Related Work

Word similarity datasets have been one of the oldest, still most prominent, benchmarks for the evaluation and comparison of semantic representation techniques. As a result, several word similarity datasets have been constructed during the past few decades; to name a few: RG-65 (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2002), YP-130 (Yang and Powers, 2005), MEN-3K (Bruni et al., 2014), SimLex-999 (Hill et al., 2015), and SimVerb-3500 (Gerz et al., 2016). Many of these English word similarity datasets have been translated to other languages to create frameworks for multilingual (Leviant and Reichart, 2015) or crosslingual (Camacho-Collados et al., 2017) semantic representation techniques. However, these datasets mostly target words that occur frequently in generic texts and, as a result, are not suitable for the evaluation of subword or rare word representation models.

One may opt for transforming a frequent-word benchmarks into an artificial rare word dataset by downsampling the dataset's words in the underlying training corpus (Sergienya and Schütze, 2015). However, this benchmark might not properly simulate a real-world rare word representation scenario (cf. Section 3.1).

2.1 Stanford RW Dataset

The Stanford Rare Word Similarity (RW) dataset is an exception as it is dedicated to evaluating infrequent word representations. The dataset has been regarded as the de facto standard evaluation benchmark for subword and rare word representation techniques. However, our analysis shows that RW suffers from multiple issues: (1) skewed distribution of the scores, (2) low-quality and inconsistent scores, and as a consequence, (3) low inter-annotator agreement.

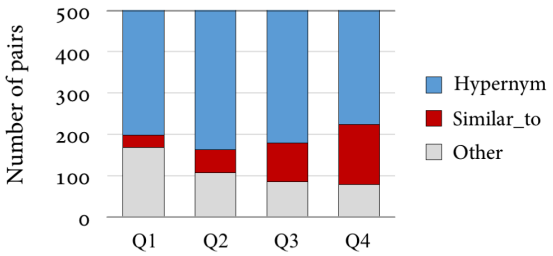

Figure 1: Distribution of relation types ("hypernymy", "similar_to", and others) across four quartiles (sorted by gold similarity scores) of the Stanford Rare Word Similarity dataset. The distribution of pairs with hypernymy relation is almost uniform across the quartiles, whereas one would expect many more pairs in the top quartiles (Q4 and Q3), given the high semantic similarity of hypernym-hyponyms.

The RW dataset comprises 2034 word pairs (i.e., $word_1$ – $word_2$). Candidates for $word_1$ were randomly sampled from Wikipedia documents, distributed across a wide range of frequencies (from 5 to 10,000) to ensure the inclusion of infrequent words. Given this automatic sampling, a measure was required to avoid noisy or junk words. To this end, a sampled word was checked in WordNet (Fellbaum, 1998) and was included only if it appeared in at least one synset. Hence, the vocabulary of the dataset is bound to that of WordNet. Words for $word_2$ were randomly picked from synsets that were directly connected to a synset of $word_1$, through various relations, such as hypernymy, holonymy, and attributes.

2.1.1 Distribution of scores

These word pairs were assigned similarity scores in $[0, 10]$. Given that all word pairs in the dataset are semantically-related according to WordNet, the scores form a skewed distribution biased towards the upper bound (see Figure 2 and Section 4.1 for more details).

2.1.2 Consistency of annotations

The scoring of the pairs has been carried out through crowdsourcing: (Amazon Mechanical) Turkers have provided ten scores for each word pair. The raters were restricted to only US-based

workers and they were asked to self-certify themselves by indicating if they "knew" the word; this was used to "discard unreliable pairs." However, our analysis of the dataset clearly indicates that the above measures have not been adequate for guaranteeing quality annotations. For instance, the word *bluejacket* is paired with *submariner* in the dataset. According to WordNet (v3.0), a submariner ("a member of the crew of a submarine") is a bluejacket ("a serviceman in the navy"; a *navy_man*, *sailor*), hence a hypernymy relationship. One would expect a word to have high semantic similarity with its hypernym. However, the gold score for this pair is just 0.43 in the scale $[0, 10]$. Other examples include "untruth" (a false statement) vs. "statement" (again, with a hypernymy relationship) with a low similarity of 1.22. Apart from not being a rigorous evaluation, the self-certification does not verify if the annotator had knowledge of *various* possible meanings of a word. For instance, *decomposition* could refer to the analysis of vectors in algebra; but, when paired with *algebra*, the assigned score is only 0.75. Such examples clearly indicate that the annotators were not aware of specialised senses of some words (e.g., the algebraic meaning of decomposition), despite "knowing" the word.

