On the Importance of Distinguishing Word Meaning Representations: A Case Study on Reverse Dictionary Mapping

Mohammad Taher Pilehvar
Tehran Institute for Advanced Studies (TeIAS), Tehran, Iran
DTAL, University of Cambridge, Cambridge, UK
mp792@cam.ac.uk

Abstract

Meaning conflation deficiency is one of the main limiting factors of word representations which, given their widespread use at the core of many NLP systems, can lead to inaccurate semantic understanding of the input text and inevitably hamper the performance. Sense representations target this problem. However, their potential impact has rarely been investigated in downstream NLP applications. Through a set of experiments on a state-of-the-art reverse dictionary system based on neural networks, we show that a simple adjustment aimed at addressing the meaning conflation deficiency can lead to substantial improvements.

1 Meaning Conflation Deficiency

Words are often the most fine-grained meaning bearing components of NLP systems. As a standard practise, particularly for neural models, the input text is treated as a sequence of words and each word in the sequence is represent with a dense distributional representation (word embedding). Importantly, this setting ignores the fact that a word can be polysemous, i.e., it can take multiple (possibly unrelated) meanings. Representing a word with all its possible meanings as a single point (vector) in the embedding space, the so-called meaning conflation deficiency (Camacho-Collados and Pilehvar, 2018), can hinder system's semantic effectiveness.

To address this deficiency, many techniques have been put forward over the past few years, the most prominent of which is sense representation or multi-prototype embedding (Schütze, 1998; Reisinger and Mooney, 2010). However, as a general trend, these representations are usually evaluated either on generic benchmarks, such as word similarity, or on sense-centered tasks such as Word Sense Disambiguation, leaving their potential impact on downstream word-based systems unknown. In this paper, we provide an analysis to highlight the importance of addressing the meaning conflation deficiency. Specifically, we show how distinguishing different meanings of a word can facilitate a more accurate semantic understanding of a state-of-the-art reverse dictionary system, reflected by substantial improvements in recall and generalisation power.

2 Reverse Dictionary

Reverse dictionary, conceptual dictionary, or concept lookup is the task of returning a word given its description or definition (Brown and McNeill, 1966; Zock and Bilac, 2004). For example, given "a crystal of snow", the system has to return the word *snowflake*. The task is closely related to the "tip of the tongue" problem where an individual recalls some general features about a word but cannot retrieve that from memory. Therefore, a reverse dictionary system can be particularly useful to writers and translators when they cannot recall a word in time or are unsure how to express an idea they want to convey.

2.1 Evaluation framework

Our experiments are based on the reverse dictionary model of Hill et al. (2016) which leverages a standard neural architecture in order to map dictionary definitions to representations of the words defined by those definitions. Specifically, they proposed two neural architectures for mapping the definition of word t to its word embedding $\mathbf{e}_t$. Let $\mathcal{D}_t$ be the sequence of words in t's definition, i.e., $\mathcal{D}_t = \{w_1, w_2, \ldots, w_n\}$, with their corresponding embeddings $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. The two models differ in the way they process $\mathcal{D}_t$. In the bag-of-words (**BoW**) model, $\mathcal{D}_t$ is taken as a bag

Proceedings of NAACL-HLT 2019, pages 2151–2156
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

of words, i.e., the representation of the definition is encoded by adding the word embeddings of all its content words, i.e., $\sum_{i=1}^{n} \mathbf{v}_i$. The model learns, using a fully-connected layer, a matrix for transforming the encoded representation to the target word's embedding $\mathbf{e}_t$. The BoW model is not sensitive to the order of words in $\mathcal{D}_t$. This might be crucial for an accurate semantic understanding. The Recurrent Neural Network (**RNN**) model alleviates this issue by encoding the input sequence using an LSTM architecture (Hochreiter and Schmidhuber, 1997). Similarly to the BoW model, a dense layer maps the encoded representation to the target word's embedding.

In both cases, the goal is to map a given definition to the corresponding target word's embedding $\mathbf{e}_t$, computed using Word2vec (Mikolov et al., 2013) and independently from the training of the main model. Two cost functions were tested: (1) the **cosine** distance between the estimated point in the target space ($\hat{\mathbf{e}}_t$) and $\mathbf{e}_t$, and (2) the **rank** loss which contrast the choice of $\mathbf{e}_t$ with a random choice for a randomly-selected word from the vocabulary other than t.

The reverse dictionary system takes advantage of a standard architecture which has proven effective in various NLP tasks. However, similarly to many other word-based models, the system ignores that the same word can have multiple (potentially unrelated) meanings. In fact, it tries to map multiple definitions, with different semantics, to the same point in the target space. For instance, the three semantically unrelated definitions of *crane*: "lifts and moves heavy objects", "large long-necked wading bird", and "a small constellation in the southern hemisphere" will have similar semantic interpretation by the system. This word-level meaning conflation can hamper the ability of the system in learning an accurate mapping function. In what follows in this paper, we will illustrate how a simple sense level distinction can facilitate a more accurate semantic understanding for the reverse dictionary system, hence leading to significant performance improvements.

