A Computational Study on Word Meanings and Their Distributed Representations via Polymodal Embedding

Joohee Park*
Korea Advanced Institute
of Science and Technology
james.joohee.park@navercorp.com

Sung-hyon Myaeng
Korea Advanced Institute
of Science and Technology
myaeng@kaist.ac.kr

Abstract

A distributed representation has become a popular approach to capturing a word meaning. Besides its success and practical value, however, questions arise about the relationships between a true word meaning and its distributed representation. In this paper, we examine such a relationship via polymodal embedding approach inspired by the theory that humans tend to use diverse sources in developing a word meaning. The result suggests that the existing embeddings lack in capturing certain aspects of word meanings which can be significantly improved by the polymodal approach. Also, we show distinct characteristics of different types of words (e.g. concreteness) via computational studies. Finally, we show our proposed embedding method outperforms the baselines in the word similarity measure tasks and the hypernym prediction tasks.

1 Introduction

Word representations based on the distributional hypothesis of Harris (1954) have become a dominant approach including word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which show remarkable performances in a wide spectrum of natural language processing. However, a question arises about a relationship between a true word meaning and its distributed representation. While the context-driven word representations seem to be able to capture word-to-word relations, for example, *men* is to *women* as *king* is to *queen*, it still remains unclear what aspects of word meaning they capture and miss. For example, a word, *coffee*, can be understood from multiple perspectives. It may be associated with a ceramic cup filled with dark brown liquid from the perceptual perspective or an emotion such as happiness or tranquility. It may provoke other related concepts like *bagel* or *awakening*. We raise the question of how well the current distributed representation captures such aspects of word meanings.

In order to help answering this question, we propose a polymodal word representation based on the theory that humans tend to use diverse sources in developing a word meaning. In particular, we construct six modules for polymodality including linear context, syntactic context, visual perception, cognition, emotion, and sentiments based on the human cognitive model proposed by Maruish and Moses (2013). They are combined to build a single word representation.

We conduct a series of experiments to examine the relationships between word meanings and their distributed representations and compare the results with other representations such as word2vec, GloVe, and meta-embedding (Yin and Schütze, 2015). We attempt to understand how well the model capture the diverse aspects of word meanings via two experiments: the property norms analysis and the sentiment polarity analysis. The result suggests that the existing embedding methods lack in capturing visual properties and sentiment polarities and show that they can be much improved by adopting polymodal approaches.

Finally, we examine distinct characteristics of different types of words via computational studies, focusing along the dimension of concept concreteness and similarity. We find that the importance of a certain module (e.g. visual perception or lexical relations) varies depending on the word properties. Our study provides some computational evidence for the heterogeneous nature of word meanings,

* Currently at Search Solutions Inc., Seongnam, 13561, Korea

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 214–223,
Taipei, Taiwan, November 27 – December 1, 2017

which has been extensively studied in the field of psycholinguistics. We briefly introduce it in the following subsection.

2 Related Work

2.1 Theoretical works

Word meanings are thought to have diverse aspects. Steels (2008) address that languages are inherently built upon our cognitive system to fulfill the purpose of communication between mutually unobservable internal representations. So many psycholinguistic theories have attempted to understand the diverse nature of word meanings by human minds. Barsalou (1999) claims that many human modalities such as conceptual/perceptual systems cooperate with each other in a complex way and influence word meanings, while Pulvermüller (1999) argues that concepts are grounded in complex simulations of physical and introspective events, activating the frontal region of the brain that coordinates the multimodal information. Studies on semantic priming (Plaut and Booth, 2000) also supports them that words can be similar to each other in various ways to foster the priming effect. The experiments in this paper are designed to provide some computational evidence on such studies on the multifaceted nature of word meanings.

2.2 Multimodal approaches

From a computational point of view, there exist a number of bimodal approaches that extend the semantic representation to include perceptual information or understandings of the world around us. Bruni et al. (2014) and Kiros et al. (2014a) propose a way to augment text-based word embeddings using public image datasets while Roller and Im Walde (2013) integrate visual features into LDA models. A recent study on Image caption generation (Xu et al., 2015) suggests an interesting way to align word embeddings and image features. Moreover, Kiros et al. (2014b) jointly trains the image abstraction network and sentence abstraction network altogether, making the visual features naturally combined into word embeddings. Similar attempts have been made not only for visual perception but also auditory (Kiela and Clark, 2015) and olfactory (Kiela et al., 2015) perception. On the other hand, Henriksson (2015) demonstrates that semantic space ensemble models created by exploiting various corpora are able to outperform any single constituent model. Yin and Schütze (2015) propose meta-Embedding that ensembles multiple semantic spaces trained by different methods with different tasks such as word2vec, GloVe, HLBL (Luong et al., 2013) and C&W (Collobert and Weston, 2008). Above works succeed to improve word embedding quality by extending the semantic representation, but it still remains unclear how those improvements are related to the word meanings.