In fact, there are numerous such pairs in the dataset. According to our estimate, 78% of the 2034 word pairs in the dataset are in a *hypernymy* or *similar_to* relationship. One would expect most of these (semantically similar) pairs to have been assigned high similarity scores which are closer to the upper bound of the similarity scale $[0, 10]$. However, as shown in Figure 1, these pairs are spread across the similarity scale, spanning from complete unrelatedness (lower bound) to identical semantics (synonymy). Having the words in the dataset sorted by their assigned gold scores, respectively, 66%, 79%, 83%, and 85% of the pairs in the first to fourth quartiles contain either "hypernymy" or "similar_to" relations (whereas one would expect most of these semantically-similar pairs to appear in the top quartiles).

Additionally, the dataset suffers from inconsistent annotations. For instance, the two almost identical pairs *tricolour-flag* and *tricolor-flag* were assigned substantially different scores, i.e., 5.80 and 0.71, respectively. This inconsistency is also reflected by high variances across annotators scores (cf. Section 4.3).

2.1.3 Inter-Annotator Agreement (IAA)

This validity metric reflects the homogeneity of the annotators' ratings and it is generally accepted as the upper bound for machine performance. IAA is widely used as a standard evaluation metric for the quality of word similarity datasets. A low IAA indicates a defective similarity scale or unreliable annotations.

In the RW dataset, "up to 10" annotations have been provided for the 2034 word pairs, each with a similarity score in $[0, 10]$ range. More precisely, 214 of the pairs are not provided with 10 scores, with the minimum number of scores for a pair being 7. The authors did not report IAA statistics for this dataset. Given that the annotators are not known for each pair in the released dataset, it is not straightforward to compute IAA.[1] According to a rough calculation, the average pairwise Spearman correlation between annotators' scores is 0.40, which is a significantly low figure compared to other existing word similarity datasets. We report an impressive IAA of 0.89 for our dataset (cf. Section 4.2).

3 The CARD-660 Dataset

3.1 Motivation

Due to a lack of reliable evaluation benchmarks, research in rare word representation has often resorted to artificial experimental setups such as corpus downsampling (Sergienya and Schütze, 2015; Herbelot and Baroni, 2017; Lazaridou et al., 2017). To this end, in order to simulate a rare word scenario, the rare word representation model is provided with only a limited number of occurrences for the target set of words, for instance by means of replacing the dataset's words with some other sequences of characters (e.g., by augmenting "UNK", such as "skyglowUNK" for "skyglow") in the training corpus. The computed representations on the "downsampled" training data are then either evaluated on a standard word similarity dataset (Sergienya and Schütze, 2015), such as RG-65, or compared against reference embeddings computed on a large training corpus (Herbelot and Baroni, 2017; Lazaridou et al., 2017).

However, due to the following three reasons, downsampling does not constitute a reliable

[1] The scores are further pruned down to only those that were within one standard deviation of the mean. This results in a further imbalanced set of scores, making the computation of IAA more challenging.

benchmark that can represent the challenging nature of the task: (1) it is unable to control the impact of second-order associations (words that frequently co-occur with the downsampled word) and cannot represent a real-world setting with novel rare usages; (2) given that morphological variations of a word (such as plural forms) are kept intact in this procedure, a subword technique can easily resort to these forms to compute the embedding for downsampled words; and (3) a constrained evaluation configuration in which the task is to estimate the embedding for a (rare) word using one or few occurrences (contexts) of it, limits the benchmark to a subset of corpus-based rare word representation techniques only. Moreover, the evaluation would require the comparison of the computed embeddings for rare words with a set of reference embeddings (computed on the full data). This dependency limits the ability of the benchmark in providing a direct evaluation of the rare word representation technique, independently from the impact of the model used to compute the reference embeddings.

