De-Conflated Semantic Representations

Mohammad Taher Pilehvar and **Nigel Collier**
Language Technology Lab
University of Cambridge
Cambridge, UK
`{mp792,nhc30}@cam.ac.uk`

Abstract

One major deficiency of most semantic representation techniques is that they usually model a word type as a single point in the semantic space, hence conflating all the meanings that the word can have. Addressing this issue by learning distinct representations for individual meanings of words has been the subject of several research studies in the past few years. However, the generated sense representations are either not linked to any sense inventory or are unreliable for infrequent word senses. We propose a technique that tackles these problems by de-conflating the representations of words based on the deep knowledge it derives from a semantic network. Our approach provides multiple advantages in comparison to the past work, including its high coverage and the ability to generate accurate representations even for infrequent word senses. We carry out evaluations on six datasets across two semantic similarity tasks and report state-of-the-art results on most of them.

1 Introduction

Modeling the meanings of linguistic items in a machine-interpretable form, i.e., semantic representation, is one of the oldest, yet most active, areas of research in Natural Language Processing (NLP). The field has recently experienced a resurgence of research interest with the new blood injected in its veins by neural network-based models that view the representation task as a language modeling problem and learn dense representations (usually referred to as embeddings) by efficiently processing massive amounts of texts. However, either in its conventional count-based form (Turney and Pantel, 2010) or the recent predictive approach, the prevailing objective of representing each word type as a single point in the semantic space has a major limitation: it ignores the fact that words can have multiple meanings and conflates all these meanings into a single representation. This objective can have negative impacts on accurate semantic modeling, e.g., semantically unrelated words that are synonymous to different senses of a word are pulled towards each other in the semantic space (Neelakantan et al., 2014).

Recently, there has been a growing interest in addressing the meaning conflation deficiency of word representations. A series of techniques tend to associate a word to multiple points in the semantic space by clustering its contexts in a given text corpus and learning distinct representations for individual clusters (Reisinger and Mooney, 2010; Huang et al., 2012). Though, these techniques usually assume a fixed number of word senses per word type, disregarding the fact that the number of senses of a word can range from one (monosemy) to dozens. Neelakantan et al. (2014) tackled this issue by allowing the number to be dynamically adjusted for each word during training. However, the approach and all the other clustering-based techniques still suffer from the fact that the computed sense representations are not linked to any sense inventory, a linking which would require large amounts of sense-annotated data (Agirre et al., 2006). In addition, because of their dependence on knowledge derived from a text corpus, these techniques are generally unable to learn accurate representations for word

senses that are infrequent in the underlying corpus.

Knowledge-based techniques tackle these issues by deriving sense-specific knowledge from external sense inventories, such as WordNet (Fellbaum, 1998), and learning representations that are linked to the sense inventory. These approaches either use sense definitions and employ Word Sense Disambiguation (WSD) to gather sense-specific contexts (Chen et al., 2014; Iacobacci et al., 2015) or take advantage of the properties of WordNet, such as synonymy and direct semantic relations (Rothe and Schütze, 2015). However, the non-optimal WSD techniques and the shallow utilization of knowledge from WordNet do not allow these techniques to learn accurate and high-coverage semantic representations for all senses in the inventory.

We propose a technique that de-conflates a given word representation into its constituent sense representations by exploiting deep knowledge from the semantic network of WordNet. Our approach provides the following three main advantages in comparison to the past work: (1) our representations are linked to the WordNet sense inventory and, accordingly, the number of senses for a word is a dynamic parameter which matches that defined by WordNet; (2) the deep exploitation of WordNet's semantic network allows us to obtain accurate semantic representations, even for word senses that are infrequent in generic text corpora; and (3) our methodology involves only minimal parameter tuning and can be effectively applied to any sense inventory that is viewable as a semantic network and to any word representation technique. We evaluate our sense representations in two tasks: word similarity (both in-context and in-isolation) and cross-level semantic similarity. Experimental results show that the proposed technique can provide consistently high performance across six datasets, outperforming the recent state of the art on most of them.

2 De-Conflated Representations

Preliminaries. Our proposed approach takes a set of pre-trained word representations and uses the graph structure of a semantic lexical resource in order to de-conflate the representations into those of word senses. Therefore, our approach firstly requires a set of pre-trained word representations (e.g.,

word embeddings). Any model that maps a given word to a fixed-size vector representation (i.e., vector space model) can be used by our approach. In our experiments, we opted for a set of publicly available word embeddings (cf. §3.1).

