A Hybrid Distributional and Knowledge-based Model of Lexical Semantics

Nikolaos Aletras
Department of Computer Science,
University College London,
Gower Street,
London WC1E 6BT
United Kingdom
nikos.aletras@gmail.com

Mark Stevenson
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello,
Sheffield S1 4DP
United Kingdom
mark.stevenson@sheffield.ac.uk

Abstract

A range of approaches to the representation of lexical semantics have been explored within Computational Linguistics. Two of the most popular are distributional and knowledge-based models. This paper proposes hybrid models of lexical semantics that combine the advantages of these two approaches. Our models provide robust representations of synonymous words derived from WordNet. We also make use of WordNet's hierarcy to refine the synset vectors. The models are evaluated on two widely explored tasks involving lexical semantics: lexical similarity and Word Sense Disambiguation. The hybrid models are found to perform better than standard distributional models and have the additional benefit of modelling polysemy.

1 Introduction

The representation of lexical semantics is a core problem in Computational Linguistics and a variety of approaches have been developed. Two of the most widely explored have been knowledge-based and distributional semantics.

Knowledge-based approaches make use of some external information source which defines the set of possible meanings for each lexical item. The most widely used information source is WordNet (Fellbaum, 1998), although other resources, such as Machine Readable Dictionaries, thesaurii and ontologies have also been used (see Navigli (2009)).

One advantage of these resources is that they represent the various possible meanings of lexical items which makes it straightforward to identify ones that are ambiguous. For example, these resources would include multiple meanings for the word *ball* including the 'event' and 'sports equipment' senses. However, the fact that there are multiple meanings associated with ambiguous lexical items can also be problematic since it may not be straightforward to identify which one is being used for an instance of an ambiguous word in text. This issue has lead to significant exploration of the problem of Word Sense Disambiguation (Ide and Véronis, 1998; Navigli, 2009).

More recently distributional semantics has become a popular approach to representing lexical semantics (Turney and Pantel, 2010; Erk, 2012). These approaches are based on the premise that the semantics of lexical items can be modelled by their context (Firth, 1957; Harris, 1985). Distributional semantic models have the advantages of being robust and straightforward to create from unannotated corpora. However, problems can arise when they are used to represent the semantics of polysemous words. Distributional semantic models are generally constructed by examining the context of lexical items in unannotated corpora. But for ambiguous words, like *ball*, it is not clear if a particular instance of the word in a corpus refers to the 'event', 'sports equipment' or another sense which can lead to the distributional semantic model becoming a mixture of different meanings without representing any of the meanings individually.

This paper proposes models that merge elements of distributional and knowledge-based approaches to lexical semantics and combines advantages of both techniques. A standard distributional semantic model

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 20–29,
Denver, Colorado, June 4–5, 2015.

is created from an unannotated corpus and then refined using WordNet. The resulting models can be viewed as enhanced distributional models that have been refined using the information from WordNet to reduce the problems caused by ambiguous terms when models are created. Alternatively, it can be used as a version of the WordNet hierarchy in which distributional semantic models are attached to synsets. Thereby creating a version of WordNet for which the appropriate synsets can be identified more easily for ambiguous lexical items that occur in text.

We evaluate our models on two standard tasks: lexical similarity and word sense disambiguation. Results show that the proposed hybrid models perform consistently better than traditional distributional semantic models.

The reminder of the paper is organised as follows. Section 2 describes our hybrid models which combine information from WordNet and a standard distributional semantic model. These models are augmented using Latent Semantic Analysis and Canonical Correlation Analysis. Sections 3 and 4 describe evaluation of the models on the word similarity and word sense disambiguation tasks. Related work is presented in Section 5 and conclusions in Section 6.

2 Semantic Models

First, we consider a standard distributional semantic space to represent words as vectors (Section 2.1). Then, we make use of the WordNet's clusters of synonyms and hierarchy in combination with the standard distributional space to build hybrid models (Section 2.2) which are augmented using Latent Semantic Analysis (Section 2.3) and Canonical Correlation Analysis (Section 2.4).