The CARD-660 dataset aims at filling the gap for rigorous generic evaluation of rare word and subword representation models. In what follows in this section, we will detail the construction procedure of the dataset which was carefully planned to guarantee a challenging and reliable dataset.

3.2 Construction Procedure

The following four-phase procedure was used to construct the dataset:

(1) A set of 660 rare words were carefully selected from a wide range of domains;
(2) For each of these initial words, a pairing word was manually selected according to a randomly sampled score from the similarity scale (Section 3.2.2);
(3) All pairs were scored by 8 annotators;
(4) A final adjudication was performed to address disagreements (Section 3.2.3).

3.2.1 Similarity scale

We adopted the five-point Likert scale used for the annotation of the datasets in SemEval-2017 Task 2 (Camacho-Collados et al., 2017). The task reported high IAA scores which reflects the well-definedness and clarity of the scale. We provided annotators with the concise guideline shown in Table 1, along with several examples. Given the continuity of the scale, the annotators were given flexibility to select values in between the five points, whenever appropriate, with a step size of 0.5.

The annotators were asked in the guidelines to make sure they were familiar with *all* common meanings of the word (as defined by WordNet or other online dictionaries). To facilitate the annotation, the annotators were provided with the definitions of some of the words that were defined in WordNet or named entities that had Wikipedia pages. For others, we asked the annotators to check the word in online dictionaries, such as WordNet browser[3] and Wiktionary[4], or encyclopediae, such as Wikipedia.

3.2.2 Word pair selection

Unlike previous work (Luong et al., 2013), we did not rely on random sampling (pruned by frequency) of initial words from a specific dictionary, to prevent the dataset from being restricted to a specific resource or vocabulary. Instead, we carefully hand-picked word pairs from a wide range of domains. To construct the 660 pairs of the dataset (each pair is denoted as $w_1 - w_2$), we first picked 660 w_1 words. Our aim was to have a dataset that can ideally reflect the performance of rare word representation techniques in downstream NLP tasks. To this end, we picked initial words (w_1s) from different common NLP datasets and resources, listed in Table 2. For each text-based resource, a frequency list was obtained and rare words were carefully picked from the long tail of the list, cross-checking the frequency of words in the Google News dataset. For the other resources (such as Wiktionary), we checked a word against a large frequency list to ensure they are not frequent words. The list was computed on the 2.8B token ukWaC+WaCkypedia corpus (Baroni et al., 2009) and comprised 16.5M unique words.

In order to have a balanced distribution of scores in the dataset, we first assigned random integer scores in $[0-4]$ to the 660 initial w_1s. Then, with the corresponding score in mind, a pairing word (w_2) was selected for each w_1. We show in Section 4.1 that this strategy resulted in a uniformly distributed set of scores in the dataset.

The dataset comprises words from a wide range of genres and domains, including slang in so-

[2] http://www.fakenewschallenge.org

[3] http://wordnetweb.princeton.edu/perl/webwn

[4] www.wiktionary.org

Score	Interpretation	Example pair
4	**Synonyms.** The two words are different ways of referring to the same concept	*car* *automobile*
3	**Similar.** The two words are of the same nature, but slightly different in details	*car* *truck*
2	**Related.** The two words are closely related but they are not similar in their nature	*car* *driver*
1	**Same domain or slight relation.** The two words have distant relationship	*car* *tarmac*
0	**Completely unrelated.** The two words have nothing in common.	*car* *sky*

Table 1: The five-point Likert similarity scale used for the annotation of the dataset.

Task. Resource
Text classification. BBC (Greene and Cunningham, 2006)
Sentiment analysis. IMDB (Maas et al., 2011), Multi-Domain Sentiment Dataset (Blitzer et al., 2007)
Machine Translation. Europarl (Koehn, 2005)
Question Answering. AQUA-RAT (Ling et al., 2017), SQuAD (Rajpurkar et al., 2016)
BioMedical (entity recognition). JNLPBA corpus (Kim et al., 2004)
Social media. Twitter
Ontologies and online glossaries. WordNet, Wiktionary
Named entities. Freebase (Bollacker et al., 2008)
Veracity assessment. FakeNews[2]

Table 2: Various datasets and resources used for rare word selection in CARD-660.

cial media (e.g., *2mrw* and *Mnhttn*), named entities (e.g., *Stephen_Hawking* and *Ursa_Major*), and domain specific terms (e.g., *erythroleukemia* and *NetMeeting*). Moreover, to have a rigorous testbed for subword representation techniques that emphasises the importance of semantic (rather than shallow) understanding of the words, the dataset contains several word pairs that have similar surface forms (hence, high string similarity) while being semantically distant, e.g., *infection-inflection* and *currency-concurrency*. There are also many compound words (e.g., *skyglow*, *musclebike*, and *logboat*) which makes the dataset particularly interesting for evaluating compositionality as well as for subword representation techniques.