Secondly, we require a lexical resource whose semantic relations allow us to view it as a graph $G = (V, E)$ where each vertex in the set of vertices V corresponds to a concept and edges in E denote lexico-semantic relationships among these vertices. Each concept $c \in V$ is mapped to a set of word senses by a mapping function $\mu(c) : c \rightarrow \{s_1, \ldots, s_l\}$. WordNet, the *de facto* community standard sense inventory, is a suitable resource that satisfies these properties. WordNet can be readily represented as a semantic graph in which vertices are synsets and edges are the semantic relations that connect these synsets (e.g., hypernymy and meronymy). The mapping function in WordNet maps each synset to the set of synonymous words it contains (i.e., word senses).

2.1 Overview of the approach

Our goal is to compute a semantic representation that places a given word sense in an existing semantic space of words. We achieve this by leveraging word representations as well as the knowledge derived from WordNet. The gist of our approach lies in its computation of a list of *sense biasing words* for a given word sense. To this end, we first analyze the semantic network of WordNet and extract a list of most representative words that can effectively pinpoint the semantics of individual synsets (§2.2). We then leverage an effective technique which learns semantic representations for individual word senses by placing the senses in the proximity of their corresponding sense biasing words (§2.3).

2.2 Determining sense biasing words

Algorithm 1 shows the procedure we use to extract from WordNet a list of sense biasing words for a given target synset y_t. The algorithm receives as its inputs the semantic graph of WordNet and the mapping function $\mu(\cdot)$, and outputs an ordered list of biasing words $\mathcal{B}_t$ for y_t. The list comprises the most semantically-related words to synset y_t which can best represent and pinpoint its meaning. We leverage a graph-based algorithm for the computation of the sense biasing words.

Algorithm 1 Get sense biasing words for synset y_t

Require: Graph $G = (V, E)$ of vertices $V = \{y_i\}_{i=1}^m$ (of m synsets) and edges E (semantic relationships between synsets)

Require: Function $\mu(y_i)$ that returns for a given synset y_i the words it contains

Require: Target synset $y_t \in V$ for which a sense biasing word sequence is required

Ensure: The sequence $\mathcal{B}_t$ of sense biasing words for synset y_t

1: $\mathcal{B}_t \leftarrow ()$
2: **for all** word w in $\mu(y_t)$ **do**
3: $\quad \mathcal{B}_t \leftarrow \mathcal{B}_t \cup (w)$
4: **for** $y_i \in V : y_i \neq y_t$ **do**
5: $\quad p_i \leftarrow \text{PERSONALIZEDPAGERANK}(y_i, y_t, G)$
6: $(y_h^*)_{h=1}^{m-1} \leftarrow \text{SORT}(V \setminus \{y_t\})$ according to scores p_i
7: **for** $h : 1$ to $m - 1$ **do**
8: $\quad$ **for all** word w in $\mu(y_h^*)$ **do**
9: $\quad\quad$ **if** $w \notin \mathcal{B}_t$ **then**
10: $\quad\quad\quad \mathcal{B}_t \leftarrow \mathcal{B}_t \cup (w)$
11: **return** sequence $\mathcal{B}_t$

#	Sense biasing words
1	dactyl, finger, toe, thumb, pollex, body_part, nail, minimus, tarsier, webbed, extremity, appendage
2	figure, cardinal_number, cardinal, integer, whole_number, numeration_system, number_system, system_of_numeration, large_integer, constituent, element, digital

Table 1: The top sense biasing words for the synsets containing the anatomical (#1) and numerical (#2) senses of the noun *digit*.

Specifically, we use the Personalized PageRank (Haveliwala, 2002, PPR) algorithm which has been extensively used by several NLP applications (Yeh et al., 2009; Niemann and Gurevych, 2011; Agirre et al., 2014). To this end, we first represent the semantic network of WordNet as a row-stochastic transition matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ where m is the number of synsets in WordNet ($|V|$). The cell M_{ij} of $\mathbf{M}$ is set to the inverse of the degree of i if there is a semantic relationship between synsets i and j and to zero otherwise. We compute the PPR distribution for a target synset y_t by using the power iteration method $\mathcal{P}^{t+1} = (1 - \sigma)\mathcal{P}^0 + \sigma \mathbf{M} \mathcal{P}^t$, where σ is the damping factor (usually set to 0.85) and $\mathcal{P}^0$ is a one-hot initialization vector with the corresponding dimension of y_t being set to 1.0. The weight p_i in line 5 is the value of the ith dimension of the PPR vector $\mathcal{P}$ computed for the synset y_t. This weight can be seen as the importance of the corresponding synset of the ith dimension (i.e., y_i) to y_t. When applied to a semantic network, such as the WordNet graph, this importance can be interpreted as semantic relevance. Hence, the value of p_i denotes the extent of semantic relatedness between y_i and y_t. We use this notion and retrieve a list of most semantically-related words to y_t. To achieve this, we sort the synsets $\{y^* \in V : y^* \neq y_t\}$ according to their PPR values

$\{p_i\}_{i=1}^{m-1}$ (line 6). We then iterate (lines 7-10) the sorted list (y^*) and for each synset y_h^* append the list $\mathcal{B}_t$ with all the words in y_h^* (i.e., $\mu(y_h^*)$). However, in order to ensure that the words in the target synset y_t appear as the most representative words in $\mathcal{B}_t$, we first assign these words to the list (line 3). Finally, the algorithm returns the ordered list $\mathcal{B}_t$ of sense biasing words for the target synset y_t.