2.1 Distributional Model

We consider a semantic space D, as a word by context feature matrix, $L \times C$. Vector representations consist of context features C in a reference corpus. We made use of pre-computed publicly available vectors[1] optimised for word similarity tasks (Baroni et al., 2014). Word co-occurrence counts are extracted using a symmetric window of two words over a corpus of 2.8 billion tokens obtained by concatenating ukWaC, the English Wikipedia and the British National Corpus. Vectors are weighted using positive Pointwise Mutual Information and the set of context features consists of the top 300K most frequent words in the corpus.

2.2 Hybrid Models

2.2.1 Synset Distributional Model

We assume that making use of information about the structure of WordNet can reduce noise introduced in vectors of D due to polysemy. We make use of all noun and verb synsets (excluding numbers and compounds) that contain at least one of the words in L to create a vector-based synset representation, H. Where H is a synset by context feature matrix, i.e. $S \times C$. Each synset vector is generated by computing the centroid of its lemma vectors in S (i.e. the sum of the lemma's vectors normalised by the number of the lemmas in the synset). For example, the vector of the synset *car.n.01* is computed as the centroid of its lemma vectors, i.e. *car, auto, automobile, machine* and *motorcar* (see Figure 1).

2.2.2 Synset Rank Model

The Synset Distributional Model provides a vector representation for each synset in WordNet which is created using information about which lemmas share synset membership. An advantage of this approach is that vectors from multiple lemmas are combined to form the synset representation. However, a disadvantage is that many of these lemmas are polysemous and their vectors represent multiple senses, not just the one that is relevant to the synset. For example, in WordNet the lemma *machine* has several possible meanings, only one of which is a member of the synset *car.n.01*.

WordNet also contains information about the relations between synsets, in the form of the synset hierarchy, which can be exploited to re-weight the importance of context features for particular synsets. We employ a graph-based algorithm that makes use of the WordNet is-a hierarchy. The intuition behind this approach is that context features that are relevant to a given synset are likely to be shared by its neighbours in the hierarchy while those that are not relevant (i.e. have been introduced via an irrelevant sense of a synset member) will not be. The graph-based algorithm increases the weight of context features

[1] `http://clic.cimec.unitn.it/composes/semantic-vectors.html`

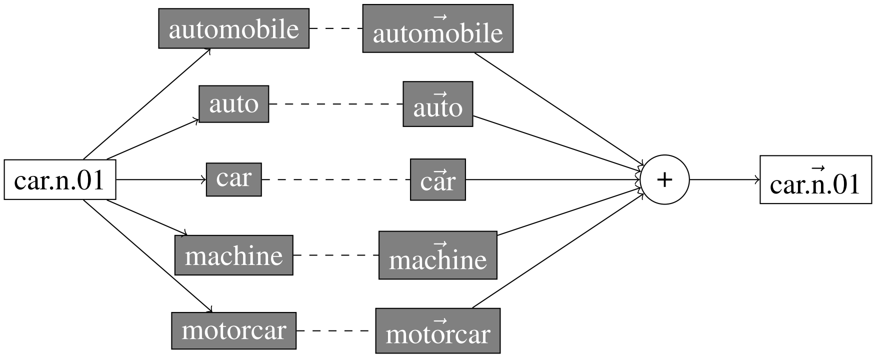

Figure 1: In the Synset Distributional Model the vector representing a synset (white box) is computed as the centroid of its lemma vectors (grey boxes)

that synsets share with neighbours and reduces those that are not shared.

PageRank (Page et al., 1999) is a graph-based algorithm for identifying important nodes in a graph that has been applied to a range of NLP tasks including word sense disambiguation (Agirre and Soroa, 2009) and keyword extraction (Mihalcea and Tarau, 2004).

