Improved Semantic Representation for Domain-Specific Entities

Mohammad Taher Pilehvar and Nigel Collier
Language Technology Lab
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{mp792,nhc30}@cam.ac.uk

Abstract

Most existing corpus-based approaches to semantic representation suffer from inaccurate modeling of domain-specific lexical items which either have low frequencies or are non-existent in open-domain corpora. We put forward a technique that improves word embeddings in specific domains by first transforming a given lexical item to a sorted list of representative words and then modeling the item by combining the embeddings of these words. Our experiments show that the proposed technique can significantly improve some of the recent word embedding techniques while modeling a set of lexical items in the biomedical domain, i.e., phenotypes.

1 Introduction

Semantic representation is one of the oldest, yet most active, research areas in Natural Language Processing (NLP) owing to the central role it plays in many applications (Pilehvar andNavigli, 2015). The field has experienced a resurgence of interest in recent years with the introduction of low-dimensional continuous space models that leverage neural networks for learning semantic representations. Word2vec (Mikolov et al., 2013) is a good example which despite its recent invention has found its way prominently into literature, mainly thanks to its ability to be quickly and effectively trained on large amounts of text.

However, since most of these corpus-based techniques base their representation only on the co-occurrence statistics derived from text corpora, they fall short of effectively modeling lexical items for which not many statistical clues can be obtained from the underlying corpus. Several attempts have been made to improve word embeddings with the help of knowledge derived from other resources (Yu and Dredze, 2014; Bian et al., 2014; Faruqui et al., 2015) or by including arbitrary contexts in the training process (Levy and Goldberg, 2014). However, most of these techniques still suffer from another deficiency of word embeddings that they inherit from their count-based ancestors: they conflate the different meanings of a word into a single vector representation. Attempts have been made to tackle the meaning conflation issue of word-level representations. A series of approaches cluster the context of a word prior to representation (Reisinger and Mooney, 2010; Huang et al., 2012; Neelakantan et al., 2014) whereas others exploit lexical knowledge bases for sense-specific information (Rothe andSchütze, 2015; Chen et al., 2014; Iacobacci et al., 2015; Camacho-Collados et al., 2015).

We propose a model that addresses both these issues through a mapping of a lexical item to a sorted list of representative words that brings about two advantages. Firstly, it pinpoints with an inherent disambiguation the meaning of the given lexical item at a deeper semantic level. Secondly, by casting the representation of the item as that of a set of potentially more frequent words, our approach can provide a more reliable representation of domain-specific items based on significantly more statistical knowledge. Our experiments show that the proposed model can provide a considerable improvement over some of the state-of-the-art word embedding approaches in a semantic similarity-based task.

Data. The final goal of this paper is to improve the semantic representation of domain-specific terms and phrases which usually have low frequencies (or are non-existent) in open-domain corpora and hence have a lower chance of being effectively modeled by existing word representation
techniques. Therefore, for our experiments we re-
trieved terms and phrases from a domain-specific
ontology in the biomedical domain. Specifically,
as our dataset in the experiments we opted for Hu-
man Phenotype Ontology (Sebastian Khler, 2014, 
HPO) which is a standardized vocabulary of phe-
notypic abnormalities encountered in human dis-
ease. Semantic modeling of phenotypes has sev-
aral applications in the biomedical domain such as profiling heritable diseases or understanding the
genetic origins of diseases (Collier et al., 2013).

2 Improved Semantic Representation

In this section we explain how our technique builds on top of pre-trained word embeddings to provide a more accurate semantic representation.

2.1 Disambiguation

As mentioned in the Introduction, one of the draw-
backs of word-level representations is that they conflate different meanings of a word into a sin-
gle vector. Our technique constructs a more accu-
rate semantic representation of a lexical item by constraining its semantics through a set of rele-
vant words. Interestingly, we achieve this on the basis of the same set of word-level representa-
tions. To this end, we first disambiguate the con-
tent word(s) in a given lexical item. In our experi-
ments, we used Babelfy (Moro et al., 2014) which 
is a state-of-the-art WSD system based on the Ba-
elNet sense inventory. Babelfy is a merger of 
Wikipedia and WordNet, among other resources 
(Navigli and Ponzetto, 2012). Let \( t \) be the phrase we are interested in 
disambiguation. The disambiguation phase transforms 
the phrase to three BabelNet concepts correspond-
ing to the intended meanings of the content words 
\{flexion, contracture, and digit\}. Disambiguating 
with respect to BabelNet provides us with an addi-
tional benefit: it links a content word to the corre-
sponding Wikipedia page of its intended meaning, 
giving us the chance to draw additional context for 
 improving its representation.