3 Polymodal word embedding

To embrace the multifaceted nature of word meanings, we propose a polymodal word embedding. More specifically, we take into account perception, sentiment, emotion, and cognition (lexical relation) derived from diverse sources, in addition to linear context and syntactic context obtained from the corpus. Note that the term *polymodal* is used to distinguish it from *bimodal* (Kiela, 2017). In bimodal approach, a single cognitive modality is used whereas more than one modalities are used in polymodal.

3.1 Modules

We describe each of the modules in detail.

Linear context refers to linear embeddings (Mikolov et al., 2013) comprising 300-dimensional vectors trained over 100 billion words from the Google News dataset using skip-gram and negative sampling.

Syntactic context takes a similar skip-gram approach as in linear context but defines the context window differently using a dependency parsing result (Levy and Goldberg, 2014). While the linear skip-gram defines the contexts of a target word w as $w_{-k}, w_{-k+1}, ..., w_{k-1}, w_k$ where k is a size of the window, syntactic context defines them as $(m_1, lbl_1), (m_2, lbl_2), ..., (m_k, lbl_k), (m_{-1}, lbl_{-1})$ where m is the modifiers of word w and lbl is the type of dependency relation.

Both linear and syntactic contexts are similar in the sense that they capture word characteristics from the corpus. However, the different definitions of the contexts make the model focus on the different aspects of word meanings. Levy and Goldberg (2014) report that linear context tends to capture topical similarity whereas syntactic context captures functional similarity. For example, the word *Florida* is close to *Miami* in linear context but close to *California* in syntactic context.

We harness both types of contexts to take into account functional and syntactic similarities.

Cognition (Lexical relation) encompasses all the relations between words, which are captured in the form of a lexicon or ontology in a cognitive system. In this paper, we mainly focus on synonym, hypernym and hyponym relations in WordNet (Miller, 1995) which contains 149k words and 935k relations between them. We train lexical-relation-specific word embedding using retro-fitting (Faruqui et al., 2015).

Specifically, let $V = \{w_1, ..., w_n\}$ be a vocabulary and Ω be an ontology that encodes semantic relations between words in V. Ω can be represented as a set of edges of undirected graph where $(w_i, w_j) \in \Omega$ if w_i and w_j holds semantic relationship of interest. The matrix $\hat{Q}$ is the collection of the vector representation of $\hat{q}_i \in \mathbb{R}^d$ for each word $w_i \in V$ where d is the length of pre-trained word vectors. In this experiment, we use GloVe as such vectors. The objective of learning is to train the matrix $Q = (q_1, ..., q_n)$ so as to make q_i close to its counterpart $\hat{q}_i$ and also to its adjacent vertices in Ω. Thus the objective function to be minimized can be written as

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i ||q_i - \hat{q}_i||^2 + \sum_{(i,j)\in\Omega} \beta_{ij} ||q_i - q_j||^2 \right]$$

where α_i and β_{ij} are hyperparameters. This procedure of training transforms the manifold of semantic space to make words in relations located more closer in Euclidean distance.

Perception is a vital component for human cognition and has a significant influence on word meanings. In this paper, we only consider visual perception. We jointly train the embeddings of images and sentences together into the multi-modal vector space to build vision-specific word embeddings (Kiros et al., 2014b).

In particular, let T be the training dataset where one image I_i is associated with a corresponding caption sentence S_i, i.e., $(I_i, S_i) \in T$. An embedding of image I_i, $x_i \in \mathbb{R}^d$, can be obtained through convolutional neural networks, in this case, 19-layer OxfordNet (Simonyan and Zisserman, 2014), where d is the size of the dimension of multimodal space. Similarly, an embedding of sentence S_i, $x_s \in \mathbb{R}^d$, can be composed through one of the sentence modeling networks, in this case, LSTM (Hochreiter and Schmidhuber, 1997). These two image and sentence modeling networks are jointly trained together to minimize the pairwise ranking loss function

$$L = \sum_{x_i} \sum_{x_{\hat{s}}} max(0, \alpha - x_i \cdot x_s + x_i \cdot x_{\hat{s}}) +$$

$$\sum_{x_s} \sum_{x_{\hat{i}}} max(0, \alpha - x_s \cdot x_i + x_s \cdot x_{\hat{i}})$$

to place correct samples closer while separating negative samples farther in the joint space. α is a hyperparameter and $x_{\hat{s}}$ and $x_{\hat{i}}$ are incorrect image and sentence pair obtained through negative sampling. We use MS COCO dataset (Lin et al., 2014) to train the network which contains 300k images and 5 captions per image. Final perception embeddings of dimension 1024 are sampled from the joint space regarding one word as a sentence.