Table 1 shows a sample of top biasing words extracted for the two senses of the noun *digit*: the numerical and the anatomical senses.[1] We explain in §2.3 how we use the sense biasing lists to learn sense-specific representations. Note that the size of the list is equal to the total number of strings in WordNet. However, we observed that taking a very small portion of the top-ranking elements in the lists is enough to generate representations that perform very similarly to those generated when using the full-sized lists (please see §3.1).

2.3 Learning sense representations

Let $\mathcal{V}$ be the set of pre-trained d-dimensional word representations. Our objective here is to compute a set $\mathcal{V}^* = \{v_{s_1}^*, \ldots, v_{s_n}^*\}$ of representations for n word senses $\{s_1, \ldots, s_n\}$ in the same d-dimensional semantic space of words. We achieve this for each sense s_i by de-conflating the representation v_{s_i} of its corresponding lemma and biasing it towards the representations of the words in $\mathcal{B}_i$. Specifically, we obtain a representation $v_{s_i}^*$ for a word sense s_i by solving:

$$\arg\min_{v_{s_i}^*} \alpha \; d(v_{s_i}^*, v_{s_i}) + \sum_{b_{ij} \in \mathcal{B}_i} \delta_{ij} \; d(v_{s_i}^*, v_{b_{ij}}) \quad (1)$$

where v_{s_i} and $v_{b_{ij}}$ are the respective word representations ($\in \mathcal{V}$) of the lemma of s_i and the jth biasing word in the list of biasing words for s_i,

---

[1]The first and third senses of the noun *digit* in WordNet 3.0.

#	Closest words
1	crappie, trout, guitar, shad, walleye, bassist, angler, catfish, trombone, percussion, piano, drummer, saxophone, jigs, fish
2	baritone, piano, guitar, trombone, saxophone, cello, percussion, tenor, saxophonist, clarinet, pianist, vocals, solos, harmonica
3	fish, trout, shrimp, anglers, fishing, bait, guitar, salmon, shark, fisherman, lakes, seafood, drummer, whale, fisheries

Table 2: Ten most similar words to the word *bass* (#1) and two of its senses: music (#2) and fish (#3).

i.e, $\mathcal{B}_i$. The distance $d(v, v')$ between vectors v and v' is measured by squared Euclidean distance $\|v - v'\|^2 = \sum_k (v_k - v'_k)^2$. The first term in Formula 1 requires the representation of the word sense s_i (i.e., $v^*_{s_i}$) to be similar to that of its corresponding lemma, i.e., v_{s_i}, whereas the second term encourages $v^*_{s_i}$ to be in the proximity of its biasing words in the semantic space. The above criterion is similar to the frameworks of Das and Smith (2011) and Faruqui et al. (2015) which, though being convex, is usually solved for efficiency reasons by an iterative method proposed by Bengio et al. (2007). Following these works, we obtain the below equation for computing the representation of a word sense s_i:

$$v^*_{s_i} = \frac{\alpha v_{s_i} + \sum_{b_{ij} \in \mathcal{B}_i} \delta_{ij} v_{b_{ij}}}{\alpha + \sum_j \delta_{ij}}. \qquad (2)$$

We define δ_{ij} as $\frac{e^{-\lambda r(i,j)}}{|\mathcal{B}_i|}$ where $r(i, j)$ denotes the rank of the word b_{ij} in the list $\mathcal{B}_i$. This is essentially an exponential decay function that gives more importance to the top-ranking biasing words for s_i. The hyperparameter α denotes the extent to which $v^*_{s_i}$ is kept close to its corresponding lemma representation v_{s_i}. Following Faruqui et al. (2015), we set α to 1. The only parameter to be tuned in our experiments is λ. We discuss the tuning of this parameter in §3.1. The representation of a synset y_i can be accordingly calculated as the centroid of the vectors of its associated word senses, i.e.,

$$\left\{ \frac{v_{y_i}}{\|v_{y_i}\|} : v_{y_i} = \sum_{s \in \mu(y_i)} \hat{v}^*_s, \hat{v}^*_s = \frac{v^*_s}{\|v^*_s\|} \right\}. \qquad (3)$$