Let $G = (V, E)$ be a graph with a set of vertices, V, denoting synsets and a set of edges, E, denoting links between synsets in the WordNet hierarchy. The PageRank score (Pr) over G for a synset (V_i) can be computed by the following equation:

$$Pr(V_i) = d \cdot \sum_{V_j \in I(V_i)} \frac{1}{O(V_j)} Pr(V_j) + (1-d)\mathbf{v} \quad (1)$$

where $I(V_i)$ denotes the in-degree of the vertex V_i and $O(V_j)$ is the out-degree of vertex V_j. d is the damping factor which is set to the default value of $d = 0.85$ (Page et al., 1999). In standard PageRank all elements of the vector $\mathbf{v}$ are the same, $\frac{1}{N}$ where N is the number of nodes in the graph.

Personalised PageRank (PPR) (Haveliwala et al., 2003) is a variant of the PageRank algorithm in which extra importance is assigned to certain vertices in the graph. This is achieved by adjusting the values of the vector $\mathbf{v}$ in equation 1 to prefer certain nodes. The values in $\mathbf{v}$ effectively initialises the graph and assigning high values to nodes in $\mathbf{v}$ makes them more likely to be assigned a high PPR score.

For each context feature c in C if $c \in LM$ where LM contains all the lemma names of synsets in S, we apply PPR to assign importance to synsets. The score of each synset S_c in the personalisation vector $\mathbf{v}$, is set to $\frac{1}{|S_c|}$ where $|S_c|$ is the number of synsets that context feature i belongs. The personalisation value of all the other sysnets is set to 0.

We apply PPR over WordNet for each context feature using UKB (Agirre et al., 2009) and obtain weights for each synset-context feature pair resulting to a new semantic space H_{p}, $S \times C$, where vector elements are weighted by PageRank values. Figure 2 shows how the synset scores are computed by applying PPR over WordNet given the context feature *car*. Note that we use the context features of the distributional model D.

2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997) has been used to reduce the dimensionality of semantic spaces leading to improved performance. LSA applies Singular Value Decomposition (SVD) to a matrix X, $W \times C$, which represents a distributional semantic space. This is a form of factor analysis where X is decomposed into three other matrices:

$$X = U \Sigma V^T \quad (2)$$

where U is a $W \times W$ matrix of row vectors where its columns are eigenvectors of XX^T, Σ is a diagonal $W \times C$ matrix containing the singular values and V is a $C \times C$ matrix of context feature vectors where its columns are eigenvectors of $X^T X$. The multiplication of the three component matrices results in the original matrix, X. Any matrix can be decomposed perfectly if the number of singular values is no smaller than the smallest dimension of X. When

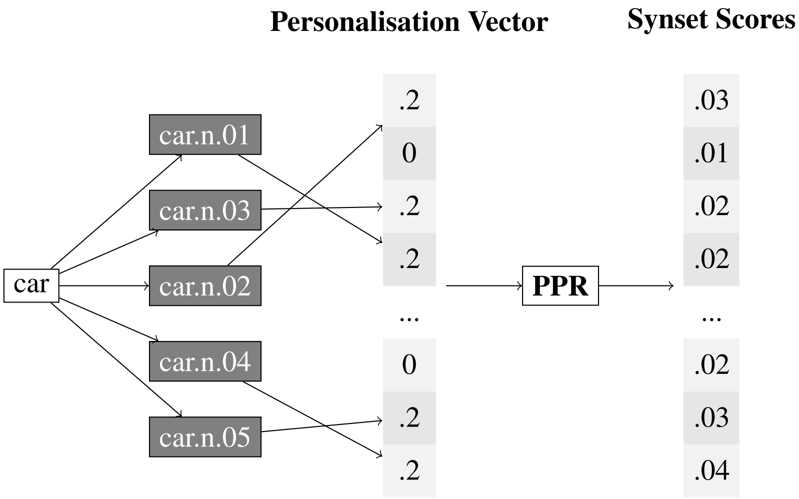

Figure 2: In the Synset Rank Model, each synset (grey boxes) is assigned with a score by computing PPR over WordNet. The personalisation vector (grey array) is initialised by assigning probabilities only to the synsets that include the context feature as a lemma name.