3.2.3 Scoring and adjudication

Once the 660 word pairs were manually selected (by the first author), the initial scores were discarded and the words were shuffled (vertically and horizontally) to dispense any potential bias from the initial round of creation. Then, the pairs were assigned to 8 annotators (including all but first authors) who independently scored each and every pair according to the annotation guidelines (see Section 3.2.1). All annotators were PhD graduates or students in Computational Linguistics or related fields and were either native or fluent English speakers.

Once all pairs were scored by the annotators, we checked for disagreements. This check was intended to improve the dataset's quality through resolving simple annotation mistakes. For each annotator, we marked the i^{th} pair if for the assigned score s_i: $s_i \geq \mu_i + 1$ or $s_i \leq \mu_i - 1$, where μ_i is the average of the other seven annotators' scores for s_i. The annotator was then asked to (more carefully) re-score the marked pair by checking for its possible meanings. They were asked to keep their initial score if not convinced otherwise. The adjudication revealed that most disagreements were due to an annotator having misread a word or not been familiar with a specific meaning of it, or missing annotations. By average, 13.8% of the pairs were re-scored by each annotator.

4 Analysis

In this section we provide an analysis on the quality of CARD-660 from three different perspectives: distribution of scores, inter-annotator agreement, and consistency among annotators. We benchmark CARD-660 against the Stanford **RW** dataset and two standard word similarity datasets (cf. Section 2): SimVerb-3500 (**SV-3500**) and SimLex-999 (**SL-999**). The latter two datasets do not target rare words; however, given that their construction strategy is similar to that employed for creating RW (based on crowdsourcing), we included them in our analysis experiments to provide better insights. For the purpose of this evaluation, all the datasets were scaled to [0,10] to make them comparable.

4.1 Score Distribution

Figure 2 shows the distribution of pairs across the similarity scale, for CARD-660 and the three other datasets. As discussed in Section 2.1, RW is heavily biased towards the upper bound of the similar-

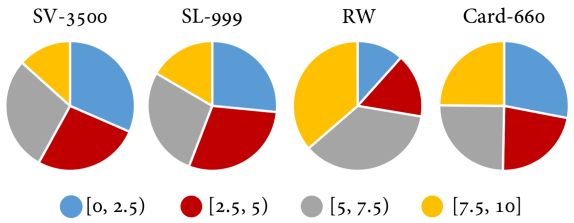

Figure 2: The distribution of word pairs across the four quartiles of the similarity scale for different datasets. A perfectly balanced dataset would have four equally sized slices.

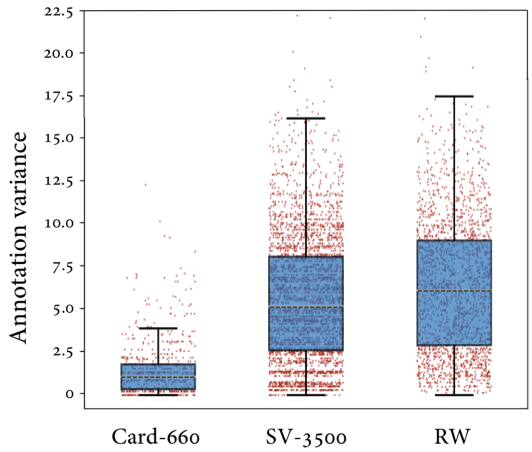

Figure 3: Annotation variance for word pairs across different datasets. Average variance for CARD-660 is 1.47, which is significantly lower than those for SV-3500 and RW: 5.64 and 6.34, respectively.

ity scale (with around 72% of the pairs in the upper half, i.e., $[5, 10]$). The skewed distribution in this dataset can be attributed to the automatic word pair selection from semantically-related words in WordNet (cf. Section 2.1). SV-3500 and SL-999 are skewed towards the lower bound, but to a smaller degree (around 59% of the pairs in $[0, 5)$). Thanks to the manual creation of CARD-660, we have a balanced set of pairs across the similarity scale (50-50% across the two halves).