As a result of this procedure, we obtain the set $\mathcal{V}^*$ of n sense representations in the same semantic space as word representations in $\mathcal{V}$. In fact, we now have a unified semantic space which enables a direct
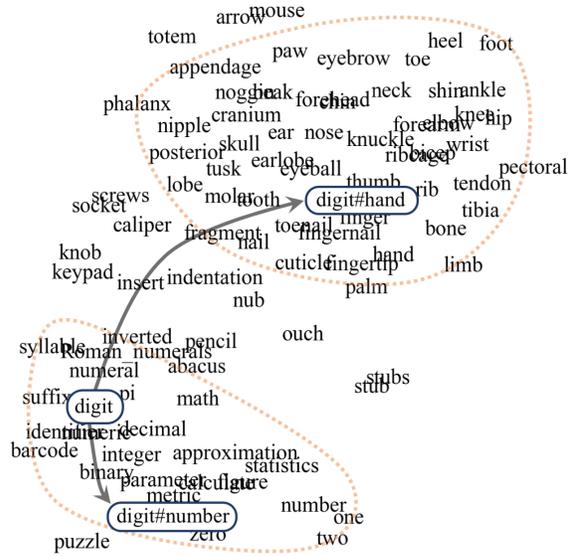

Figure 1: The illustration of the word *digit* and two of its computed senses in our unified 2-d semantic space.

comparison of the two types of linguistic items. In §3.3 we evaluate our approach in the word to sense similarity measurement framework.

We show in Table 2 the closest words to the word *bass* and two of its senses, music and fish,[2] in our unified semantic space. We can see in row #1 a mixture of both meanings when the word representation is used whereas the closest words to the senses (rows #2 and #3) are mostly in-domain and specific to the corresponding sense.

To exhibit another interesting property of our sense representation approach, we depict in Figure 1 the word *digit* and its numerical and anatomical senses (from the example in Table 1) in a 2-d semantic space, along with a sample set of words in their proximity.[3] We can see that the word *digit* is placed in the semantic space in the neighbourhood of words from the numerical domain (lower left of the figure), mainly due the dominance (Sanderson and Van Rijsbergen, 1999) of this sense in the general-domain corpus on which the word embeddings in our experiments were trained (cf. §3.1). However, upon de-conflation, the emerging anatomical sense of the

---

[2] The first and fourth senses in WordNet 3.0, respectively defined as "the lowest part of the musical range" and "the lean flesh of a saltwater fish of the family Serranidae."

[3] We used the t-SNE algorithm (van der Maaten and Hinton, 2008) for dimensionality reduction.

word is shifted towards the region in the semantic space which is occupied by anatomical words (upper right of the figure). A clustering-based sense representation technique would have failed in accurately representing the infrequent anatomical meaning of *digit* by analyzing a general domain corpus (such as the one used here). But our sense representation technique, thanks to its proper usage of knowledge from a sense inventory, is effective in unveiling and accurately modeling less frequent or domain-specific senses of a given word.

Please note that any vector space model representation technique can be used for the pre-training of word representations in $\mathcal{V}$. Also, the list of sense biasing words can be obtained for larger sense inventories, such as FreeBase (Bollacker et al., 2008) or BabelNet (Navigli and Ponzetto, 2012). We leave the exploration of other ways of computing sense biasing words to the future work.

3 Experiments

We benchmarked our sense representation approach against several recent techniques on two standard tasks: word similarity (§3.2), for which we evaluate on both in-isolation and in-context similarity datasets, and cross-level semantic similarity (§3.3).

3.1 Experimental setup

Pre-trained word representations. As our word representations, we used the 300-d Word2vec (Mikolov et al., 2013) word embeddings trained on the Google News dataset[4] mainly for their popularity across different NLP applications. However, our approach is equally applicable to any count-based representation technique (Baroni and Lenci, 2010; Turney and Pantel, 2010) or any other embedding approach (Pennington et al., 2014; LeCun et al., 2015). We leave the evaluation and comparison of various word representation techniques with different training approaches, objectives, and dimensionalities to the future work.

Parameter tuning. Recall from §2.3 that our procedure for learning sense representations needs only one parameter to be tuned, i.e., λ. We did not perform an extensive tuning on the value of this parameter and set its value to ⅕ after trying four differ-

---

ent values (1, ½, ⅕, and 1/10) on a small validation dataset. We leave the more systematic tuning of the parameter and the choice of alternative decay functions (cf. §2.3) to the future work.