2.2 Sense Integration

Let t be an ambiguous word with three meanings; hence, three distinct definitions $\mathcal{D}_{t_1}$, $\mathcal{D}_{t_2}$, and $\mathcal{D}_{t_3}$. The original model of Hill et al. (2016) maps all these definitions to $\mathbf{e}_t$. We mitigate the meaning conflation deficiency through a sense-specific mapping function that obtains distinct interpretations for individual definition, hence mapping them to different points in the target space: $\mathbf{s}_{t_1}$, $\mathbf{s}_{t_2}$, and $\mathbf{s}_{t_3}$. Specifically, in our experiments we leveraged DeConf (Pilehvar and Collier, 2016). DeConf is a WordNet-based sense representation technique which receives a set of pre-trained word embeddings and generates embeddings for individual word senses in the same semantic space, hence generating a combined space of words and word senses.

DeConf performs a set of random walks on WordNet's semantic network and extracts for each sense a set of *sense biasing* words $\mathcal{B}_\mathbf{s}$. A sense biasing word for the i^{th} meaning of a target word t is a semantically related word to that specific sense of the word ($\mathbf{s}_{t_i}$). For each word sense in WordNet we obtain the corresponding $\mathcal{B}_\mathbf{s}$. Then, the embedding for a specific word sense $\mathbf{s}$ is computed as:

$$\mathbf{s} = || \, \mathbf{e}_w + \alpha \sum_{b \in \mathcal{B}_\mathbf{s}} \exp(-\delta_i) \, \mathbf{e}_b ||, \qquad (1)$$

where δ is a decay parameter and $\mathbf{e}_w$ is the embedding of corresponding lemma of sense $\mathbf{s}$. In our experiments, as for word embeddings we used the 300-dimensional Word2vec embeddings, trained on the Google News corpus.[1] The same set was used as input to DeConf. As a result of this process, around 207K additional word senses were introduced in the space for the 155K unique words in WordNet 3.0.

2.2.1 Supersenses

It is widely acknowledged that sense distinctions in WordNet inventory are too fine-grained for most NLP applications (Hovy et al., 2013). For instance, for the noun *star*, WordNet 3.0 lists eight senses, among which two *celestial body* senses (as an "astronomical object" and that "visible, as a point of light, from the Earth"), and three *person* senses ("skillful person", "lead actor", and "performing artist"). This fine level of sense distinction is often more than that required by the target downstream application (Rüd et al., 2011; Severyn et al., 2013; Flekova and Gurevych, 2016). In our experiments, we used WordNet's lexicographer files (lexnames[2]) in order to reduce sense granularity. Created by the curators of WordNet

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

[2] https://wordnet.princeton.edu/man/lexnames.5WN.html

			WN-seen		WN-unseen		Concept Mapping	
			top-10	*top-100*	*top-10*	*top-100*	*top-10*	*top-100*
Supersense	RNN	cosine	0.656	0.824	0.150	0.310	0.230	0.480
		ranking	0.694	0.836	0.162	0.352	0.335	0.630
	BoW	cosine	0.642	0.820	0.250	0.416	0.280	0.590
		ranking	0.706	**0.872**	**0.310**	**0.474**	**0.390**	**0.735**
Sense	RNN	cosine	**0.742**	0.854	0.164	0.336	0.275	0.505
		ranking	0.668	0.840	0.180	0.372	0.325	0.615
	BoW	cosine	0.678	0.826	0.290	0.456	0.300	0.620
		ranking	0.692	0.848	0.292	0.470	0.380	**0.735**
Word	RNN	cosine	0.462	0.652	0.056	0.162	0.215	0.400
		ranking	0.534	0.728	0.086	0.188	0.190	0.475
	BoW	cosine	0.446	0.652	0.136	0.264	0.175	0.465
		ranking	0.562	0.740	0.160	0.292	0.320	0.600
Baseline	-	-	0.104	0.346	0.054	0.158	0.065	0.300

Table 1: Accuracy performance (@10/100) of the original (word-based) reverse dictionary system and its sense- and supersense-based improvements on different datasets. See Section 2.1 for system configurations.

during its development, these files organize WordNet synsets into 45 groups (such as food, animal, event, and emotion) according to their syntactic and logical properties. These groupings are usually referred to as *supersenses*. Using supersenses, the *celestial* and *person* meanings of *star* are grouped into two main groups. A supersense embedding e_{ss} in our experiments is simply computed as a normalized average (centroid) of its contained sense embeddings, i.e., $e_{ss} = ||\sum_{s \in ss} e_s||$. This reduces the average number of senses for polysemous words in WordNet from 2.9 to 1.8.