fewer singular values are used then the matrix product is an approximation of the original matrix. LSA reduces the dimensionality of the SVD by deleting coefficients in the diagonal matrix Σ starting with the smallest. The approximation of matrix X retaining the K largest singular values, $\tilde{X}$, is then given by:

$$\tilde{X} \approx U_K \Sigma_K V_K^T \tag{3}$$

where U_K is a $W \times K$ matrix of word vectors, Σ_K is a $K \times K$ diagonal matrix with singular values and V_K is a $K \times C$ matrix of context feature vectors.

We apply LSA on the Synset Distributional Model, H and the Synset Rank model, H_{p} to obtained the reduced semantic spaces $\tilde{H}$ and $\tilde{H}_{\text{p}}$ respectively.

2.4 Joint Representation using CCA

Recent work has demonstrated that distributional models can benefit from combining alternative views of data (see Section 5). H and H_{p} provide two different views of the synsets and we incorporate evidence from both to learn a joint representation using Canonical Correlation Analysis (CCA) (Hardoon et al., 2004). Given two multidimensional variables $\mathbf{x}$ and $\mathbf{y}$, CCA finds two projection vectors by maximising the correlations of the variables onto these projections. The function to be maximised is:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} \tag{4}$$

The dimensionality of the projection vectors is lower or equal to the dimensionality of the original variables.

The computation of CCA directly over H and H_{p} is computationally infeasible because of their high dimensionality (300K). We apply CCA over the reduced spaces learned using LSA, $\tilde{H}$ and $\tilde{H}_{\text{p}}$ to obtain two joint semantic spaces following a similar approach to Faruqui and Dyer (2014). These are the spaces H^*, resulting from the projection of the Synset Distributional Model $\tilde{H}$, and H_{p}^*, resulting from the projection of the Synset Rank Model $\tilde{H}_{\text{p}}$.

3 Word Similarity

3.1 Computing Similarity

Since hybrid models represent words as synset vectors, similarity between two words can be computed following two ways. First, we compute similarity between two words as the **maximum** of their pairwise synset similarity. On the other hand, similarity can be computed as the **average** pairwise synset similarity using the synsets that the two words belong. Similarity is computed as the cosine of the angle between word or synset vectors.

3.2 Data

We make use of six standard data sets that have been widely used for evaluating lexical similarity and relat-

Max						
Model	**WS-353**	**WS-Sim**	**WS-Rel**	**RG**	**MC**	**MEN**
Distributional Model						
D	0.62	0.70	**0.59**	0.79	0.72	**0.72**
Hybrid Models - Full						
H	0.49	0.60	0.36	0.69	0.64	0.58
H_{p}	0.58	0.67	0.49	0.82	**0.86**	0.63
Hybrid Models - LSA						
$\tilde{H}$	0.55	0.69	0.42	0.71	0.71	0.54
$\tilde{H}_{\text{p}}$	0.58	0.68	0.46	0.85	**0.86**	0.55
Hybrid Models - CCA						
H^*	**0.67**	**0.76**	0.57	0.81	0.79	**0.72**
H^*_{p}	0.52	0.62	0.41	**0.86**	0.80	0.56

Table 1: Spearman's correlation on various data sets. Maximum similarity between pairs of synsets.

edness. First, we make use of **WS-353** (Finkelstein et al., 2001) which contains 353 pairs of words annotated by humans. Furthermore, we make use of the similarity (**WS-Sim**) and relatedness (**WS-Rel**) pairs of words created by Agirre et al. (2009) from the original WS-353 data set.