The size of the sense biasing words lists. Also recall from §2.2 that the extracted lists of sense biasing words were originally as large as the total number of unique strings in WordNet (around 150K in ver. 3.0). But, given that we use an exponential decay function in our learning algorithm (cf. §2.3), the impact of the low-ranking words in the list is negligible. In fact, we observed that taking a very small portion of the top-ranking words, i.e., the top 25, produces similarity scores that are on par with those generated when the full lists were considered. Therefore, we experimented with the down-sized lists which enabled us to generate very quickly sense representations for all word senses in WordNet.

3.2 Word similarity

Comparison systems. We compared our results against nine other sense representation techniques: the WordNet-based approaches of Pilehvar and Navigli (2015), Chen et al. (2014), Rothe and Schütze (2015), Jauhar et al. (2015), and Iacobacci et al. (2015) and the clustering-based approaches of Huang et al. (2012), Tian et al. (2014), Neelakantan et al. (2014), and Liu et al. (2015) (please see §4 for more details). We also compared against the approach of Faruqui et al. (2015) which uses knowledge derived from WordNet for improving word representations. From the different configurations presented in (Faruqui et al., 2015) we chose the system that uses GloVe (Pennington et al., 2014) with *all* WordNet relations which is their best performing monolingual system. As for the approach of Jauhar et al. (2015), we show the results of the EM+RETERO system which performs most consistently across different datasets.

Benchmarks. As our word similarity benchmark, we considered five datasets: RG-65 (Rubenstein and Goodenough, 1965), YP-130 (Yang and Powers, 2005), MEN-3K (Bruni et al., 2014), SimLex-999 (Hill et al., 2015, SL-999), and Stanford Contextual Word Similarity (Huang et al., 2012, SCWS). The latter benchmark provides for each word a context that triggers a specific meaning of the word, making

Dataset	Approach	Sense-based score		Word-based score	
		r	ρ	r	ρ
MEN-3K	Iacobacci et al. (2015)	–	**80.5**	–	66.5
	DECONF	**78.0**	78.6	72.3	73.2
	Faruqui et al. (2015)	–	75.9	–	73.7
	Pilehvar and Navigli (2015)	61.7	66.6	–	–
RG-65	DECONF	**90.5**	**89.6**	77.2	76.1
	Iacobacci et al. (2015)	–	87.1	–	73.2
	Faruqui et al. (2015)	–	84.2	–	76.7
	Pilehvar and Navigli (2015)	80.2	84.3	–	–
YP-130	Pilehvar and Navigli (2015)	**79.0**	**71.0**	–	–
	DECONF	72.9	69.5	58.0	55.9
	Iacobacci et al. (2015)	–	63.9	–	34.3
SimLex-999	DECONF	54.2	51.7	45.4	44.2
	Pilehvar and Navigli (2015)	43.4	43.6	–	–

Table 3: Pearson ($r \times 100$) and Spearman ($\rho \times 100$) correlation scores on four standard word similarity benchmarks. For each benchmark, we show the results reported by any of the comparison systems along with the scores for their corresponding initial word representations (word-based).

Approach	Score	
	AvgSim	AvgSimC
DECONF	**70.8**	**71.5**
Rothe and Schütze (2015) (best)	68.9	69.8
Neelakantan et al. (2014) (best)	67.3	69.3
Chen et al. (2014)	66.2	68.9
Liu et al. (2015) (best)	–	68.1
Huang et al. (2012)	62.8	65.7
Tian et al. (2014) (best)	–	65.7
Iacobacci et al. (2015)	62.4	–
Jauhar et al. (2015)	–	58.7
Initial word vectors		65.1

Table 4: Spearman correlation scores ($\rho \times 100$) on the Stanford Contextual Word Similarity (SCWS) dataset. We report the AvgSim and AvgSimC scores (cf. §3.2) for each system, where available.

it very suitable for the evaluation of sense representation techniques. For each of the datasets, we list the results that are reported by any of our comparison systems.

Similarity measurement. For the SCWS dataset, we follow the past works (Reisinger and Mooney, 2010; Huang et al., 2012) and report the results according to two system configurations: (1) AvgSim: where the similarity between two words is computed as the average of all the pairwise similarities between their senses, and (2) AvgSimC: where each pairwise sense similarity is weighted by the rele-

vance of each sense to its corresponding context. For all the other datasets, since words are not provided with any context (they are in isolation), we measure the similarity between two words as that between their most similar senses. In all the experiments, we use the cosine distance as our similarity measure.

3.2.1 Experimental results

Tables 4 and 3 show the results of our system, DE-CONF, and the comparison systems on the SCWS and the other four similarity datasets, respectively. In both tables we also report the word vectors baseline, whenever they are available, which is computed by directly comparing the corresponding word representations of the two words ($\in \mathcal{V}$). Please note that the word-based baseline does not apply to the approach of Pilehvar and Navigli (2015) as it is purely based on the semantic network of WordNet and does not use any pre-trained word embeddings.

