Cognitively Motivated Distributional Representations of Meaning

Elias Iosif[1,2], **Spiros Georgiladakis**[1], **and Alexandros Potamianos**[1,2]
[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]"Athena" Research and Innovation Center, Greece
iosife@central.ntua.gr, sgeorgil@central.ntua.gr, potam@central.ntua.gr

Abstract

Although meaning is at the core of human cognition, state-of-the-art distributional semantic models (DSMs) are often agnostic to the findings in the area of semantic cognition. In this work, we present a novel type of DSMs motivated by the dual–processing cognitive perspective that is triggered by lexico–semantic activations in the short–term human memory. The proposed model is shown to perform better than state-of-the-art models for computing semantic similarity between words. The fusion of different types of DSMs is also investigated achieving results that are comparable or better than the state-of-the-art. The used corpora along with a set of tools, as well as large repositories of vectorial word representations are made publicly available for four languages (English, German, Italian, and Greek).

Keywords: distributional semantic models, activation based cognitive models, semantic similarity, multilinguality

1. Introduction

Distributional semantic models (DSMs) aim to represent the meaning of lexical entities by encoding linguistic features extracted from text corpora. Word-level representations are the basis for more complex tasks such as phrase– and sentence–level representation and similarity computation (Mitchell and Lapata, 2010; Agirre et al., 2012). Related applications include paraphrase detection (Androutsopoulos and Malakasiotis, 2010), affective text analysis (Malandrakis et al., 2013), etc.

Adopting the standpoint of (Potamianos, 2014) we argue that the aforementioned tasks should be driven by the fundamental properties of the human cognitive system. Indeed, empirical evidence provided by cognitive sciences and psycholinguistics constitute a fertile source of inspiration for DSM design. In this spirit, we propose a novel type of word–level DSMs motivated by the dual–processing cognitive perspective (Kahneman, 2013) that is triggered by lexico–semantic activations in the short–term human memory. The proposed model is shown to perform better than state-of-the-art models for the problem of word similarity computation with respect to various datasets in four languages (English, German, Italian, and Greek). The fusion of different types of DSMs is also investigated exceeding the performance of individual models. We make publicly available the used corpora along with a set of tools, as well as large repositories of vectorial word representations for all four languages.

2. Related Work

Word-level representations constitute the core aspect of DSMs typically constructed from co-occurrence statistics of word tuples. Word-level DSMs can be broadly categorized into unstructured and structured with respect to the extraction of contextual features. The bag-of-words model is the most widely used approach (e.g., see (Grefenstette, 1994)), lacking however some desirable characteristics such as "order sensitivity" (Turney, 2012). Unlike unstructured models, the order of extracted features is taken into account in the framework of structured DSMs via the exploitation of syntactic relationships (e.g., argument structures and modifications) (Pado and Lapata, 2007). A comparison of unstructured and structured DSMs was conducted in (Agirre et al., 2009) for the task of semantic similarity estimation between words, where slightly higher performance was achieved by unstructured DSMs. The multiple senses of words are typically not directly encoded in DSMs. For addressing this issue, exemplar models were proposed where the meaning of a word was represented by a set of stereotypical corpus sentences instead of a single feature vector (Erk and Padó, 2010). Recently, the computation of contextual features was posed in a learning-based framework where the goal is to estimate the context in which the words of interest are expected to occur (Bengio et al., 2003; Huang et al., 2012; Mikolov et al., 2013a). This advancement is discussed in (Baroni et al., 2014), where it is compared with traditional DSMs for various tasks of lexical semantics. A similar comparative study is also presented in (Huang et al., 2012).