2.2 Representative list

Let the set of disambiguated concepts for a lex-
ical item \( t \) be \( C_t \). We further enrich this set by adding all the BabelNet concepts that have a se-
manic link (in the semantic network of BabelNet) 
to any of the concepts in \( C_t \). Let the enriched 
set of concepts be \( C_t^* \). Our goal here is to map 
\( C_t^* \) to a set of most relevant words that can repre-
sent its semantics. We achieve this by exploiting 
the fact that these concepts are linked to relevant 
Wikipedia articles. Let \( D_t \) be the set of Wikipedia 
articles retrieved for \( t \) (i.e., the set of articles that 
are associated with the concepts in \( C_t^* \)). We ana-
lyze the textual content of these articles by lever-
aging the method proposed by Camacho-Collados 
et al. (2015) and retrieve a sorted list of salient 
words. Specifically, we use lexical specificity and contrast word frequency statistics between \( D_t \) and 
all articles in Wikipedia. Lexical specificity (La-
fon, 1980) is a statistical measure based on the 
hypergeometric distribution which can be used to 
compute the semantic importance of an arbitrary 
vocabulary word \( w \) for \( D_t \) as:

\[
Spec(H; h; G; g) = -\log_{10} P(X \geq g) \quad (1)
\]

where \( H \) and \( h \) are the respective aggregate fre-
cuencies of all words in all Wikipedia articles and 
\( D_t \), and \( G \) and \( g \) are the respective frequencies of 
\( w \) in all Wikipedia articles and \( D_t \). For a given 
lexical item \( t \), we construct the set of semantically 
representative words \( R_t \) by keeping the words that 
are relevant to \( D_t \) with a minimum confidence of 
99% according to the hypergeometric distribution, 
i.e., \( P(X \geq 0.01) \).

For our example phenotype flexion contracture 
of digit, the representative list \( R_t \) comprises of 
around 1300 weighted words, with the top ones 
being muscle, finger, spasticity, toe, hand, pa-
tient, and spastic. Please note that our technique 
mapped an ambiguous term digit to a set of more semantically constrained keywords such as finger, 
toe, and hand. This enables us to construct a 
sense-specific representation of the word by lever-
aging word-level representations.

2.3 Vector construction

So far, we mapped a given lexical item \( t \) to a set of 
relevant concepts \( C_t^* \) and obtained for this set the 
sorted list \( R_t = \{r_1, ..., r_m\} \) of the most seman-
tically representative words. The final step is to 
construct a vector representation \( V_t \) for \( t \). We do 
this by combining the vectors for the words in \( R_t \). 
Let \( \mathcal{V}(x) \) be the vector representation given by a 
model such as Word2vec for the word \( x \). We com-
pute the weight for the \( i^{th} \) dimension of the vector 
\( V_t \), i.e., \( v_i \), as:

\[
v_i = \sum_{j=1}^{m} e^{-\lambda j} \mathcal{V}(r_j)_i \quad i = 1, ..., n \quad (2)
\]
where $V(r_j)_i$ is the weight of the $j^{th}$ dimension of the base vector for the $i^{th}$ word in $R_t$ and $e^{-\lambda i}$ is a decay function (with the decay constant $\lambda$) that gives more importance to the higher ranking terms in $R_t$. In our experiments, we did not perform a tuning on the value of $\lambda$ which was set to $\frac{1}{5}$. Please note that the dimensionality of $V_t$ is identical to that of the base word representations, i.e., $n$. Table 1 shows the top-3 most similar phenotypes for four phenotypes in the HPO ontology when Word2vec was used as the base representation.

### 3 Experiments

We evaluate our model in the semantic representation of phenotypes in the HPO ontology.

#### 3.1 Dataset

As of February 2016, the HPO ontology comprises of 11,591 phenotypic abnormalities. Each of these concepts is provided with a title (with an average length of four words) and about 35% of all these concepts are associated with synonymous titles (by average, each of these concepts has 1.94 synonyms). For example, Keratoconjunctivitis sicca is a phenotype for which three synonymous titles are provided by the ontology: Dry eye syndrome, Keratitis sicca, and Xerophthalmia.

#### 3.2 Tasks

Based on the ontological structure of HPO, we propose two tasks in the framework of semantic similarity measurement.

**Synonym identification.** Let $\mathcal{P}$ be the set of all phenotypes in the HPO ontology. Let $\mathcal{P}^* = \{p_1, ..., p_k\} (\subset \mathcal{P})$ be the subset of $k$ phenotypes for which at least one synonymous phenotype is provided in HPO and $s_{p_i} = \{s_{p_i}^1, ..., s_{p_i}^l\}$ be the set of $l$ synonymous phenotypes for phenotype $p_i$. Given a $s_{p_i}$, the task here is simply to identify the corresponding phenotype (i.e., $p_i$). In other words, the system has to identify the set of synonymous phenotypes to a given phenotype. Specifically, we compare the representation of $s_{p_i}$ with those of all the phenotypes in $\mathcal{P}$, obtaining a sorted list of most similar phenotypes. Ideally, the concept containing the synonymous title should appear at the top of this list. The higher the rank of $p_i$, the better has the system captured the semantics of the phenotypes. For this task we have 7193 synonymous titles ($\sum_{i=1}^{s} |s_{p_i}|$) that are to be matched with their corresponding phenotypes (among a total of 11,591 phenotypes).