We also made use of the **RG** (Rubenstein and Goodenough, 1965) and **MC** (Miller and Charles, 1991) data sets which contain 65 and 30 pairs of nouns respectively. Finally, we make use of the larger **MEN** data set (Bruni et al., 2012) which contains 3,000 pairs of words that has been used as image tags. Annotations are obtained using croudsourcing.

3.3 Model Parameters

The parameters we need to tune are the number of the top components in LSA spaces, $\tilde{H}$ and $\tilde{H}_{\text{p}}$, and CCA spaces, H^* and H^*_{p}. For the LSA spaces, we tune the number of the top k components in RG. We set $k \in \{50, 100, ..., 1000\}$ and select the value that maximises performance which is $k = 700$ for $\tilde{H}$ and $k = 650$ for $\tilde{H}_{\text{p}}$. For the joint spaces learned using CCA, we also tune the number of the top l correlated features in RG. We set $l \in \{10, 20, ..., 650\}$ and select the value that maximises performance which is $l = 250$ for H^* and $l = 40$ for H^*_{p}.

3.4 Evaluation Metric

Performance is measured as the correlation between the similarity scores returned by each proposed method and the human judgements. This is the standard approach to evaluate word and text similarity tasks, e.g. (Budanitsky and Hirst, 2001; Agirre et al., 2009; Agirre et al., 2012). Our experiments use Spearman's correlation coefficient.

3.5 Results

Table 1 shows the Spearman's correlation of similarity scores generated by each model and human judgements of similarity across various data sets by taking the maximum pairwise similarity score of two words' synsets. The first row of the table shows the results obtained by the word distributional model of Baroni et al. (2014). The full hybrid models H and H_{p} perform consistently worse than the original distributional model D across data sets. The main reason is that a large number of synsets contain only one lemma name which might be polysemous. For example, the only lemma name of the synsets 'ball.n.01' (*'round object that is hit or thrown or*

Average						
Model	**WS-353**	**WS-Sim**	**WS-Rel**	**RG**	**MC**	**MEN**
Distributional Model						
D	0.62	0.70	0.59	0.79	0.72	0.72
Hybrid Models - Full						
H	0.61	0.71	0.52	0.72	0.65	0.64
H_{p}	0.65	0.73	0.56	0.79	0.81	0.58
Hybrid Models - LSA						
$\tilde{H}$	0.59	0.70	0.48	0.68	0.68	0.63
$\tilde{H}_{\mathrm{p}}$	0.65	0.73	0.56	**0.81**	**0.86**	0.58
Hybrid Models - CCA						
H^*	**0.70**	**0.77**	**0.64**	0.78	0.84	**0.74**
H^*_{p}	0.61	0.69	0.52	0.72	0.76	0.62

Table 2: Spearman's correlation on various data sets. Average pairwise similarity between pairs of synsets.

kicked in games') and 'ball.n.04' (*'the people assembled at a lavish formal dance'*) is 'ball'. In this case, the synset vector in H and the lemma vector in D are identical and still polysemous. This problem does not hold in H_{p} and therefore the correlations are higher for that semantic space but still lower than those obtained for D. Applying LSA on H and H_{p} improves results but correlations are still lower than those obtained using D[2]. On the other hand, the joint space learned by applying CCA, H^*, produces consistently better similarity estimates than D while outperforms all the other models in the majority of the data sets. That confirms our main assumption than incorporating information obtained from a large corpus and a knowledge-base improves word vector representations.

Table 2 shows the Spearman's correlation of similarity scores generated by each model and human judgements of similarity across various data sets by taking the average pairwise similarity score of two words' synsets. Results show that using the average rather than the maximum system similarity improves results for almost all data sets. For example, the best hybrid model, H^*, achieves correlations that are between 2% and 12% than D for the majority of data sets, although performance is 1% lower for the RG data set. This improved performance suggest that human judgements of word similarity are based on the relation between all the senses of two given words rather than just the most similar ones.