Word-level representations are the building blocks for phrase- and sentence-level models. Such models are motivated by the principle of semantic compositionality (Pelletier, 1994), while they are meant to address a number of properties that characterize the compositional aspects of meaning, namely, "linguistic creativity", "order sensitivity", "adaptive capacity", and "information scalability" (Turney, 2012). The work presented in (Mitchell and Lapata, 2010) constitutes one

of the earlier approaches in the area of compositional DSMs, where a series of algebraic operations are applied over the word-level feature vectors for representing the semantics of two-word phrases divided in three categories, namely, noun-noun (NN), adjective-noun (AN), and verb-object (VO). In (Baroni and Zamparelli, 2010), the compositional aspects of AN phrases were investigated where the semantic transformation triggered by adjectives was modeled via a function (implemented as matrix) operating over the vectorial representation of nouns. The phrase-level approach of (Baroni and Zamparelli, 2010) was generalized in (Socher et al., 2012) according to which sentence-level representations are constructed in a recursive bottom-up fashion. A comparison of the major compositional approaches reported in the literature is presented in (Dinu et al., 2013). In (Turney, 2012), the compositionality task is investigated in the space of similarities instead of features. Specifically, two distinct models are built, namely, domain and function spaces.

Despite the fact that meaning is at the core of human cognition, incorporating findings from the area of semantic cognition into the design of DSMs is not straightforward. The relation between cognitive semantics and the information encoded by DSMs is discussed in (Lenci, 2008). Motivated by evidence related to the low-dimensional geometry of thought (Gärdenfors, 2004) in (Athanasopoulou et al., 2014) word-level semantics are locally represented via a series of manifolds, while global operations (e.g., computation of word semantic similarity) are decomposed into local (i.e., manifold-level) operations. In (Georgiladakis et al., 2015) a two-layer model motivated by semantic priming (McNamara, 2005) and the dual cognitive processing theory (Kahneman, 2013) was adopted for representing the semantics of compositional phrases. Other related approaches include the mapping of DSMs to feature-normed spaces (Făgărăşan et al., 2015), the extension of word embeddings with cognitively motivated lexico-semantic resources such as WordNet (Rothe and Schütze, 2015).

3. Distributional Semantic Models

Here, two types of DSMs are described, which rely on the distributional hypothesis of meaning (Harris, 1954). Following the terminology coined in (Baroni et al., 2014) we refer to them as *context-counting DSMs* and *context-predicting DSMs*.

Context-counting DSMs (CDSMs). This type can be regarded as the mainstream paradigm of DSMs. The core idea is that the representation of word meaning is implemented by considering the context in which the word occurs. A context window of size $2H+1$ words is centered on the word of interest w_i and the contextual features that fall within the window are extracted. A high-dimensional space is built –also known as vector space model (VSM)– where the value of each dimension is set according to the co-occurrence counts of words and their contextual features. The transformation of raw co-occurrence counts was found to improve the properties (e.g., smoothness) and performance of DSMs. Such transformations aim to encode feature saliency, e.g., positive point–wise mutual information (PPMI) (Bullinaria and Levy, 2007). High-dimensional word–feature matrices constitute the mainstream implementation of VSM. The dimensionality of such representations can be reduced by applying techniques such as Singular Value Decomposition (SVD). A survey of CDSMs is provided in (Turney and Pantel, 2010).

Context-predicting DSMs (PDSMs). PDSMs have been recently proposed. Their key difference compared to CDSMs deal with the computation of contextual features (Bengio et al., 2003; Huang et al., 2012; Mikolov et al., 2013a). The basic operation in CDSMs, is the counting of contextual features for a target word w_i given a context. The key idea that underlies PDSMs is to formulate the distributional hypothesis of meaning in a probabilistic framework and learn the model parameters. For example, in (Mikolov et al., 2013a) this is investigated in the framework of statistical language modeling where the word–feature conditional probabilities are learnt (word embeddings).

For both CDSMs and PDSMs the cosine of feature vectors is the most widely–used metric for computing word semantic similarity. In (Lin, 1998; Turney and Pantel, 2010), more metrics are discussed.