**Hypernym identification.** Similarly to the previous experiment, a system’s task here is to identify the hypernym of a given phenotype. The aim of this experiment is to have a broader evaluation that can also cover all those concepts that do not provide synonymous titles (the dataset comprises of 11,590 phenotypes that have a hypernym).

#### 3.3 Baselines

As baseline, we benchmark our improved representations against Word2vec. We use the 300-dimensional vectors trained on the Google News corpus (about 100B tokens). We also report results for the Word2vec vectors when retrofitted using the approach of Faruqui et al. (2015) to the Paraphrase Database (Ganitkevitch et al., 2013, PPDB) and SNOMED-CT\(^1\). The latter is a comprehensive clinical terminology from which we extracted 108K synonymous sets, each comprising an average of 2.7 synonyms. We also compare our representations against the 300-dimensional GloVe vectors (Pennington et al., 2014) trained on the Wikipedia 2014 + Gigaword 5 corpus (6B tokens).

We were also interested in verifying how Word2vec and GloVe would perform if trained on

---

\(^1\)https://www.nlm.nih.gov/snomed/
Table 2: Evaluation results for the synonym identification task. We report mean and median rank (lower better) and the percentage of phenotypes for which the rank was equal to one (first match; higher better).

| System | Description | Mean rank | Median rank | First match |
|--------|-------------|-----------|-------------|-------------|
| Word2vec | Trained on open-domain data (Google News) | 1343.6 | 11 | 22% |
| Word2vec (2nd order) | | 664.1 | 6 | 28% |
| Word2vec | Trained on in-domain data (PubMed) | 224.1 | 4 | 32% |
| Word2vec (2nd order) | | 198.2 | 3 | 36% |
| GloVe | Trained on open-domain data (Wikipedia + Gigaword) | 1326.4 | 9 | 24% |
| GloVe (2nd order) | | 673.5 | 6 | 28% |
| GloVe | Trained on in-domain data (PubMed) | 701.4 | 4 | 34% |
| GloVe (2nd order) | | 493.5 | 3 | 36% |
| Word2vec | Trained on Google News, retrofitted to PPDB | 1357.4 | 8 | 26% |
| Word2vec | Trained on Google News, retrofitted to SNOMED-CT | 1346.2 | 9 | 25% |
| Random baseline | Random selection of the synonymous phenotype | 5473.0 | 5473.0 | 0% |

3.4 Results and discussion

Table 2 shows the evaluation results. We report mean and median rank of the target phenotype in the sorted list of most semantically similar phenotypes as well as the percentage of target phenotypes for which this rank was equal to one, i.e., the synonymous title was computed as the most similar item (first match in the table). As a reference, we also report the performance of a baseline which randomly picks the target phenotype.

We can see that a considerable performance improvement was gained when our technique was used for improving Word2vec and GloVe representations trained on open-domain corpora. Interestingly, even when the vectors were trained on an in-domain corpus (PubMed) that covers a large portion of the phenotypes with high frequencies, our model was still able to provide statistically significant improvements according to mean rank over the vanilla Word2vec and GloVe models, which left around 5% of all the phenotypes with no representation. The token coverage raised to 91% when the two models were trained on PubMed, resulting in the generation of representations for 99.7% of all phenotypes. In this setting, the respective relative mean rank improvements of 11.4% and 29.7% of our approach with respect to Word2vec and GloVe should be attributed to the additional semantic information that our model introduces to the vectors as well as the more accurate representation of concepts, thanks to the disambiguation phase and the semantically constraining keywords.

For the hypernym identification task we observed a very similar trend where our model improved Word2vec and GloVe from the respective mean ranks of 1034.1 and 1021.5 to 606.1 and 556.7 on the open-domain corpus and from 317.2 and 424.9 to 277.6 and 309.5 on PubMed.

4 Conclusions and future work

We proposed an approach for enhancing the representation capability of existing word modeling techniques in specific domains and showed that consistent improvement can be gained over Word2vec and GloVe even when they are trained on domain-specific corpora. We plan to enhance our technique by making it sensitive to syntax and different parts of speech, such as in the manner of Baroni and Zamparelli (2010). We also plan to carry out a deeper analysis to better understand...
the potential of our model and to identify places in which it can be improved.

Acknowledgments

The authors gratefully acknowledge the support of the MRC grant No. MR/M025160/1 for PheneBank.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In Proceedings of EMNLP, pages 1183–1193, Cambridge, Massachusetts.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, pages 132–148, Nancy, France.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Unified Multilingual Semantic Representation of Concepts. In Proceedings of ACL-IJCNLP, pages 741–751, Beijing, China.

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

Nigel Collier, Anika Oellrich, and Tudor Groza. 2013. Toward knowledge support for analysis and interpretation of complex traits. Genome Biology, 14(9):1–11.

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 NAACL-HLT, pages 1606–1615, Denver, Colorado.

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

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, pages 873–882, Jeju Island, Korea.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In Proceedings of ACL-IJCNLP, pages 95–105, Beijing, China.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1:127–165.