4 Word Sense Disambiguation

4.1 Data

We test the efficiency of our hybrid models on the English All Words tasks of Senseval-2 (Palmer et al., 2001) and Senseval-3 (Snyder and Palmer, 2004), two standard data sets for evaluating WSD. Our experiments focus on the disambiguation of nouns in these data sets.

4.2 Word Sense Tagging

A simple approach to all-words WSD was implemented in which each sense of an ambiguous word is compared against its context and the most similar chosen.

For example suppose that we want to disambiguate

[2] Note that Baroni et al. (2014) found that applying SVD to D did not improve performance over using the full space.

Nouns	**Senseval-2**		**Senseval-3**	
	Precision	Recall	Precision	Recall
Hybrid Models - Full				
H	0.46	0.45	0.37	0.36
H_{p}	**0.65**	**0.63**	**0.50**	**0.48**
Hybrid Models - LSA				
$\tilde{H}$	0.45	0.44	0.39	0.37
$\tilde{H}_{\text{p}}$	0.60	0.58	0.46	0.45
Hybrid Models - CCA				
H^*	0.44	0.43	0.36	0.34
H^*_{p}	0.61	0.60	0.48	0.46

Table 3: Results obtained by hybrid models on SenseEval-2 and SenseEval-3 data sets (nouns only).

the word *bank* in the following sentence:

"Banks provide payment services."

Assume that the word *bank* consists of two senses *'bank.n.01'* and *bank.n.02* defined as *'sloping land (especially the slope beside a body of water)'* and *"a financial institution that accepts deposits and channels the money into lending activities'* respectively.

First we consider the vectors of all the possible noun synsets containing the word *bank* as a synset name. Then for each context word (*provide*, *payment* and *service*) that exists in our semantics spaces we compute a centroid vector from its constituent senses. Finally, we compute a context vector for the entire context by summing up all the context word vectors. We select the synset of the target word that its vector has the highest cosine similarity to the context vector.

4.3 Model Parameters

The parameters we need to tune are the same as for the word similarity task and we use the best settings obtained for that task. We also experimented with varying the number of surrounding sentences used as context by testing values between ± 1 and ± 4. The best performance was obtained using a context created from the sentence containing the target word and ± 1 sentences surrounding it.

4.4 Evaluation Metrics

Word sense disambiguation systems are evaluated by computing precision and recall. Precision measures the proportion of disambiguated words that have been correctly assigned with a sense. Recall measures the proportion of words disambiguated correctly out of all words available for disambiguation.

4.5 Results

Table 3 shows the results obtained by using our hybrid models on the two word sense disambiguation data sets. The full Synset Rank model H_p is consistently better method in terms of precision and recall in both data sets. On the other hand, it is somewhat surprising that dimensionality reduction and integration of semantic spaces do not help in improving performance. That is the $\tilde{H}_p$ and H^*_p models achieve lower precision and recall than the fuller H_p.

The Synset Distributional models H, $\tilde{H}$ and H^* consistently fail to perform well. The difference in precision and recall compared to the Synset Rank models is between 12% and 19%. This suggests that the knowledge-based weighting of the context features generates less noisy vectors for sense tagging.

The pattern of results observed for the WSD task is somewhat different to those obtained for word similarity, where applying LSA and CCA improved performance (see Section 3). The most likely explanation of this difference is that WSD requires the model to represent the possible senses of each ambiguous word. It is also important that these senses correspond to the ones used in the relevant lexicon (WordNet in this case). The Synset Rank model H_p does this by making use of information from WordNet. However, these synset representations are disrupted by LSA and CCA which compress the semantic space by extracting general features from them. This is not a problem for word similarity since there is no need to model the senses found in the lexicon.