4. Activation-based DSMs (ADSMs)

In this section, cognitively motivated DSMs are proposed, which are termed *activation-based DSMs*. The underlying hypothesis is that the representation of a target word w_i takes the form of a set of semantically related attributes (words) that are activated when w_i is processed (Rogers and McClelland, 2004). This hypothesis is justified by the psycholinguistic phenomenon of *semantic priming* according to which the presence of a word facilitates the cognitive processing of another word (Collins and Loftus, 1975; McNamara, 2005). Given a geometrical representation where the words of a vocabulary V are arranged according to their respective pairwise semantic distances (i.e., a space G), the activated attributes of w_i can be regarded as a sub-space G_i (also referred to as *relative activation area*) that represents the semantics of w_i. Such areas are expected to be of small size (compared to V) given the limited capacity[1] of human short–term memory (Cowan, 2001). A number of cognitive factors (e.g., semantic concreteness (Barber et

[1]The capacity should not be regarded as fixed, e.g., it was empirically found to be task–depended (Cowan, 2001).

al., 2013), visuospatial characteristics (Yao et al., 2013)) play a role to the size of activation areas that varies[2] across words.

In this work, we advance some of the design principles of the aforementioned CDSMs and PDSMs by putting into a computational framework the core implications of activation areas. The notion of the maximum activation area $G_i^{\max}$ of word w_i is central to the proposed model. $G_i^{\max}$ is defined as the list of all words with which w_i co-occurs in our corpus. What we attempt to models here is the largest possible list of words that w_i can form "direct" cognitive connections with. The activation area G_i for word w_i is defined as a fraction of the maximum activation area $G_i^{\max}$ for this word.

The two main novel aspects of the proposed model – which are not directly encoded in CDSMs and PDSMs– are:

1. The injection of sparsity in word representations via the exploitation of small activation areas G_i,

2. The conditioning of sparsity on the maximum activation area $G_i^{\max}$. The size of activations is not the same for all words.

Each activation area is formulated as a vectorial representation enabling the construction of VSM on top of which well-established similarity metrics, e.g., cosine of vectors, can be applied. In particular, the words of the entire vocabulary V are used for defining a $|V|$–dimensional vector for a target word w_i. Only those dimensions that correspond to the words included in G_i are allowed to take non–zero values. An advantageous characteristic of this construction is the representational fusion of ADSMs with existing DSMs (see next section), since VSM constitutes the most common implementation in DSMs. Overall, the proposed model is a two-tier system meant for activation and similarity computation adopting the paradigm presented in (Iosif and Potamianos, 2015). In the first tier, the activation area G_i of target word w_i is computed by identifying and filtering the most similar vocabulary words to w_i according to a similarity metric. The second layer is used for the computation of semantic similarity of word pairs based on their respective activations.

The theoretical foundations of the two-tier architecture originate from a generic cognitive model that was empirically found to apply to numerous semantic and behavioral tasks (Kahneman, 2013). Given a stimulous (e.g., a word or any other perceptual entity), the first layer rapidly activates (at the expense of accuracy) a group of similar/related entities. The refinement of relations that exist between the stimulous and the activated entities is passed to a second layer with slower but more accurate responses.

5. Fusion of Models

Two fusion schemes are proposed next: 1) an early fusion scheme where representations of target words are concatenated into a single hybrid representation and 2) a late fusion scheme where similarity scores estimated by different models are linearly combined.

Feature-level fusion (early). Given a word w_i, let $v_{i,l}$ be its semantic representation computed via model l. Assuming that this representation can be encoded as a vector, and that M models are available, the hybrid representation of w_i, denoted as $\hat{v}_i$, can be computed as:

$$\hat{v}_i = v_{i,1} \oplus v_{i,2} \oplus \cdots \oplus v_{i,M}, \quad (1)$$

where $\oplus$ stands for the vector–concatenate operator. Then, similarity between two words, w_1 and w_2, can be estimated by using their hybrid representations $\hat{v}_1$ and $\hat{v}_2$ and applying standard similarity metrics, e.g., cosine of vectors. Since the vectors in (1) come from different models, normalization should take place before applying the $\oplus$ operations. In this work, we propose the use of Z normalization[3] (Cohen, 1995):

$$v'_{i,l} = \frac{v_{i,l} - \mu_{i,l}}{\sigma_{i,l}}, \quad (2)$$

where $\mu_{i,l}$ and $\sigma_{i,l}$ are the arithmetic mean and the standard deviation of feature values of vector $v_{i,l}$, respectively.