3 Experiments

We carried out evaluations on the three reverse dictionary datasets created by Hill et al. (2016): WordNet definitions and "single-sentence descriptions" written for a set of frequent words (*concept mapping*). They proposed two different versions of the WordNet dataset: *WN-seen*, in which a test instance is already observed during training, and *WN-unseen*, in which test instances are excluded from the training data. The former dataset is targeted at evaluating the ability of the system to recall a previously encoded information.

We experimented with three variants of the reverse dictionary system: the original word-based model and the two proposed sense-based variants, based on WordNet senses and supersenses.[3]

Table 1 reports accuracy performance for four different configurations of the system (BoW and RNN definition composition and cosine and ranking loss; cf. Section 2.1) on the three datasets. In the last row, we also report results for the unsupervised baseline of Hill et al. (2016) which adds the embedding of words in the input definition and finds the nearest embedding in the target space.

Results reported in the Table clearly highlight that addressing the meaning conflation deficiency in the system has led to significant performance improvements (word vs. sense and supersense settings). This is observed consistently across all the three datasets and for both sense-based models. The better semantic understanding of the system is reflected by its better recall of seen test instances (WN-seen) and better generalisation to unseen and out-of-domain data (WN-unseen and concept mapping). The absolute top-10 accuracy improvements of the ranking-BoW supersense model over the best corresponding word-based configurations are: 14.4% (WN-seen), 15% (WN-unseen), and 7% (concept mapping).

Among the two proposed systems, supersenses prove to be more effective, suggesting that the fine-grained sense distinctions in WordNet might not be necessary for an accurate reverse dic-

[3]The experiments are based on the implementation available at https://github.com/fh295/DefGen2.

tionary mapping, corroborating previous findings (Flekova and Gurevych, 2016). Our results are also in line with the findings of Hill et al. (2016) that the reverse dictionary system performs best with the bag-of-words (BoW) input encoding and the ranking loss. One of the fundamental differences between the two input encodings lies in their sensitiveness to order: RNNs are sensitive to the order of words in a given sequence whereas permuting words in the sequence does not alter BoW's encoding. Hill et al. (2016) suggested that it is often possible to retrieve a concept even if the words in its corresponding definition are shuffled. This can partly explain the strikingly good relative performance of the BoW model.

4 Analysis

During our analysis of system outputs, we observed many examples in which the word-based model was unable to retrieve an ambiguous word since the definition was referring to one of its less frequent meanings. For instance, the word *dressing* might refer to different concepts such as "getting dressed" or "savory dressing for salads". Having a conflated understanding of *dressing*, the word-based model was unable to retrieve the salad meaning.

dressing *savory dressings for salads; basically of two kinds: either the thin french or vinaigrette type or the creamy mayonnaise type*
word: mayonnaise, marinade, sauce **sense:** *dressing*, mayonnaise, mayo **baseline:** or, either, type

Other similar examples include infrequent senses of *party*, defined as "an organization to gain political power", and *partition*, defined as "a vertical structure that divides or separates". In both cases, the sense-based model improves the original word-based one, in which the system is unable to retrieve the intended word. Numerous such examples were observed during our analysis of the results, highlighting the important limitation of word-based models for their inherent bias towards more frequent usages.

Moreover, as a side benefit, sense embeddings provide parts of speech distinction, unlike common pre-trained word embeddings which conflate all parts of speech to a single token. For instance, the word-based model is unable to recall the nominal *bear* because it has a conflated understanding of the word which includes all its senses, particularly the dominant verb meaning.[4]

bear *massive plantigrade carnivorous or omnivorous mammals with long shaggy coats and strong claws*
word: critter, rabbit, squirrel, wolf **sense:** *bear*, mustelid, bruin **baseline:** carnivorous, omnivorous.

The same applies to the "open land" meaning of *common*, which is a less frequent (nominal) meaning of the word which is usually used as an adjective for concepts such as "ordinary" or "usual".

common *a piece of open land for recreational use in an urban area*
word: park, plaza, entryway, courtway **sense:** park, green, *common* **baseline:** for, area, in, recreational

Additionally, word embeddings are insensitive to fine-grained semantic distinctions, such as antonymy, due to their construction nature. However, the sense representations used in our experiments (DeConf) were constructed by exploiting the knowledge encoded in WordNet. Hence, they benefit from the rich semantic and ontological knowledge provided by the resource (such as relation types). Some of the improvements can be attributed to this property of sense embeddings.

unanticipated *not anticipated*
word: unavoidable, inevitable, plausible **sense:** unforeseen, *unanticipated*, unpredicted **baseline:** not, anticipated, expected

However, there are cases in which the word-based model provided more accurate results. For instance:

service *work done by one person or group that benefits another*
word: *service*, caring **sense:** organisation, dependant, programme

Our analysis showed that most of these errors were due to fine-grained sense distinctions in WordNet or obscure meanings. For instance, one of the senses[5] of *organisation* is semantically re-

[4] In our analysis, we found that improvements are mostly due to addressing semantic conflation rather than ambiguities in parts of speech.