5 Related Work

Dealing with polysemy in distributional semantics is a fundamental issue since the various senses of a word type are conflated in a single vector. Previous work tackled the problem through vector adaptation, clustering and language models (Erk, 2012). Vector adaptation methods modify a traditional (i.e. polysemous) target word vector by applying pointwise operations such as addition or multiplication to that and the surrounding words in a sentence (Mitchell and Lapata, 2008; Erk and Padó, 2008; Thater et al., 2011; Van de Cruys et al., 2011). Alternatively, clustering methods have been used to cluster together the different contexts a target word appears assuming that each cluster of contexts captures a different sense of the target word (Dinu and Lapata, 2010; Erk and Pado, 2010; Reisinger and Mooney, 2010). Language models have also been used to remove polysemy from word vectors by predicting words that could replace the target word given a context (Deschacht and Moens, 2009; Washtell, 2010; Moon and Erk, 2013). More recently, Polajnar and Clark (2014) applied context selection and normalisation to improve the quality of word vectors. Our hybrid models are related to the vector adaptation methods since we modify the synset vectors using its lemmas' vectors to remove noise.

Our work is also inspired by recent work on improving classic distributional vector representations of words by incorporating information from different modalities. For example, researchers have developed methods that make use of both visual and contextual information to improve word vectors (Bruni et al., 2011; Silberer et al., 2013; Lazaridou et al., 2014). Following a similar direction, Faruqui and Dyer (2014) found that learning joint spaces from multilingual vector spaces using CCA improves the performance of standard monolingual vector spaces on semantic similarity. Fyshe et al. (2014) showed that integrating textual vector space models with brain activation data when people are reading words achieves better correlation to behavioural data than models of one modality.

Our hybrid models are also closely related to a supervised method proposed by Faruqui et al. (2015). Their method refines distributional semantic models using relational information from various semantic lexicons, including WordNet, by making linked words in these lexicons to have similar vector representations. While our models are also based on using information from WordNet for refining vector representations, they are fundamentally different. They create synset vectors in an unsupervised fashion and more importantly can be used for sense tagging.

6 Conclusions

This paper proposed hybrid models of lexical semantics that combine distributional and knowledge-based approaches and offer advantages of both techniques. A standard distributional semantic model is created from an unannotated corpus and then refined by (1) using WordNet synsets to create synset vectors; and (2) applying a graph-based technique over WordNet to reweight synset vectors. The resulting hybrid models can be viewed as enhanced distributional models using the information from WordNet to reduce the problems caused by ambiguous terms when models are created. Results show that our models perform better than traditional distributional models on lexical similarity tasks. Unlike standard distributional approaches the techniques proposed here also model polysemy and can be used to carry out word sense disambiguation.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 33–41, Athens, Greece.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '09)*, pages 19–27, Boulder, Colorado.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. In *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics (GEMS '11)*, pages 22–32, Edinburgh, UK, July.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145, Jeju Island, Korea.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the workshop on "WordNet and other Lexical Resources" at the Second Annual Meeting of the North American Association for Computational Linguistics*, Pittsburgh, PA.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the Latent Words Language Model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 21–29, Singapore.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, MA.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii.

Katrin Erk and Sebastian Pado. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of NAACL*, Denver, Colorado.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, London.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, New York, USA. ACM Press.

J. Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*.

Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2014. Interpretable semantic vectors from a joint model of brain- and text- based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 489–499, Baltimore, Maryland.

David Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Z. Harris. 1985. Distributional structure. In J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, New York.

Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003-35, Stanford InfoLab.

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

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

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pages 404–411, Barcelona, Spain.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Taesun Moon and Katrin Erk. 2013. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):42.

Roberto Navigli. 2009. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24, Toulouse, France.

Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238, Gothenburg, Sweden. Association for Computational Linguistics.

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.

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

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand.

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

Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1022, Edinburgh, Scotland, UK.

Justin Washtell. 2010. Expectation vectors: A semiotics inspired approach to geometric lexical-semantic representation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics (GEMS)*, pages 45–50, Uppsala, Sweden.