Similarity-level fusion (late). The similarity between two words, w_1 and w_2, denoted as $\hat{S}(w_1, w_2)$, is estimated as follows:

$$\hat{S}(w_1, w_2) = \lambda_0 + \sum_{j=1}^{M} \lambda_j \; S_j(w_1, w_2), \quad (3)$$

where $S_j(w_1, w_2)$ stands for the similarity between w_1 and w_2 computed via model j, while λ_j is are trainable weights.

6. Experimental Settings

Corpora creation. For each language a web-harvested corpus was created according to the following procedure. Starting from a dictionary, a web search query was formulated for each dictionary entry and sent to the Yahoo! search engine. For each query, up to the 1000 top ranked results (document snippets) were retrieved. The overall corpus was built by aggregating the retrieved snippets.

Context-counting DSMs (CDSMs). Several window sizes ($H = 1, \ldots, 5$) were applied for extracting the

[2] This issue is also considered in (Georgiladakis et al., 2015) for the compositional representation of phrases.

[3] Other normalization schemes, e.g., min–max, yielded significantly lower performance.

contextual features, which were weighted according to PPMI. The dimensions of the word–feature matrix were reduced by applying SVD. Here, we report results for 300 dimensions for which the best results were obtained.

Context-predicting DSMs (PDSMs). The `word2vec` tool was used (Mikolov et al., 2013a; Mikolov et al., 2013c). We applied the CBOW approach as being more computationally efficient (Mikolov et al., 2013b; Baroni et al., 2014) for $H = 1,\ldots,5$. The dimensions of the resulting the word–feature matrix was set to 300.

Activation-based DSMs (ADSMs). First, for a target word, w_i, the size of $G_i^{\max}$ (denoted as $N_i^{\max}$) was estimated as the number of the vocabulary words that co-occur with w_i within sentence boundaries. Next, the G_i area (i.e., relative activation area) was computed by selecting the $\alpha N_i^{\max}$ most semantically similar[4] words to w_i, for $\alpha = 1...100\%$. For the computation of the feature vector of w_i in the second layer, we experimented with several context windows $H = 1,\ldots,5$. For a given H, only those words that were included in G_i were considered as valid features weighted according to the respective similarity scores (i.e., those used for G_i). A low–dimensional variation of ADSMs was also implemented by applying SVD over the word–feature matrix retaining 300 dimensions.

The λ weights of (3) were computed using least squares estimation[5]. For all the aforementioned DSMs (CDSMs, PDSMs, ADSMs), the cosine of feature vectors was used for computing word similarity.

7. Evaluation Datasets and Results

The task of noun semantic similarity computation was used for evaluation purposes. We used the datasets presented in Table 1, which deal with four languages, namely: English, German, Italian, and Greek. The Spearman correlation coefficient against human ratings was used as the evaluation metric. The correlation scores[6] obtained by ADSMs are presented in Figure 1(a) and (b) for $H = 1$ and $H = 5$, respectively,[7] as a function of the size of relative activation areas. It is observed that, when a narrow window size is used ($H = 1$) the best performance is obtained for activation sizes greater than 50%. For $H = 5$ the highest correlation is achieved for activation size $\alpha = 3 - 10\%$, except for the case of the Greek dataset for which the best results are obtained for activation size $\alpha = 20\%$. Overall, the highest correlation scores are achieved for the English datasets for all window sizes. The low-dimensional ADSMs (SVD-based) were found to yield significantly lower performance than ADSMs across all datasets, languages, and H values.