We can see from the tables that our sense representations obtain considerable improvements over those of words across the five datasets. This highlights the fact that the de-conflation of word representations into those of their individual meanings has been highly beneficial. On the SCWS dataset, DECONF outperforms all the recent state-of-the-art sense representation techniques (in their best settings) which proves the effectiveness of our approach in capturing the semantics of specific mean-

ings of the words. The improvement is consistent across both system configurations (i.e., AvgSim and AvgSimC). Moreover, the state-of-the-art WordNet-based approach of Rothe and Schütze (2015) uses the same initial word vectors as DECONF does (cf. §3.1). Hence, the improvement we obtain indicates that our approach has made better use of the sense-specific knowledge encoded in WordNet.

As seen in Table 3 our approach shows competitve performance on the other four datasets. The YP-130 dataset focuses on verb similarity, whereas SimLex-999 contains verbs and adjectives and MEN-3K has word pairs with different parts of speech (e.g., a noun compared to a verb). The results we obtain on these datasets exhibit the reliability of our approach in modeling non-nominal word senses.

3.2.2 Discussion

The similarity scale of the SimLex-999 dataset is different from our other word similarity benchmarks in that it assigns relatively low scores to antonymous pairs. For instance, *sunset-sunrise* and *man-woman* in this dataset are assigned the respective similarities of 2.47 and 3.33 (in a $[0, 10]$ similarity scale) which is in the same range as the similarity between word pairs with slight domain relatedness, such as *head-nail* (2.47), *air-molecule* (3.05), or *succeed-try* (3.98). In fact we observed that tweaking the similarity scale of our system in a way that it diminishes the similarity scores between antonyms can result in significant performance improvement on this dataset. To this end, we performed an experiment in which the similarity of a word pair was simply divided by five whenever the two words belonged to synsets that were linked by the antonymy relation. We observed that the performance on the SimLex-999 dataset increased to 61.1 (from 54.2) and 59.0 (from 51.7) according to Pearson ($r \times 100$) and Spearman ($\rho \times 100$) correlation scores, respectively.

3.3 Cross-Level semantic similarity

In addition to the word similarity benchmark, we evaluated the performance of our representations in the cross-level semantic similarity measurement framework. To this end, we opted for the SemEval-2014 task on Cross-Level Semantic Similarity (Jurgens et al., 2014, CLSS). The word to sense similarity subtask of this task, with 500 instances in its test set, provides a suitable benchmark for the evaluation of sense representation techniques.

For a word sense s and a word w, we compute the similarity score according to four different strategies: the similarity of s to the most similar sense of w (**MaxSim**), the average similarity of s to individual senses of w (**AvgSim**), the direct similarity of s to w when the latter is modeled as its word representation (Sense-to-Word or **S2W**) or as the centroid of its senses' representations (Sense to aggregated word senses or **S2A**). For this task, we can only compare against the publicly-available sense representations of Iacobacci et al. (2015), Rothe and Schütze (2015), Pilehvar and Navigli (2015) and Chen et al. (2014) which are linked to the WordNet sense inventory.

3.3.1 Experimental results

Table 5 shows the results on the word to sense dataset of the SemEval-2014 CLSS task, according to Pearson (r) and Spearman (ρ) correlations and for the four strategies. As can be seen from the low overall performances, the task is a very challenging benchmark with many WordNet out-of-vocabulary or slang terms and rare usages. Despite this, DE-CONF provides consistent improvement over the comparison sense representation techniques according to both measures and for all the strategies.

Across the four strategies, S2A proves to be the most effective for DECONF and the representations of Rothe and Schütze (2015). The representations of Chen et al. (2014) perform best with the S2W strategy whereas those of Iacobacci et al. (2015) do not show a consistent trend with relatively low performance across the four strategies. Also, a comparison of our results across the S2W and S2A strategies reveals that a word's aggregated representation, i.e., the centroid of the representations of its senses, is more accurate than its original word representation.