[5] The 6^{th} sense of *organisation* in WordNet 3.0, defined as "the activity or result of distributing or disposing persons or things properly or methodically".

lated (also close in WordNet's graph) to the meaning of *service* in the example. This would suggest the need for more accurate sense representations and highlight the fact that the fine-granularity of senses should be better adjusted to the underlying task. Moreover, it corroborates our finding that the coarse-grained supersenses are more suitable in the task of reverse dictionary mapping. We leave the experiments with other sense representation techniques to future work.

5 Related work

Sense representations address the meaning conflation deficiency of their word-based counterparts by computing distinct representations for individual meanings of words, usually referred to as word senses. Sense distinctions might be given by an external sense inventory, such as WordNet (Fellbaum, 1998). An inventory-based sense representation technique exploits the knowledge encoded in the resource to construct representations (Rothe and Schütze, 2015; Jauhar et al., 2015; Pilehvar and Collier, 2016). Alternatively, senses can be automatically induced in an unsupervised manner by analyzing the diversity of contexts in which a word appears (Schütze, 1998; Reisinger and Mooney, 2010; Huang et al., 2012; Neelakantan et al., 2014; Guo et al., 2014; Šuster et al., 2016).

Regardless of how senses are obtained, the integration of sense representations into NLP systems is not a straightforward process. Hence, they have often been evaluated on artificial tasks such as word similarity. This is also due to lack of suitable evaluation benchmarks for sense representation techniques. Pilehvar and Camacho-Collados (2019) recently proposed a dataset, The Word-in-Context (WiC), which provides a challenging, yet reliable, benchmark for the purpose.

Few attempts have been made at the integration of sense representation into downstream applications. Li and Jurafsky (2015) experimented with unsupervised sense representations in tasks such as part-of-speech tagging and named entity recognition, with mixed results. Also related to our work are the proposals of Flekova and Gurevych (2016) and Pilehvar et al. (2017) to disambiguate the input text and replace word embeddings with sense embeddings for the intended senses. Our results for supersenses corroborates the findings of Pilehvar et al. (2017) who found reducing fine-granularity of senses beneficial to some settings.

A more recent branch of research investigates the construction of dynamic word embeddings that can adapt according to the context in which they appear (Peters et al., 2018; Devlin et al., 2018). One of the objectives of this research has been to bypass the integration difficulties of sense representations into downstream models. These so-called contextualised word embeddings can easily be replaced with conventional static word embeddings in neural-based NLP systems. This integration has proven beneficial to a wide range of NLP applications. Pilehvar and Camacho-Collados (2019) carried out an analysis on the sense distinguishing capability of contextualised embeddings, showing that, despite their successful application to downstream applications, these embeddings are not very powerful in capturing distinct meanings of words.

6 Conclusions

We provided an analysis on the impact of addressing the meaning conflation deficiency of word embeddings on the performance of a downstream NLP application, i.e., reverse dictionary mapping. Through a set of experiments we showed that a simple migration from words to senses can significantly improve the ability of the system in semantic understanding, leading to consistent performance boost. In future work, we plan to evaluate sense integration in other NLP applications, such as Machine Translation, in the light of (Liu et al., 2018), and question answering.

References

R. Brown and D. McNeill. 1966. The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5:325–337.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research* 63:743–788.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

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

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, pages 2029–2041.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*. pages 497–507.

Felix Hill, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics* 4:17–30.

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

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence* 194:2–27.

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 ACL*. Jeju Island, Korea, pages 873–882.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL*. Denver, Colorado, pages 683–693.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of EMNLP*. Lisbon, Portugal, pages 683–693.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling Homographs in Neural Machine Translation. In *Proceedings of NAACL*. New Orleans, LA, USA.

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

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*. Doha, Qatar, pages 1059–1069.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*. New Orleans, LA, USA.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream nlp applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 1857–1869.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of EMNLP*. Austin, TX, pages 1680–1690.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of ACL*. pages 109–117.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*. Beijing, China, pages 1793–1803.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: Using search engines for robust cross-domain named entity recognition. In *Proceedings of ACL-HLT*. Portland, Oregon, USA, pages 965–975.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–124.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Learning semantic textual similarity with structural representations. In *Proceedings of ACL (2)*. Sofia, Bulgaria, pages 714–718.

Simon Šuster, Ivan Titov, and Gertjan van Noord. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*. San Diego, California, pages 1346–1356.

Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations: From an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*. ElectricDict '04, pages 29–35.