Language	Dataset	# pairs
English (EN)	ws353 (Finkelstein et al., 2002)	353
English (EN)	men1K men3K (Bruni et al., 2014)	1000 3000
German (GE)	ws350ge (Leviant and Reichart, 2015)	350
Italian (IT)	ws350it (Leviant and Reichart, 2015)	350
Greek (GR)	gr200 (Zervanou et al., 2014)	200

Table 1: Evaluation datasets.

In Table 2, we present the performance for a number of relative activation sizes (for $H = 5$) with respect to similarity and relatedness using the respective subsets of word pairs from the English dataset ws353 (Agirre et al., 2009). The dominant relation for highly similar words is synomymy, e.g., "midday-noon". Regarding relatedness, various relations of associative nature may hold, e.g., "news-report". The highest correlation score is obtained for the case of similarity. For both relation types the best scores are achieved for activation size $\alpha = 3\%$. The performance[8] of early fusion (computed according to (1) and (2)) is shown in Table 3 along with the performance of the three individual models (i.e., CDSMs, PDSMs, and ADSMs). Context window $H = 5$ was applied for all fused models, while the 10% of the activation size was utilized for ADSMs. In addition, the performance is reported in terms of average correlation considering the correlation scores obtained for all datasets (ws353, men1K, men3K, ws350ge, ws350it, and gr200). We observe that the best results are achieved by the fusion scheme across all datasets and languages (on average, 0.72 correlation). For the majority of datasets the correlation obtained by the best individual model is im-

Type	Relative activation size			
	1%	3%	10%	20%
Similarity	0.63	**0.77**	0.74	0.68
Relatedness	0.59	0.72	0.70	0.62

Table 2: Performance wrt similarity vs. relatedness.

[4] In this work, we used PPMI, however, any similarity metric, e.g., cosine, can be applied.

[5] Very similar results were obtained when applying other methods, e.g., ridge regression.

[6] Results are shown only for men3K, since almost identical performance was observed for both men3K and men1K.

[7] Results for $1 < H < 5$ are omitted as being the middle ground of $H = 1$ and $H = 5$.

[8] For the results reported next, no distinction is made regrading the relations (e.g., similarity vs. relatedness) of the experimental word pairs.

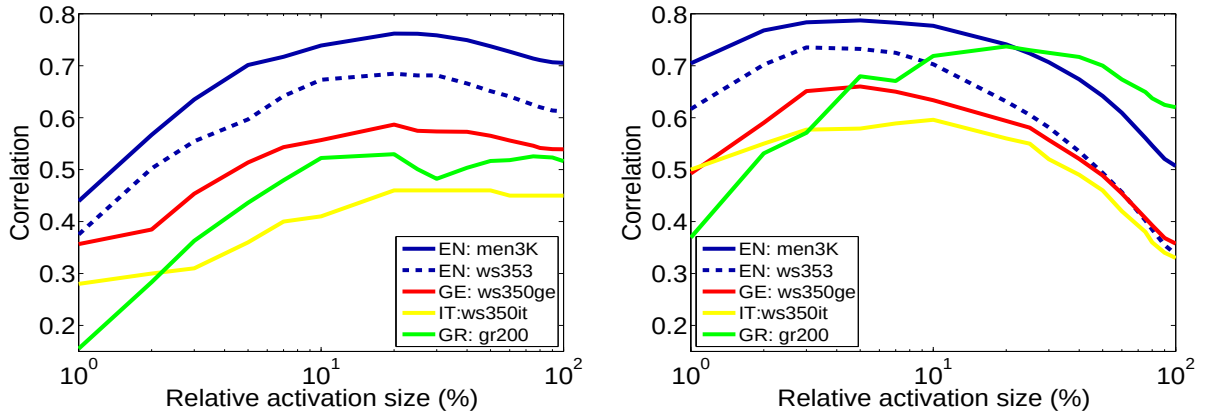

Figure 1: Performance of activation-based DSMs. Correlation as a function of the % of the activation areas for context window sizes: (a) $H=1$, (b) $H=5$. Results are shown for English (EN), German (GE), Italian (IT), and Greek (GR).