Our analysis showed that the performances of the approaches of Rothe and Schütze (2015) and Iacobacci et al. (2015) were hampered partly due to their limited coverage. In fact, the former was unable to model around 35% of the synsets in WordNet 1.7.1, mainly for its shallow exploitation of knowledge from WordNet, whereas the latter approach did not cover around 15% of synsets in WordNet 3.0. Chen et al. (2014) provide near-full coverage for

System	MaxSim		AvgSim		S2W		S2A	
	r	ρ	r	ρ	r	ρ	r	ρ
DECONF*	**36.4**	**37.6**	**36.8**	**38.8**	**34.9**	**35.6**	**37.5**	**39.3**
Rothe and Schütze (2015)*	34.0	33.8	34.1	33.6	33.4	32.0	35.4	34.9
Iacobacci et al. (2015)*	19.1	21.5	21.3	24.2	22.7	21.7	19.5	21.1
Chen et al. (2014)*	17.7	18.0	17.2	16.8	27.7	26.7	17.9	18.8
DECONF	35.5	36.4	36.2	38.0	34.9	35.6	36.8	38.4
Pilehvar and Navigli (2015)	19.4	23.8	21.2	26.0	—	—	—	—
Iacobacci et al. (2015)	19.0	21.5	20.9	23.2	22.3	20.6	19.2	20.4

Table 5: Evaluation results on the word to sense similarity test dataset of the SemEval-14 task on Cross-Level Semantic Similarity, according to Pearson ($r \times 100$) and Spearman ($\rho \times 100$) correlations. We show results for four similarity computation strategies (see §3.3). The best results per strategy are shown in bold whereas they are underlined for the best strategies per system. Systems marked with $*$ are evaluated on a slightly smaller dataset (474 of the original 500 pairs) so as to have a fair comparison with Rothe and Schütze (2015) and Chen et al. (2014) that use older versions of WordNet (1.7.1 and 1.7, respectively).

word senses in WordNet. However, the relatively low performance of their system shows that the usage of glosses in WordNet and the automated disambiguation have not resulted in accurate sense representations. Thanks to its deep exploitation of the underlying resource, our approach provides full coverage over all word senses and synsets in WordNet.

The three best-performing systems in the task are Meerkat_Mafia (Kashyap et al., 2014) ($r = 37.5$, $\rho = 39.3$), SimCompass (Banea et al., 2014) ($r = 35.4$, $\rho = 34.9$), and SemantiKLUE (Proisl et al., 2014) ($r = 17.9$, $\rho = 18.8$). Please note that these systems are specifically designed for the cross-level similarity measurement task. For instance, the best-ranking system in the task leverages a compilation of several dictionaries, including The American Heritage Dictionary, Wiktionary and WordNet, in order to handle slang terms and rare usages, which leads to its competitive performance (Kashyap et al., 2014).

4 Related Work

Learning semantic representations for individual senses of words has been an active area of research for the past few years. Based on the way they view the problem, the recent techniques can be classified into two main branches: (1) those that, similarly to our work, extract knowledge from external sense inventories for learning sense representations; and (2) those techniques that cluster the contexts in which a word appears in a given text corpus and learn distinct representations for individual clusters.

Examples for the first branch include the approaches of Chen et al. (2014), Jauhar et al. (2015) and Rothe and Schütze (2015), all of which use WordNet as an external resource and obtain sense representations for this sense inventory. Chen et al. (2014) uses the content words in the definition of a word sense and WSD. However, the sole usage of glosses as sense-distinguishing contexts and the non-optimal WSD make the approach inaccurate, particularly for highly polysemous words with similar senses and for word senses with short definitions. Similarly, Rothe and Schütze (2015) use only polysemy and synonymy properties of words in WordNet along with a small set of semantic relations. This significantly hampers the reliability of the technique in providing high coverage (discussed further in §3.3.1). Our approach improves over these works by exploiting deep knowledge from the semantic network of WordNet, coupled with an effective training approach. ADW (Pilehvar and Navigli, 2015) is another WordNet-based approach which exploits only the semantic network of this resource an obtains interpretable sense representations. Other work in this branch include SensEmbed (Iacobacci et al., 2015) and Nasari (Camacho-Collados et al., 2015; Camacho-Collados et al., 2016) which are based on the BabelNet sense inventory (Navigli and Ponzetto, 2012). The former technique first disambiguates words in a given corpus with the help of a knowledge-based WSD system and then uses the generated sense-annotated corpus as training data for Word2vec. Nasari combines structural knowledge from the semantic network of BabelNet

with corpus statistics derived from Wikipedia for representing BabelNet synsets. However, the approach falls short of modeling non-nominal senses as Wikipedia, due to its very encyclopedic nature, does not cover verbs, adjectives, or adverbs.

The second branch, which is usually referred to as *multi-prototype* representation, is often associated with clustering. Reisinger and Mooney (2010) proposed one of the recent pioneering techniques in this branch. Other prominent work in the category include topical word embeddings (Liu et al., 2015) which use latent topic models for assigning topics to each word in a corpus and learn topic-specific word representations, and the technique proposed by Huang et al. (2012) which incorporates "global document context." Tian et al. (2014) modified the Skip-gram model in order to learn multiple embeddings for each word type. Despite the fact that these techniques do not usually take advantage of the knowledge encoded in structured knowledge resource, they generally suffer from two disadvantages. The first limitation is that they usually make an assumption that a given word has a fixed number of senses, ignoring the fact that polysemy is highly dynamic across words that can range from monosemous to highly ambiguous with dozens of associated meanings (McCarthy et al., 2016). Neelakantan et al. (2014) tackled this issue by estimating the number of senses for a word type during the learning process. However, all techniques in the second branch suffer from another disadvantage that their computed sense representations are not linked to any sense inventory, a linking which itself would require the existence of high coverage sense-annotated data (Agirre et al., 2006).