4.2 Inter-Annotator Agreement

As mentioned in Section 2.1, IAA has been extensively used as a quality metric for word similarity datasets. Following standard practise, we measure two sets of IAA scores: (1) **Pairwise** is the averaged pairwise correlation between all possible rater pairings, and (2) **Mean** is the averaged correlation of each rater against the average of others.

Table 3 reports IAA statistics for CARD-660. Thanks to the manual scoring of the pairs by experts (as opposed to turkers), the IAA values for

	Mean		**Pairwise**	
	r	ρ	r	ρ
Initial	88.0±2.3	87.9±1.9	80.2±2.9	80.6±2.6
Final	93.5±1.4	93.1±1.2	88.9±1.7	88.9±1.7

Table 3: Inter-annotator agreement (IAA) scores before (*initial*) and after (*final*) adjudication (± standard deviation). IAA is shown in terms of Pearson r and Spearman ρ percentage correlations. The *final* scores are representative of the dataset's quality.

the dataset are very high, placing it among the best word similarity datasets in the literature. This is particularly interesting considering that, compared to standard word similarity datasets which contain mostly common words, our dataset comprises words that are semantically difficult to annotate due to their rare nature. The pairwise IAA score of 88.9 is significantly higher than the crowdsourced RW, with the estimated pairwise IAA score of around 40.0 (cf. Section 2.1). The same applies to other recent crowdsourced word similarity datasets for common words which usually report pairwise IAA scores below 70.0 (e.g., $\rho = 67.0$ for SL-999)[5].

4.3 Consistency of Annotations

Despite being suitable for measuring linear relationships between scores, correlation cannot fully reflect the consistency between annotators. Two annotators can have perfect correlation, i.e., 1.0, even if they consistently provide different scores for the same pairs (therefore, having different average assigned scores). To check the consistency among annotators, i.e., if they had the same interpretation of the similarity scale, we compute variance across annotators for individual pairs.

The box and whisker (over scatter) plot in Figure 3 shows the distribution of annotator variances for the pairs in different datasets. Clearly, the score variances for CARD-660 are significantly lower than those for the two crowdsourced datasets, i.e., SimVerb-3500 and RW.[6] Specifically, for the majority of pairs in CARD-660 the annotation variance is lower than the other two datasets' first quartile (bottom of the blue square

[5] SimVerb-3500 reports a pairwise ρ of 84.0; however, our calculation did not agree with this figure. Personal communication with the authors revealed an issue in the computation of their IAA. The correct figure is instead 61.2.

[6] We are not able to report results for SimLex-999 since individual annotators' scores are not released for this dataset.

Embedding set	$\|V\|$	Missed words		Missed pairs		Pearson r		Spearman ρ	
		RW	CARD	RW	CARD	RW	CARD	RW	CARD
Glove Wikipedia-Gigaword (300d)	400K	7%	55%	12%	74%	34.9	15.1	34.4	15.7
Glove Common Crawl - uncased (300d)	1.9M	1%	36%	1%	50%	36.5	29.2	37.7	27.6
Glove Common Crawl - cased (300d)	2.2M	1%	29%	2%	44%	44.0	33.0	45.1	27.3
Glove Twitter (200d)	1.2M	29%	60%	48%	79%	17.7	13.7	15.3	11.7
Word2vec GoogleNews (300d)	3M	6%	48%	10%	75%	43.8	13.5	45.3	7.4
Word2vec Freebase (1000d)	1.4M	100%	85%	100%	92%	0.0	17.3	0.0	4.6
Dependency-based Wikipedia (300d)	174K	22%	60%	36%	80%	17.4	6.4	19.7	3.3
LexVec Common Crawl (300d)	2M	1%	41%	1%	55%	47.1	25.9	48.8	18.5
LexVec Wikipedia-NewsCrawl (300d)	370K	8%	58%	14%	78%	35.6	11.8	34.8	7.8
ConceptNet Numberbatch (300d)	417K	5%	37%	10%	53%	53.0	36.0	53.7	24.7
ConceptNet + Word2vec Freebase	1.6M	1%	22%	2%	45%	44.0	42.6	45.1	31.3
Glove cased CC + Word2vec Freebase	3.4M	11%	21%	10%	39%	53.0	38.8	53.7	32.7

Table 4: Pearson r and Spearman ρ correlation percentage performance of mainstream pre-trained word embeddings on the RW and CARD-660 datasets. Column $|V|$ shows the size of vocabulary for the corresponding embedding set.

which splits the lower 25% of the data from the top 75%). This indicates that our annotators had significantly higher degrees of agreement, reflecting the well-definedness of the similarity scale as well as the reliability of expert-based annotation (as opposed to crowdsourcing).

5 Evaluations

In the remainder of this paper, we provide two sets of experiments to showcase the challenging nature of our dataset. Specifically, in Section 5.1 we report the performance of common pretrained word embeddings on CARD-660, and in Section 5.2 we provide experimental results for state-of-the-art rare word representation techniques. In all experiments, we used the cosine similarity for comparing pairs of word embeddings.

5.1 Pre-trained Embeddings

As was mentioned earlier in the Introduction, it is not possible to enumerate the entire vocabulary of a natural language, even if massive corpora are used. A challenging rare word benchmark should ideally reflect this phenomenon. To verify this in our dataset, we experimented with a set of commonly used word embeddings trained on corpora with billions of tokens.

Table 4 provides correlation performance results for different embedding sets on the RW and CARD-660 datasets. Specifically, we considered different variants of Word2vec[7] (Mikolov et al., 2013) and Glove[8] (Pennington et al., 2014), two commonly-used word embeddings that are trained on massively large text corpora; Dependency-based embeddings[9] (Levy and Goldberg, 2014) which extends the Skip-gram model to handle dependency-based contexts; LexVec[10] (Salle et al., 2016) which improves the Skip-gram model to better handle frequent words; and ConceptNet Numberbatch[11] (Speer et al., 2017) which exploits lexical knowledge from multiple resources, such as Wiktionary and WordNet, and was the best performing system in SemEval 2017 Task 2. In the last two rows of the Table we also report results for two hybrid embeddings constructed by combining the pre-trained Freebase Word2vec, which mostly comprises named entities, with two of the best performing embeddings evaluated on the dataset. Given that the word embeddings are not comparable across two different spaces, we only compute the similarity between a pair only if both words are covered in the same space (with priority given to the non-Freebase embedding).

As can be seen in the Table, many of the embeddings yield high coverage for the RW dataset, with those trained on the Common Crawl (CC) corpus providing near full coverage. This highlights the limited vocabulary of the dataset (which is bound to that of WordNet). Also, many of

[7] https://code.google.com/archive/p/word2vec/

[8] https://nlp.stanford.edu/projects/glove/

[9] http://u.cs.biu.ac.il/~yogo/data/syntemb/deps.words.bz2

[10] https://github.com/alexandres/lexvec

[11] https://github.com/commonsense/conceptnet-numberbatch

the embeddings attain performance around 40.0 on RW, which is higher than the estimated IAA of the dataset. In contrast, CARD-660 proves to be significantly more challenging, with the highest coverage model (Glove CC and Word2vec Freebase hybrid model) missing around 40% of the pairs. Also, the best performance of 42.6 (r) and 32.7 (ρ) are substantially (around 50.0) lower than the IAA for the dataset (see Table 3).

5.2 Rare Word Representation Techniques

Rare and unseen word representation has been an active field of research during the past few years, with many different techniques proposed. In this experiment, we evaluate the performance of some of recent models on our dataset. These techniques can be broadly classified into two categories. The first group exploits the knowledge encoded for a rare word in external lexical resources (Section 5.2.1), whereas the second induces embeddings for rare words by extending the semantics of its subword units (Section 5.2.2).

5.2.1 Resource-based models

The basic assumption here is that a lexical resource, such as dictionary, provides high coverage for words in a language, even if they are rare. Resource-based models usually rely on WordNet as their external resource and estimate the embedding for a rare word by exploiting different types of lexical knowledge encoded for it in the resource. The **definition centroid** model of Lazaridou et al. (2017) takes WordNet word glosses (definitions) as semantic clue. An embedding is induced for an unseen word by averaging the content words' embeddings in its definition.[12] The **definition LSTM** strategy of Bahdanau et al. (2017) extends the centroid model by encoding the definition using an LSTM network (Hochreiter and Schmidhuber, 1997), in order to better capture the semantics and word order in the definition. **SemLand** (Pilehvar and Collier, 2017) also uses WordNet, but takes a different approach which benefits from the graph structure of WordNet. For an unseen word, SemLand extracts the set of its semantically related words from WordNet and induces an embedding for the unseen word by combining pre-trained embeddings for the related words.

[12]The original model is multimodal (text and images). Given that our focus is on texts, we follow Herbelot and Baroni (2017) and use the text modality only.

5.2.2 Subword models

Resource-based models fall short of inducing embeddings for words that are not covered in the lexical resource. Subword models alleviate this limitation by breaking the word into its subword (Pinter et al., 2017; Bojanowski et al., 2017) or morphological units (Luong et al., 2013; Botha and Blunsom, 2014; Soricut and Och, 2015) and induce an embedding by composing the information available for these. **FastText** (Bojanowski et al., 2017) is one of the popular approaches of this type. The model first splits the unseen word into character ngrams (by default, 3- to 6-grams) and then computes the unseen word's embedding as the centroid of the embeddings of these character n-grams (which are available as a result of a specific training). We also report results for **Mimick** (Pinter et al., 2017), one of the most recent subword models. The technique learns a mapping function from strings to embeddings by training a Bi-LSTM network that encodes character sequences of a word to its pre-trained embedding.

5.3 Experimental Setup

We report results for the five techniques discussed in Sections 5.2.2 and 5.2.1. We used two of the best performing embedding sets, i.e., Glove cased CC and ConceptNet Numberbatch, to train the models (except FastText for which we use the pre-trained WikiNews subword embeddings[13]). In fact, the models were expected to provide improvements over these baseline embeddings by filling their gaps for unseen words.

Mimick was trained with the default parameters,[14] except for the hidden units which we set to 100, instead of the original 50, since the target embeddings in our experiments were larger (300d compared to 128d of the original model). For the Definition LSTM model, the input definitions were represented as sequences of 50d word embeddings, encoded using an LSTM layer of 100 units, and then passed to a dense layer with 300 neurons with linear activation function. The training was carried out with Mean Squared Error loss and the RMSprop optimizer, for 100 epochs with batch size 64.

[13]https://fasttext.cc/docs/en/english-vectors.html
[14]https://github.com/yuvalpinter/Mimick/

Model	Missed words		Missed pairs		Pearson r		Spearman ρ	
	RW	CARD	RW	CARD	RW	CARD	RW	CARD
ConceptNet Numberbatch (300d)	5%	37%	10%	53%	53.0	36.0	53.7	24.7
+ Mimick (Pinter et al., 2017)	0%	0%	0%	0%	56.0	34.2	57.6	**35.6**
+ Definition centroid (Herbelot and Baroni, 2017)	0%	29%	0%	43%	59.1	42.9	60.3	33.8
+ Definition LSTM (Bahdanau et al., 2017)	0%	25%	0%	39%	58.6	41.8	59.4	31.7
+ SemLand (Pilehvar and Collier, 2017)	0%	29%	0%	43%	**60.5**	**43.4**	**61.7**	34.3
Glove Common Crawl - cased (300d)	1%	29%	2%	44%	44.0	33.0	45.1	27.3
+ Mimick (Pinter et al., 2017)	0%	0%	0%	0%	**44.7**	23.9	45.6	29.5
+ Definition centroid (Herbelot and Baroni, 2017)	0%	21%	0%	35%	43.5	35.2	45.1	31.7
+ Definition LSTM (Bahdanau et al., 2017)	0%	20%	0%	33%	24.0	23.0	22.9	19.6
+ SemLand (Pilehvar and Collier, 2017)	0%	21%	0%	35%	44.3	**39.5**	**45.8**	**33.8**
FastText (Bojanowski et al., 2017)	0%	3%	0%	5%	46.3	19.0	48.2	20.4

Table 5: Correlation performance of different rare and unseen word representation techniques on the Stanford RW and CARD-660 datasets (the best performance in each batch shown in bold; the overall best underlined).

5.4 Experimental Results

Table 5 reports the performance of different rare word representation techniques. Both pre-trained embeddings outperform the IAA of RW, with Glove covering 98% of the pairs. This severely limits the room for further meaningful experiments on the dataset. In contrast, on CARD-660 and similarly to the previous experiment, there are substantial gaps between IAA (cf. Table 3) and the best-performing models: SemLand and Mimick, with the respective figures of 45.5 (r) and 53.3 (ρ). These gaps suggest a difficult dataset which can serve future research in subword and rare word representation as a reliable benchmark.

The definition centroid model proves effective, despite its simplicity, whereas the WordNet-based SemLand provides the best results in most of the settings. Being constrained to the vocabulary of WordNet, the RW dataset does not constitute a challenging benchmark for WordNet-based models, with most of them providing near full coverage. However, these techniques are not as effective on our dataset, with the best WordNet-based model still missing around 33% of the pairs (with Glove pre-trained embeddings).

The CARD-660 dataset also proves a very difficult benchmark for subword models. Despite providing near full coverage, these models are unable to consistently improve the pre-trained word embedding baseline. This would suggest that the simple strategy of backing off to a word's characters might not always provide reliable means of estimating its semantics (e.g., the single-morpheme word *galaxy*, or the exocentric compound *honeymoon*). The results encourage further research on a more semantically-oriented handling of subwords, through learning more effective splitting and composition techniques.

6 Conclusions

Thanks to a carefully designed procedure and an expert-based curation, CARD-660 provides multiple advantages over existing benchmarks, including a very high IAA (average pairwise correlation of around 0.90). A series of experiments was carried out on the dataset, leading to two main conclusions: (1) the dataset proved a very challenging benchmark, with the best pre-trained embedding model still missing around 40% of the word pairs and the best rare word representation model hardly crossing into 40.0s (correlation performance); and (2) knowledge-based models are not enough to provide high coverage whereas subword models, which provide near-full coverage, are not semantically as effective. The significant gap between state of the art and IAA (around 50.0) encourages future research to take this dataset as a challenging, yet reliable, evaluation benchmark.

Acknowledgments

We gratefully acknowledge the funding support of EPSRC (N. Collier and D. Kartsaklis - Grant No. EP/M005089/1) and MRC (M. T. Pilehvar) Grant No. MR/M025160/1 for PheneBank. We would also like to thank Andreas Chatzistergiou, Costanza Conforti, Gamal Crichton, Milan Gritta, and Ehsan Shareghi for their contribution in creating the dataset.

References

Dzmitry Bahdanau, Tom Bosc, Stanislaw Jastrzebski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *CoRR*, abs/1706.00286.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

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.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, Vancouver, Canada.

Jan A. Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of ICML*, pages 1899–1907, Beijing, China.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Lev Finkelstein, Gabrilovich Evgenly, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2002. Placing search in context: The concept revisited. *ACM Transactions of Information Systems*, 20(1):116–131.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, pages 70–75, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41(S4):677–705.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR*, abs/1508.00106.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113, Sofia, Bulgaria.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT*, pages 142–150, Portland, Oregon, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR*, Scottsdale, Arizona.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, Doha, Qatar.

Mohammad Taher Pilehvar and Nigel Collier. 2017. Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 388–393, Valencia, Spain.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword rnns. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Alexandre Salle, Aline Villavicencio, and Marco Idiart. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–424, Berlin, Germany. Association for Computational Linguistics.

Irina Sergienya and Hinrich Schütze. 2015. Learning better embeddings for rare words using distributional representations. In *Proceedings of EMNLP*, pages 280–285.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of NAACL-HLT*, pages 1627–1637, Denver, Colorado.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Dongqiang Yang and David M. W. Powers. 2005. Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the Twenty-eighth Australasian Conference on Computer Science*, volume 38, pages 315–322, Newcastle, Australia.