Language	Dataset	Type of DSMs			
		Context-counting (CDSMs)	Context-predicting (PDSMs)	Activation-based (ADSMs)	Early fusion
EN	ws353	0.70	0.68	0.70	**0.74**
EN	men1K	0.76	0.74	0.78	**0.80**
EN	men3K	0.76	0.75	0.78	**0.80**
GE	ws350ge	0.56	0.60	0.63	**0.66**
IT	ws350it	0.55	0.50	0.60	**0.62**
GR	gr200	0.58	0.53	**0.72**	**0.72**
Average		0.65	0.63	0.70	**0.72**

Table 3: Correlation scores for feature-level fusion (early): CDSMs, PDSMs and ADSMs ($H=5$, $\alpha=10\%$).

proved by 0.02-0.03 via fusion, with the exception of the gr200 dataset where no improvement is shown over the ADSM performance. Regarding the individual models, the proposed ADSMs are shown to yield consistently higher performance compared to CDSMs and PDSMs across all languages. Also, CDSMs perform slightly better than PDSMs for all languages except German.

The performance obtained via the late fusion scheme (computed according to (3)) is presented in Table 4 along with the performance of the three individual models (i.e., CDSMs, PDSMs, and ADSMs). As in the case of early fusion, context window $H=5$ was applied for all fused models, while the 10% of the activation size was used for ADSMs. For the case of English datasets (ws353 and men1K), we used a subset of the largest dataset (men3K) for training purposes (i.e., learning the λ weights of (3)) by excluding the pairs of ws353 and men1K. For the case of German, Italian, and Greek datasets (ws350ge, ws350it, and gr200, respectively) the entire men3K English dataset was used for training. We observe that the late fusion improves the performance only in the monolingual train/test setting, i.e., when English is used in both train and test datasets. In this case, the highest improvement is observed for the ws353 dataset (from 0.70 to 0.75). Overall, the achieved performance is comparable or better compared to the state-of-the-art results reported in the literature: 0.75 for ws353 (Baroni et al., 2014), 0.60 and 0.57 for ws350ge and ws350it[9] (Leviant and Reichart, 2015), and 0.55 for gr200 (Zervanou et al., 2014).

8. Corpora, Tools, and Resources

The English, German, Italian, and Greek corpora are provided that consist of 116, 84, ,83, and 66 millions of web document snippets, respectively. The first tool constitutes an implementation of CDSMs characterized by fast corpus indexing, while the second tool implements the proposed ADSMs. For each language, a pool of vectorial word representations is made available, computed by the aforementioned DSMs (i.e., CDSMs, PDSMs, ADSMs)

[9]The relative difference in performance between the three ws353-based datasets may be attributed to factors related to the translation of the English version to German and Italian.

Train		Test		Type of DSMs			
Lang	Dataset	Lang	Dataset	Context-counting (CDSMs)	Context-predicting (PDSMs)	Activation-based (ADSMs)	Late fusion
EN	menTrain	EN	ws353	0.70	0.68	0.70	**0.75**
EN	menTrain	EN	men1K	0.76	0.74	0.78	**0.79**
EN	men3K	GE	ws350ge	0.56	0.60	**0.63**	0.61
EN	men3K	IT	ws350it	0.55	0.50	**0.60**	0.43
EN	men3K	GR	gr200	0.58	0.53	**0.72**	0.62

Table 4: Correlation scores for similarity-level (late) fusion: CDSMs, PDSMs and ADSMs ($H = 5$, $\alpha = 10\%$).

for $H = 1, 3, 5$. For each combination, 100K vectors are provided for a total of 1.2 millions of vectors.

9. Conclusions

The proposed activation-based DSMs were found to perform better than the mainstream context-counting DSMs, as well as the recently proposed context-predicting DSMs. This was observed for all four languages and datasets. Small activation sizes (i.e., $3 - 10\%$) contain adequate information for building semantic representations in the framework of activation-based DSMs. In addition, it was shown that feature-level fusion performs better than late fusion at the level of similarity scores. A key operation regarding the fusion of representations is their Z normalization that was found to perform best for this task exceeding the performance of the individual models. Late fusion improves the performance only when both train and test datasets are in the same language. Last but not least, multilingual resources (corpora and word semantic representations) and tools are made publicly available[10].

10. Acknowledgements

This work has been partially supported by the SpeDial project supported by the EU FP7 with grant # 611396, BabyRobot project supported by the EU Horizon 2020 Programme with grant # 687831, and the BabyAffect project supported by the Greek General Secretariat for Research and Technology with grant # 3610.

11. References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–27.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proc. of the Sixth International Workshop on Semantic Evaluation (SemEval)*, pages 385–393.

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Athanasopoulou, G., Iosif, E., and Potamianos, A. (2014). Low-dimensional manifold distributional semantic models. In *Proc. of International Conference on Computational Linguistics (COLING)*.

Barber, H. A., Otten, L. J., Kousta, S.-T., and Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, 125:47–53.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proc. of EMNLP*, pages 1183–1193.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of ACL*, pages 238–247.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

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

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, 39(3):510–526.

Cohen, P. R. (1995). *Empirical Methods for Artificial Intelligence*. MIT Press.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Cowan, N. (2001). The magical number 4 in short-term

[10] `http://www.telecom.tuc.gr/~iosife/downloads/adsms/actindex.html`

memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 21:87–114.

Dinu, G., Pham, N. T., and Baroni, M. (2013). General estimation and evaluation of compositional distributional semantic models. In *Proc. of Workshop on Continuous Vector Space Models and their Compositionality*.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

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

Făgărăşan, L., Vecchi, E. M., and Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proc. of International Conference on Computational Semantics (IWCS)*.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. MIT press.

Georgiladakis, S., Iosif, E., and Potamianos, A. (2015). Fusion of compositional network-based and lexical function distributional semantic models. In *Proc. of NAACL–HLT Workshop on Cognitive Modeling and Computational Linguistics*, pages 39–47.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Huang, E., Socher, R., Manning, C., and Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882.

Iosif, E. and Potamianos, A. (2015). Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering*, 21(01):49–79.

Kahneman, D. (2013). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Lenci, A. (2008). Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, 20:131.

Leviant, I. and Reichart, R. (2015). Separated by an uncommon language: Towards judgment language informed vector space modeling. In *Preprint pubslished on arXiv. arxiv:1508.00106*.

Lin, D. (1998). An information-theoretic definition of similarity.

Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2013). Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

McNamara, T. P. (2005). *Semantic priming: Perspectives from Memory and Word Recognition*. Psychology Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *http://arxiv.org/abs/1301.3781/*.

Mikolov, T., Le, Q., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. In *http://arxiv.org/abs/1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34:1388–1429.

Pado, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13:11–24.

Potamianos, A. (2014). Cognitive multimodal processing: from signal to behavior. In *Proc. of Workshop on Roadmapping the Future of Multimodal Interaction Research*.

Rogers, T. T. and McClelland, J. L. (2004). *Semantic Cognition. A Parallel Distributed Processing Approach*. The MIT Press.

Rothe, S. and Schütze, H. (2015). AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proc. of ACL and IJCNLP*, pages 1793–1803.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proc. of EMNLP-CoNLL*, pages 1201–1211.

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

Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Yao, B., Vasiljevic, M., Weick, M., Sereno, M. E., ODonnell, P., and Sereno, S. C. (2013). Semantic size of abstract concepts: It gets emotional when you can't see it. *PLoS ONE*, 8:47–53.

Zervanou, K., Iosif, E., and Potamianos, A. (2014). Word semantic similarity for morphologically rich languages. In *Proc. of Ninth International Conference on Language Resources and Evaluation (LREC)*.