Another notable line of research incorporates knowledge from external resources, such as PPDB (Ganitkevitch et al., 2013) and WordNet, to improve word embeddings (Yu and Dredze, 2014; Faruqui et al., 2015). Neither of the two techniques however provide representations for word senses.

5   Conclusions

We put forward a sense representation technique, namely DECONF, that provides multiple advantages in comparison to the recent state of the art: (1) the number of word senses in our technique is flexible and the computed representations are linked to word senses in WordNet; (2) DECONF is effective in providing accurate representation of word senses, even for those senses that do not usually appear frequently in generic text corpora; and (3) our approach is general in that it can be readily applied to any set of word representations and any semantic network without the need for extensive parameter tuning. Our experimental results showed that DECONF can outperform recent state of the art on several datasets across two tasks. We release our computed representations for around 118K synsets and 205K word senses in WordNet 3.0 at `https://github.com/pilehvar/deconf`. As future work, we plan to investigate the possibility of using larger semantic networks, such as FreeBase and BabelNet, which would also allow us to apply the technique to languages other than English. We also plan to evaluate the performance of our approach with other decay functions as well as with other initial word representations.

Acknowledgments

References

[Agirre et al.2006] Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 89–96.

[Agirre et al.2014] Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

[Banea et al.2014] Carmen Banea, Di Chen, Rada Mihalcea, Claire Cardie, and Janyce Wiebe. 2014. Simcompass: Using deep learning word embeddings to assess cross-level similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 560–565, Dublin, Ireland.

[Baroni and Lenci2010] Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

[Bengio et al.2007] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 2007. *Semi-Supervised Learning*. MIT Press. chapter Label Propagation and Quadratic Criterion.

[Bollacker et al.2008] 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.

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

[Camacho-Collados et al.2015] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

[Camacho-Collados et al.2016] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. NASARI: Integrating explicit knowledge and corpus statistics for amultilingual representation of concepts and entities. *Artificial Intelligence*.

[Chen et al.2014] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP 2014*, pages 1025–1035, Doha, Qatar.

[Das and Smith2011] Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1435–1444, Portland, Oregon, USA.

[Faruqui et al.2015] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

[Ganitkevitch et al.2013] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia.

[Haveliwala2002] Taher H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web*, pages 517–526.

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

[Huang et al.2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, pages 873–882, Jeju Island, Korea.

[Iacobacci et al.2015] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China.

[Jauhar et al.2015] Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 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, Denver, Colorado.

[Jurgens et al.2014] David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland, August.

[Kashyap et al.2014] Abhay L. Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya W. Satyapanich, Sunil R Gandhi, and Tim Finin. 2014. Meerkat Mafia: Multilingual and Cross-Level Semantic Textual Similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 416–423, August.

[LeCun et al.2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

[Liu et al.2015] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2418–2424.

[McCarthy et al.2016] Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, in press.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop*

at International Conference on Learning Representations, Scottsdale, Arizona.

[Navigli and Ponzetto2012] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Neelakantan et al.2014] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP 2014*, pages 1059–1069, Doha, Qatar.

[Niemann and Gurevych2011] Elisabeth Niemann and Iryna Gurevych. 2011. The people's Web meets linguistic knowledge: Automatic sense alignment of Wikipedia and Wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 205–214.

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

[Pilehvar and Navigli2015] Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

[Proisl et al.2014] Thomas Proisl, Stefan Evert, Paul Greiner, and Besim Kabashi. 2014. Semantiklue: Robust semantic similarity at multiple levels using maximum weight matching. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 532–540, Dublin, Ireland.

[Reisinger and Mooney2010] Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California.

[Rothe and Schütze2015] Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China.

[Rubenstein and Goodenough1965] Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[Sanderson and Van Rijsbergen1999] Mark Sanderson and C. J. Van Rijsbergen. 1999. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions of Infmation Systems*, 17(4):440–465.

[Tian et al.2014] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, Dublin, Ireland, August.

[Turney and Pantel2010] Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

[van der Maaten and Hinton2008] L.J.P van der Maaten and G.E. Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9: 25792605.

[Yang and Powers2005] 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.

[Yeh et al.2009] Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49.

[Yu and Dredze2014] Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland.