Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5802–5809
Marseille, 11–16 May 2020

MSD-1030: A Well-built Multi-Sense Evaluation Dataset for Sense Representation Models

Ting-Yu Yen[1], Yang-Yin Lee[1], Yow-Ting Shiue[1], Hen-Hsen Huang[2,3], Hsin-Hsi Chen[1,3]
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2]Department of Computer Science, National Chengchi University, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{tyyen,yylee}@nlg.csie.ntu.edu.tw, orina1123@gmail.com, hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

Sense embedding models handle polysemy by giving each distinct meaning of a word form a separate representation. They are considered improvements over word models, and their effectiveness is usually judged with benchmarks such as semantic similarity datasets. However, most of these datasets are not designed for evaluating sense embeddings. In this research, we show that there are at least six concerns about evaluating sense embeddings with existing benchmark datasets, including the large proportions of single-sense words and the unexpected inferior performance of several multi-sense models to their single-sense counterparts. These observations call into serious question whether evaluations based on these datasets can reflect the sense model's ability to capture different meanings. To address the issues, we propose the Multi-Sense Dataset (MSD-1030), which contains a high ratio of multi-sense word pairs. A series of analyses and experiments show that MSD-1030 serves as a more reliable benchmark for sense embeddings. The dataset is available at `http://nlg.csie.ntu.edu.tw/nlpresource/MSD-1030/`.

Keywords: semantics, evaluation methodologies, crowdsourcing

1. Introduction

Word embeddings, or distributed word representations, have attracted much attention in recent years. Previous studies show that word embedding models are capable of learning semantic and syntactic information from a large unannotated corpus (Mikolov et al., 2013; Pennington et al., 2014). However, one essential issue in word embeddings is that each word form is represented by only one vector. That is, multiple senses of a word form are indistinguishable, which is problematic for applications involving word ambiguity such as word sense disambiguation (WSD) and semantic relation identification.

Sense embedding models, in which each sense of a word form is represented by its own vector (Reisinger and Mooney, 2010; Huang et al., 2012; Jauhar et al., 2015; Bartunov et al., 2016; Lee and Chen, 2017; Lee et al., 2018), have been proposed to address the polysemy issue mentioned above. Camacho-Collados and Pilehvar (2018) provide an extensive review of previous studies in sense embeddings. More recently, pre-trained contextualized word representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) handle polysemy by assigning a vector representation conditioned on the specific context to every word in a sentence. Different from those approaches, sense embeddings can be grounded in an ontology such as WordNet (Miller, 1995) or BabelNet (Navigli and Ponzetto, 2012), which can support operations such as queries and making inference over the ontology or any connected knowledge base. On the other hand, sense embeddings can be trained with smaller amounts of data and further enhanced by external ontologies, making them extremely useful in low-resource scenarios where there are well-established knowledge bases.

Similar to word embeddings, sense embeddings can be evaluated intrinsically or extrinsically (task-based). Although the latter may more closely reflect how useful a model would be in practical applications, this kind of evaluation is usually time-consuming. Furthermore, unlike word embeddings, sense embeddings can not be directly adopted in an application without some sense selection process, so it would be hard to decouple the quality of sense embeddings and the performance of WSD. Thus, intrinsic evaluation benchmarks are still important for efficient model and parameter selection. Researchers commonly use word embedding benchmarks, including semantic similarity datasets (Bruni et al., 2014; Radinsky et al., 2011; Finkelstein et al., 2002; Luong et al., 2013), the contextual word similarity dataset (Huang et al., 2012), and synonym selection datasets (Turney, 2001; Landauer and Dumais, 1997; Jarmasz and Szpakowicz, 2004), to evaluate sense embeddings. In this work, we argue that these evaluation benchmarks do not provide a solid base for testing sense embeddings in a polysemy scenario. We examined eight datasets and found that every dataset contains a great number of single-sense words, where there is no ambiguity to resolve. Moreover, most of these benchmarks have a biased distribution of human-annotated scores, leading to concerns about reliability when evaluating sense embeddings.

To address this, we propose MSD-1030, a novel multi-sense dataset with 1,030 English word pairs designed to facilitate more reliable evaluations of sense embeddings. Compared to existing benchmarks, these are the characteristics of MSD-1030: (1) Most of the words in MSD-1030 are multi-sense words. (2) The distribution of human annotator scores is controlled across the dataset. (3) MSD-1030 does not contain phrases (e.g., short sleep, smooth surface) or rare words that may obscure the focus of sense embedding evaluations. (4) From the experimental results, MSD-1030 is more suitable for evaluating sense embeddings than existing datasets. (5) According to our error analysis, even state-of-the-art sense embedding models may not be able to assign appropriate similarity scores to some multi-sense word pairs in MSD-1030, which shows that there is still room for improvement in modern sense embedding models.

2. Previous Datasets

Many datasets have been constructed to evaluate the quality of lexical semantics models. These datasets can be

	MEN	MTurk	RW	WS353	ESL-50	TF-80	RD-300	SCWS	MSD
# single-sense words	438	335	2053	244	113	278	1197	812	210
Total # words	751	499	2951	437	224	395	1464	1713	1030
Ratio	58.3%	67.1%	70.0%	55.8%	50%	70%	82%	47.4%	20.4%

Table 1: Number of single-sense words and their ratio in datasets.

categorized into three types: semantic similarity, contextual word similarity, and synonym selection.

2.1 Semantic Similarity

The task of semantic similarity/relatedness[1] is the most common way to evaluate the quality of a word embedding model. A number of semantic similarity datasets have been proposed, including MEN (3,000 word pairs) (Bruni et al., 2014), MTurk (287 word pairs) (Radinsky et al., 2011), WordSim-353 (WS353, 353 word pairs) (Finkelstein et al., 2002), and Rare Words (RW, 2,034 word pairs) (Luong et al., 2013). In these datasets, each word pair has a score in terms of their similarity or relatedness. A higher score indicates a higher similarity/relatedness between the words in the pair. MEN, MTurk, and RW were constructed by crowdsourcing on the Amazon Mechanical Turk platform. The performance of a semantic model is typically measured by the Spearman or Pearson correlation between the human-rated scores and the scores given by the model.

2.2 Contextual Word Similarity

Huang et al. (2012) extended the semantic similarity task and constructed the Stanford's Contextual Word Similarities (SCWS) dataset, which consists of 2,003 word pairs (1,713 unique words) together with crowd-sourced semantic similarity scores. For each word in a given pair, its context and part-of-speech tag are given. The context is word sequences around the target word.

The Word-in-Context (WiC) dataset (Pilehvar et al., 2019) also aims at evaluating a model's ability to handle different senses. However, each question in this dataset asks whether a certain word has the same meaning in two given contexts, which is a binary classification task. Since this formulation is very different from that of word similarity, we do not include this dataset in our analyses in the next section.

2.3 Synonym Selection

The task of synonym selection is also adopted to evaluate word embedding models. Commonly used datasets include ESL-50 (English as a Second Language) (Turney, 2001), RD-300 (Reader's Digest Word Power Game) (Jarmasz and Szpakowicz, 2004), and TOEFL-80 (TF-80, Test of English as a Foreign Language) (Landauer and Dumais, 1997). The number following the name of each dataset indicates the number of questions. Each question is composed of a word as the question stem and four words as alternatives. The task is to select the word most similar to the question stem from the four alternatives. For instance, let *swear* be the question stem, and *vow*, *explain*, *think*, and *describe* be the alternatives. The correct answer to this question would be *vow*. The performance of a model is determined by the accuracy.

3. Concerns of Previous Datasets

The aforementioned datasets commonly used to evaluate word embeddings are also used by researchers to evaluate and compare sense embedding models. However, in the following subsections, we show a number of concerns of existing datasets that make them unsuitable for benchmarking sense-level models.

3.1 Large Portion of Single-sense Words

We examine the number of single-sense words in all the datasets. To identify single-sense words, we utilize Roget's 21st Century Thesaurus (Kipfer, 1993) (Roget), in which each word has one or more senses. Here the senses of a word are defined as the categories of synonyms listed for that word in Roget[2]. To determine whether a word is single-sense or multi-sense, we calculate the number of senses it has. According to Table 1, all the datasets contain more than 47% single-sense words; MTurk, RW, TF-80, and RD-300 contain more than 67% one-sense words. Due to the low proportion of multi-sense words in these benchmarks, sense embedding models that excel at handling polysemy cannot demonstrate their strengths. In contrast, our MSD-1030, which will be introduced in Section 4, has a much lower ratio of single-sense words.

Note that the WiC dataset does not have this problem because it only contains questions about ambiguous words. However, it does not require a model to be able to compare multiple meanings across different words.

3.2 Better Performance of Sense Embedding Models with One Sense

The most important advantage of sense embeddings over word embeddings in the semantic similarity task is that specific meanings of a word can be considered when we compute the similarity. Thus, we hypothesize that given a sense embedding model, the performance will decline if we utilize only the first sense vector of each word.

To verify this, we conduct an experiment with the GenSense (Lee et al., 2018)[3] and SenseRetro (Jauhar et al., 2015)[4] sense embedding models, both of which perform post-processing with external ontologies. In this experiment, both models are based on GloVe embeddings trained on a corpus consisting of 6 billion tokens[5]. The

[1] In this study we adopt the general sense of similarity, which includes both conceptual similarity (e.g., *car* and *truck*) and relatedness (e.g., *car* and *wheel*).

[2] Note that in Roget the synonymy relation is directed. Specifically, if a word a has n senses $S_1^a, \ldots, S_n^a$, then there may exist another word b such that $b \in S_i^a$ but $a \notin S_j^b \ \forall j$.

[3] https://github.com/y95847frank/GenSense

[4] https://github.com/sjauhar/SenseRetrofit

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

	MEN	MTurk	RW	WS353	ESL-50	TF-80	RD-300	SCWS
GenSense	67.6	64.1	33.8	50.5	**64.6**	**87.2**	69.4	54.8
GenSense-1	**69.0**	**65.0**	**35.0**	**51.9**	52.1	82.1	**70.6**	**57.6**
SenseRetro	46.9	43.3	24.0	27.3	**64.6**	**83.3**	**74.1**	49.7
SenseRetro-1	**51.3**	**44.9**	**27.4**	**28.7**	47.9	75.6	70.6	**51.6**

Table 2: Performance (all sense vectors v.s. only first sense vectors) of 50-dimensional GenSense and SenseRetro (Spearman correlation ($\rho \times 100$) for MEN, MTurk, RW, WS353, and SCWS; accuracy $\times$ 100 for ESL-50, TF-80, and RD-300). Roget is the external ontology for both sense embedding models.

GenSense w value	(***bank*$_0$, money**)	(*bank*$_0$, *shore*)	(*bank*$_1$, *money*)	(***bank*$_1$, shore**)
1.00	71.8	54.7	**55.3**	76.4
0.75	72.8	57.4	51.9	79.1
0.50	**75.0**	**59.1**	46.1	**81.6**

Table 3: Similarity scores of GenSense sense vectors when decreasing weight of original word vector.

w value	MEN	MTurk	RW	WS353	SCWS
1.00	**67.6**	**64.1**	**33.8**	**50.5**	**54.8**
0.75	67.3	63.4	33.6	49.6	54.6
0.50	65.8	61.6	32.6	46.7	54.7

Table 4: $\rho \times 100$ of GenSense (50-dimensional) when decreasing weight of original word vector.

dimension of the sense embeddings is set to 50. The external ontology for both models is Roget.

To measure the similarity between a pair of words with sense models, we adopt the MaxSim similarity metric (Reisinger and Mooney, 2010). For the semantic similarity datasets, we calculate the Spearman correlation between the human-rated scores and the MaxSim scores to measure the performance. Unless otherwise stated, MaxSim is always used throughout this paper[6]. In the testing phase of synonym selection, we also adopt MaxSim to compute scores between the question stem and the four alternatives in the answer sets. The alternative with the maximum score is selected as the model's answer.

To evaluate sense embeddings in the contextual word similarity task, we adopt MaxSimC (Reisinger and Mooney, 2010), which is the cosine similarity between the pair of senses that maximizes their similarities with their corresponding contexts. Similarly, the performance is measured by the Spearman correlation between the human-rated scores and the MaxSimC scores.

The results are shown in Table 2. Sense embedding models with '-1' indicate that for each word only the vector of the first sense (according to Roget) is used; the other sense vectors are ignored. Surprisingly, the performance of GenSense and SenseRetro is enhanced when the vectors of senses other than the first sense are discarded (GenSense-1 and SenseRetro-1) in the semantic similarity datasets and the contextual word similarity dataset. That is, for most word pairs, the human scores might be closer to the model scores computed based on only one sense per word than the maximum possible similarities across multiple senses. As this contradicts our hypothesis in the first paragraph of this subsection, we suspect that these datasets cannot evaluate multi-sense models effectively. Although for the synonym selection datasets, there is no such a contradiction in most of the experiments, we will discuss concerns about them in Sections 3.5 and 3.6.

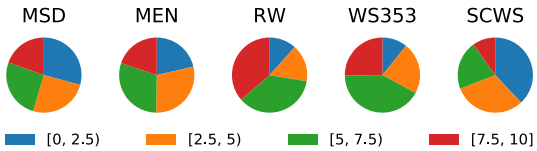

Figure 1: The score distribution of word pairs in four parts of the scale.

3.3 Performance of GenSense with Different Weights

In GenSense, parameter w controls the importance of the original word vectors in the retrofitting process. When decreasing w, the generated sense vectors become closer to their synonym neighbors in the external ontology than to their pre-trained word vector. We use Roget as the external ontology. Consider the word *bank*, whose senses include *bank*$_0$ (money) and *bank*$_1$ (shore). Table 3 shows the cosine similarities of several word pairs under three weights of GenSense. When the weight decreases (from $w = 1.0$ to $w = 0.5$), the cosine similarity score between *bank*$_0$ and *money* increases, and the similarity score between *bank*$_1$ and *shore* also increases. These results are reasonable because in Roget, *bank*$_0$ has synonyms *stock* and *treasury*, and *bank*$_1$ has synonyms like *coast* and *reef*.

We further examine how the separation of different senses affects the performance on four semantic similarity datasets and one contextual similarity dataset. The Spearman correlation results are shown in Table 4. Surprisingly, the performance drops significantly on every dataset as the weight of the original word vectors is decreased. That is, these existing benchmarks may favor dominating senses in the training corpus, making it unnecessary for the sense embedding models to deal with different senses separately. As a result, we cannot properly evaluate a sense embedding model's ability of handling multiple diverse senses using these datasets.

3.4 Skewed Score Distribution

To understand the distribution of the similarity score in the semantic similarity datasets and the SCWS dataset, we perform an analysis similar to that used by Pilehvar et al. (2018). We divide each dataset's score scale into four equal bins on the interval [1, 10] and assign the similarity scores

[6] Though other similarity metrics exist, we will explain in Section 4.3 that with MaxSim we can explicitly require both the model and human annotators to select the closest pair of senses, making the two judgments more comparable.

Determine similarity scale of two words based on the following criterion.

Score	Type	Interpretation	Example Pair	
4	Synonym	The two words are different ways of referring to the same concept of second word.	*decision*	*choice, resolution*
3	Similar	The two words are of the same nature, but slightly different in details.	*decision*	*preference, firmness*
2	Related	The two words are closely related to second word, but they are not similar in their nature.	*decision*	*adjudicature, purposiveness*
1	Same domain or slight related	The two words have distant relationship.	*decision*	*result*
0	Completely unrelated	The two words have nothing in common.	*decision*	*sky*

1) Some words have multiple senses. Thus, first read through the word definitions below, and then give your answer. For example, although "DELIVER" is usually used for carrying and bringing things, it could also be a synonym of "EXPRESS". Therefore, the similarity score of "DELIVER" and "EXPRESS" should be 4.

2) These HITs will be quality controlled. If you are unsure whether or not you are doing them well, we recommend you do a small number of questions and wait for the feedback before continuing. Thank you in advance!

Figure 2: Annotation guidelines.

to their corresponding bins. The results in Figure 1 show that except for MEN and the proposed MSD (to be introduced later), the score distributions are significantly imbalanced. The skewed distributions suggest that the datasets cannot support evaluations on word pairs with similarity across a variety of levels.

3.5 High Proportion of Multi-words (Synonym Selection Datasets)

Given the goal of evaluating sense embeddings, the existence of multi-words in the candidates of synonym selection questions is problematic. Since some of the existing sense embedding models themselves are unable to handle multi-words, questions with at least one multi-word alternative are usually ignored in evaluations. In RD-300, 70% of the questions contain at least one multi-word. After deleting questions that include multi-words, only 91 questions remain. In fact, all three synonym selection datasets contain less than one hundred questions after deleting multi-words. This concern renders them unable to provide a convincing evaluation for embedding models.

3.6 Lack of Pairs with Medium or Low Similarity (Synonym Selection Datasets)

The goal for the synonym selection task is to select the most semantically synonymous one among the alternatives. A vital drawback is that the synonym selection datasets thus cannot be used to evaluate whether embeddings can discern between word pairs that are only vaguely similar, or even dissimilar.

4. Construction of MSD-1030

Given the above analyses, previous semantic similarity datasets focus mainly on one sense per word. Synonym selection datasets suffer from an insufficient quantity of questions and an inability to distinguish medium- or low-similarity word pairs. To facilitate a reliable evaluation of sense embeddings, we propose a new dataset that takes these concerns into account.

4.1 Selection of Initial Words

The initial words are selected from WordNet and Roget. To ensure the words are non-rare, we exclude the words that are not in the top 10,000 frequent word list in the `wordfreq` package (Speer et al., 2018) to form word pool V. We then select a subset V_{multi} of V such that each word in the subset contains more than one sense: $V_{multi} = \{w | w \in V \; and \; |\mathcal{S}^w| > 1\}$, where $\mathcal{S}^w = \{S_1^w, \dots, S_n^w\}$ is the set of n senses of w in Roget.

4.2 Selection of Pairing Words

Our goal is to create a more balanced distribution of word pair similarities. For each word w in the initial multi-sense word pool V_{multi}, we randomly select a sense $S_i^w \in \mathcal{S}^w$. Then, we randomly select one of the five relation types listed in Figure 2. To obtain words having the selected relation with w, we use Roget's high, mid and low relevance for *synonym* (4), *similar* (3), and *related* (2) types, respectively, and use WordNet's hypernym relation for the *same domain or slight related* (1) type. The *completely unrelated* (0) type contains all the remaining words that does not belong to the aforementioned types. We then randomly select a word w_t among the words having the selected relation type t to the sense S_i^w of w to formulate the word pair (w, w_t).

Through this process, our dataset consists of a balanced distribution of these five similarity types. Since we sample word pairs with the goal of choosing multi-sense words and balancing the counts of pairs with different degrees of similarities, one limitation is that the sense distribution of words in our dataset may not reflect those in real-world applications.

4.3 Annotate Word Pairs

We recruited 11 human annotators on the Amazon Mechanical Turk crowdsourcing platform to assign similarity scores for each word pair. Compared to annotation by experts, the crowdsourcing approach allows faster data collection, making feasible the construction of a dataset consisting of thousands of word pairs. In fact, among the datasets that we introduced in Section 2, those of size at this scale are all built by crowdsourcing. We recruited only native English speakers with approval rates exceeding 95% and who had already completed more than 1,000 tasks. We introduced additional quality control mechanisms that be described later.

For word pair similarity annotation, we adopted the five-point Likert scale. In the annotation of the datasets in SemEval-2017 Task 2 (Camacho-Collados et al., 2017), the utilization of this scale and the clear definition of every score value led to extremely low disagreement. As displayed in Figure 2, we provided the annotators with the definition of each score scale, the corresponding example pairs, and additional notes. Considering that the annotators

might feel that the similarity of a pair falls between two consecutive scales, we allowed them to select scores with a step size of 0.5. Note that the annotators had no access to the information about which relation type a pair belonged to in our word selection step, which can prevent flavoring models utilizing the resources we used for constructing the dataset.

As indicated in the first note in Figure 2, when determining the score of a word pair, the annotator was required to consider the pair of meanings that results in the highest similarity. This principle more closely aligns the similarity judgments of humans and models when MaxSim is used for evaluation. Moreover, this can be considered more aligned with practical applications, as in most contexts an occurrence of a word is associated with only one meaning.

To follow the above principle, the annotators were to fully understand all definitions of the two words; if they were to forget one important meaning of a word, they would provide poor answers. To facilitate annotation, we provided several essential meanings of each word from WordNet and Dictionary.com, and provided links to word definitions in other online dictionaries. To discourage annotators from skipping the definition part, we set a minimum time: workers were to spend at least 30 seconds browsing definitions of the words before answering. The annotators were paid 0.1 USD per pair. It was estimated that they could annotate three pairs in one minute, so they earned 18 USD per hour.

The Spearman correlation between the annotated scores and the category scores of word pairs according to Figure 2 is 0.31, which is not quite high. This shows that even with a given scale, the subjective nature of similarity still makes the judgment not clear-cut in some cases. However, we will show in Section 5 that our data construction process does lead to better consistency and a more balanced score distribution.

4.4 Post-processing

To enhance the quality of our dataset, we adopted a method similar to SCWS (Huang et al., 2012): we discarded all ratings from workers with significantly low performance. We used two references to determine worker performance.

The first reference was a validation set of fifty gold pairs annotated by three supervisors who were either native or fluent English speakers familiar with our evaluation goal. A supervisor got 6.5 USD for annotating the validation set. As an inter-annotator agreement (IAA) metric, the average Spearman correlation of all pairs of supervisors is 0.71. This fairly strong correlation indicates that our annotation task is well-defined. For each gold pair, the mean of the three scores was regarded as the golden standard score.

The second reference was the comparison among peers. Large discrepancies between a worker's score and that of other workers who annotated the same word pair indicate a low-quality response. Using these two references, we manually excluded ratings from low-performing annotators.

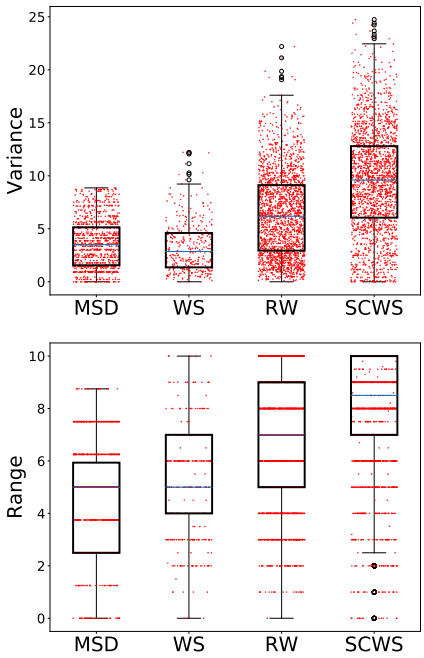

Figure 3: The distribution of annotations variance and range of each word pair in MSD-1030 and three other datasets.

After filtering out low-quality ratings, pairs had different numbers of scores. Pairs with fewer than seven scores were deleted. For pairs with more than seven scores, we removed extremely high or low values until there were seven scores left. The final score of each pair was the average score. After the above post-processing step, the Spearman correlation between the crowd scores and the gold scores on the validation set is 0.81, showing that our dataset has near-expert-level quality.

Finally, there are a total of 1,030 word pairs in our dataset. Our way of selecting words (Sections 4.1 and 4.2) ensures that at least one of the two words in every pair is multi-sense. The average number of Roget senses per word is 2.97 (SD = 2.36). The ratio of multi-sense words is 79.6%. In other words, less than 21% of the words in MSD-1030 are single-sense. This ratio is significantly lower than those shown in Table 1. Therefore, our dataset does not have the issue of dominance of single-sense words. We will examine whether other issues of previous datasets have been resolved in MSD-1030 in the following section.

5. Analyses of MSD-1030

We analyzed MSD-1030 from different aspects and compared it with one contextual word similarity dataset (SCWS) and three semantic similarity datasets (MEN, RW, and WS353)[7]. For a fair comparison, we linearly adjusted the scoring scale of every dataset to [0, 10].

5.1 Annotation Consistency

Annotation consistency is an important indicator of a crowdsourced dataset's quality. Extremely low labeling consistency for a word pair suggests that the similarity of this pair is too difficult to determine for human annotators.

[7] Though other similarity metrics exist, we explain in Section 4.3 that with MaxSim we can explicitly require both the model and human annotators to select the closest pair of senses, making the two judgments more comparable.

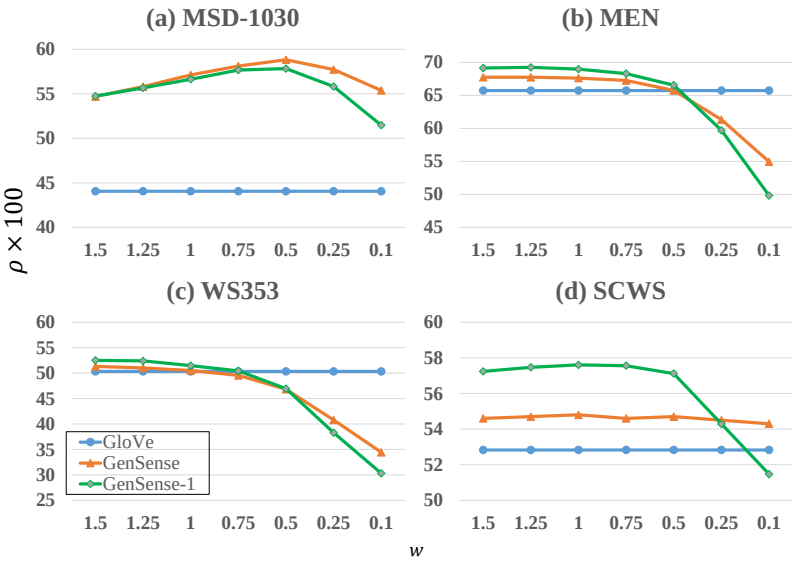

Figure 4: ρ × 100 as a function of GenSense's parameter on MSD-1030 and other datasets.

	MSD	MEN	SCWS
GenSense	**57.4**	67.6	54.8
GenSense-1	56.6	**69.0**	**57.6**
SenseRetro	**51.4**	46.9	49.7
SenseRetro-1	48.4	**51.3**	**51.6**

Table 5: ρ × 100 of 50-dimensional GenSense and SenseRetro on MSD-1030, MEN, and SCWS (all senses v.s. only first sense vectors).

Furthermore, generally low labeling consistencies of pairs in a dataset might be a result of unclear annotation guidelines.

We used the variance and range of the scores of each pair to measure annotation consistency. Figure 3 illustrates the box-and-whisker plots that represent the distribution of the scores' variance and range for the pairs in MSD-1030 and three other datasets. We do not include the MEN result in Figure 3 because its individual annotator ratings are not available. Generally, MSD-1030 and WS353 have much lower variances than the other two datasets. The mean of variances in MSD-1030 is 3.63, which indicates that its consistency is better than RW and SCWS. Although WS353 has a slightly lower mean of variance, it only contains 353 word pairs. Moreover, in terms of the range of annotation scores, MSD-1030 has the lowest mean among all four datasets (4.57, Figure 3 bottom). This indicates that the comprehensive annotation guidelines and well-designed construction procedure led to better consistency in MSD-1030.

Although a high consistency may indicate that difficult pairs have been removed at the post-processing step (Section 4.4), as we show in Section 6.1, the sense embedding models performed worse on our dataset than on previous datasets of similar size. Therefore, we conclude that there is a good balance between agreement and difficulty in the proposed MSD-1030.

5.2 Score Distribution

In Figure 1, we show the score distribution of word pairs in four bins of the scale for MSD-1030 and the other four datasets. As mentioned in Section 3.4, the distributions of RW, WS353, and SCWS are strongly biased. There are 72% and 67% of word pairs in the upper half of the similarity scale in RW and WS353, respectively. SCWS is seriously biased towards the lower half, with 69% of word pairs in [0, 5]. The score distribution of MSD-1030 is generally more balanced, which is a result of the method for selecting word pairs described in Section 4.2.

5.3 Experiments with Multi-sense Models

We conducted the experiment described in Section 3.2 again on MSD-1030. The results are shown in Table 5. As mentioned in Section 3.2, GenSense-1 and SenseRetro-1, in which only first sense vectors are utilized, outperform their multi-sense counterparts in MEN, RW, WS353, and SCWS. On the other hand, on MSD-1030 an opposite pattern is shown. The performance of GenSense-1 and SenseRetro-1 drop compared to the multi-sense versions, confirming that sense embedding models can take advantage of multiple senses to excel on the MSD-1030 dataset. Furthermore, this is consistent with the hypothesis stated in Section 3.2: adopting only the vector of the first sense of each word harms the performance. Thus, we have confirmed that the lower performance of the multi-sense models on previous datasets was not a result of the models' inability to handle non-major senses, but a result of the biases toward single-sense words and major senses in those datasets.

We performed a further detailed experiment described in Section 3.3 with MSD-1030. Using GenSense, we observed the performance changes in these datasets when adjusting the parameter w. In Figure 4, the performance changes on MEN and WS353 are in similar same patterns. The performance decreases as the weight decreases. Moreover, models that applied the first sense (GenSense-1) outperform the original model when the weight was larger. On SCWS, GenSense's performance exhibits a

	# pairs	GenSense	SenseRetro	sensegram	AdaGram	MUSE
MSD-1030	1,030	60.7 / 59.9	53.9 / 53.2	44.5 / 44.1	54.8 / 52.4	57.2 / 55.6
MEN	3,000	77.3 / 76.9	72.3 / 71.1	60.6 / 59.6	70.5 / 68.1	73.9 / 72.5
WS353	353	63.0 / 64.4	54.9 / 56.5	47.4 / 43.5	69.9 / 65.0	69.5 / 65.7
SCWS	2,003	57.2 / 60.4	59.2 / 61.4	51.7 / 52.3	64.7 / 63.4	66.8 / 62.6

Table 6: Spearman correlation ($\rho \times 100$) / Pearson correlation ($\gamma \times 100$) of sense embedding models (300-dimensional) on MSD-1030, other semantic similarity and contextual word similarity datasets.

slightly different pattern, staying constant with different original vector weights. Nevertheless, this result still contradicts our expectation: emphasizing multi-sense information from the external ontology should enhance performance.

In contrast to these three datasets, the performance on MSD-1030 confirms our expectations. There is a rise in the Spearman correlation when the original word vector weight is lowered from 1.5 to 0.5. Moreover, GenSense is generally superior to GenSense-1 across different weight values. As the parameter declines, the performance gap also widens, showing that models that emphasize multi-sense information from the ontology are stronger.

6. Evaluating Sense Embedding Models on MSD-1030

6.1 Performance of Sense Embeddings

We evaluate two knowledge-based sense embedding models: GenSense and SenseRetro, and three unsupervised models: sensegram (Pelevina et al., 2016), AdaGram (Bartunov et al., 2016), and MUSE (Lee and Chen, 2017). When training the sense embeddings, knowledge-based models take advantage of knowledge bases such as WordNet, Wikipedia, and Roget, while unsupervised models learn sense representations directly from text corpora. For all models, we downloaded the off-the-shelf 300-dimensional pre-trained vectors and reported their best results.

Table 6 shows the Spearman and Pearson correlation coefficients of various embedding models on four datasets. While these embedding models yielded high performance on previous semantic similarity datasets, their performance on MSD-1030 was generally the worst among all datasets. More specifically, given that MSD-1030 contains only common words and MEN is almost three times as large as MSD, the performance gap between the two datasets indicates that our multi-sense dataset introduces new challenges for sense embedding models.

6.2 Error Analysis

In this subsection, we analyze the word pairs that the embedding models failed to handle. GenSense and MUSE are selected as they achieve high correlation scores on MSD-1030. For comparison, we also include the results of the word-level embedding model GloVe.

Table 7 shows the word pairs with the largest similarity rank difference between model and ground-truth. We note that problems emerge when the sense that the ground-truth score is based on is rarely seen in text. For instance, the word *distribute* mostly means "disperse through a space", but it also means "divide" and is a synonym of *partition*. As word embedding models like GloVe only have one vector for each word form, a pair involving a rare sense might not be given a sufficiently high similarity score. Table 7 suggests that knowledge-based sense embedding models such as GenSense are still unable to handle this type of word pair. Based on this observation, MSD-1030 can support research on solving problems of existing sense embedding models, including the inability to handle rare senses (although all words in MSD-1030 are common).

	GloVe.6B	MSD	|D|
(distribute, partition)	915	78	837
(pat, perfectly)	886	72	814
(double, duplicate)	859	47	812
	GenSense	MSD	|D|
(distribute, partition)	863	78	785
(visit, weekend)	15	767	752
(harvest, accumulate)	862	118	744
	MUSE	MSD	|D|
(pat, perfectly)	975	72	903
(register, join)	949	89	860
(descent, falling)	933	79	854

Table 7: Word pairs with the largest differences |D| between the similarity rank of MSD-1030 and that given by the embedding models.

7. Conclusions

In this paper, we raise six concerns about existing word embedding benchmarks. When we exploit these datasets to evaluate sense embedding models, these problems are pronounced. Thus, we present MSD-1030, a high-quality and well-built semantic similarity dataset for more reliable sense embedding evaluations. In-depth analyses and experiments show that MSD-1030 has at least two merits over existing benchmarks: (1) MSD-1030 contains many multi-sense word pairs that are challenging even for state-of-the-art sense embedding models. (2) MSD-1030 has a more balanced score distribution and higher annotation consistency, compared to the other datasets. With its high reliability, MSD-1030 can support future research in evaluating sense embedding models.

Although MSD-1030 is an English dataset, we present a solid method that can be adopted in the construction of similar datasets of other languages or specific domains, given the availability of an ontology that indicates different levels of relations between lexical units. Our main future directions include investigating the relationship between evaluation results on our dataset and performance in downstream applications, which requires consideration of WSD methods, as well as developing similar datasets for contextualized representations.

8. Acknowledgements

This research was partially supported by the Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-108-2634-F-002-008-, and MOST 108-2218-E-009-051- and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

9. Bibliographical References

Bartunov, S., Kondrashkin, D., Osokin, A., and Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 130–138.

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

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882.

Jarmasz, M. and Szpakowicz, S. (2004). Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 212–219.

Jauhar, S. K., Dyer, C., and Hovy, E. (2015). Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 683–693.

Kipfer, B. A. (1993). *Roget's 21st century thesaurus in dictionary form: the essential reference for home, school, or office*. Laurel.

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

Lee, G.-H. and Chen, Y.-N. (2017). MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Lee, Y.-Y., Yen, T.-Y., Huang, H.-H., Shiue, Y.-T., and Chen, H.-H. (2018). GenSense: A generalized sense retrofitting model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1662–1671.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 104-113.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

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

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pelevina, M., Arefiev, N., Biemann, C., and Panchenko, A. (2016). Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1267–1273.

Pilehvar, M. T., Kartsaklis, D., Prokhorov, V., and Collier, N. (2018). Card-660: Cambridge rare word dataset-a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 109–117.

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). LuminosoInsight/wordfreq: v2.2.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*, pages 491–502.