Sentiment, either positive or negative, is determined for words that have sentiment orientations depending on their inherent meanings, usages, backgrounds etc. To capture the sentiment polarity of words (positive and negative), we use SentiWordNet3.0 (Baccianella et al., 2010), a lexical resource that automatically annotates the degree of positivity, negativity, and neutrality of English words. It is a one-dimensional value and if a word has multiple senses, we take the difference between the maximum positivity and the minimum negativity.

Emotion are considered by using NRC Emotion Lexicon (Mohammad and Turney, 2013) to reflect the emotional characteristics of words. It contains 15k words that are annotated with 10 emotion categories: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive. We built 10-dimensional one-hot emotion vectors based on this dataset.

Note that some embedding sets may not cover every word in our set of test vocabulary. In that case, out-of-vocabulary (OOV) words are initialized to zero for the missing modules. All embeddings are L2-normalized.

3.2 Ensemble methods

While the most rudimentary way for the amalgamation of several vectors is a concatenation with weights, other ensemble methods are expected to produce the vectors with improved quality (Henriksson, 2015). Faruqui and Dyer (2015) suggest that singular value decomposition (SVD) can be a promising way to merge the information by approximating the original matrix. Motivated by

their work, we examine two matrix factorization techniques, SVD and non-negative matrix factorization (NMF). In addition, we explore an unsupervised ensemble method via autoencoder (AE) networks. The details of these methods are illustrated below. Hyperparameters such as dimension d are selected to obtain the highest Spearmans correlation score in the RG-65 dataset (Rubenstein and Goodenough, 1965), which is used as a development set to minimize the interference on the test set. Note that before applying SVD, NMF, and AE, embeddings from different modules are concatenated with weights.

Concatenation (CONC) is used as the first step for ensembling multiple vectors of different dimensions. That is, let S be a set of n semantic spaces and s_i be a single vector space in S. $e_{id} \in s_i$ is a representation of word w_d in the semantic space $s_i \in S$. Then the resulting concatenated embedding e_d of word w_d is

$$e_d = \alpha_1 e_{d1} \oplus ... \oplus \alpha_i e_{di} \oplus ... \oplus \alpha_n e_{dn}$$

where $\oplus$ is the concatenation operator and $\sum_i \alpha_i = 1$. RG-65 is used as a development set to tune the weights α_i of particular embedding e_{di}.

Singular Value Decomposition (SVD) is a generalization of eigenvalue decomposition to any $m \times n$ matrix where it is reported to be effective in signal processing (Sahidullah and Kinnunen, 2016). Let V be the set of m words and k is the dimension of word embedding e_i for word $w_i \in V$. The dictionary matrix M is a $m \times k$ matrix where each row vector m_i of M is an embedding vector of e_i of word w_i. Then this matrix M is decomposed into $M = U\Sigma V^T$ where U and V are $m \times m$ and $n \times n$ real unitary matrices respectively, and Σ is a $m \times n$ non-negative real rectangular diagonal matrix. u_{id} is the first d dimension of i-th row vector u_i of U and we use it as a representation of word w_i. d is 230 for SVD. The size of vocabulary m is 20150.

Non-negative matrix factorization (NMF) has been reported to be effective method in various research areas including bioinformatics (Taslaman and Nilsson, 2012), signal denoising (Schmidt et al., 2007), and topic modeling (Arora et al., 2013). Two non-negative matrix W and H are optimized to approximate the dictionary matrix $M^T \approx WH$ by minimizing the frobenius norm $||M^T - WH||_F$ where $W, H \geq 0$. NMF has an inherent property of clustering the column vectors of the target matrix. To make M^T non-negative, we normalize the values of each embedding into the [0,1]. Let s_{id} be the first d dimension of i-th column vector s_i of W. Then we use s_{id} as a representation of word w_i. d is 200 for NMF.

Autoencoder (AE) is a neural network used for unsupervised learning of efficient coding for data compression or dimensionality reduction (Hinton and Sejnowski, 1986). Previous work suggests that an autoencoder may be able to learn relationships between the modules and result in higher-level embeddings (Silberer and Lapata, 2014). Our autoencoder consists of simple feedforward network. We trained two matrices W_{enc} of size $k \times d$ and W_{dec} of size $d \times k$ to learn efficient coding of word representation where k is the dimension of original word embedding and d is the dimension of compressed representation. Parameters are optimized to minimize cosine proximity loss:

$$L = \sum_{x \in T} 1 - \frac{\tilde{x} \cdot x}{||\tilde{x}|| \cdot ||x||}$$

where x is a k-dimensional word embedding, T is a training data set of size 20150 words, $\tilde{x} = f(W_{dec} f(W_{enc} x + b_{enc}) + b_{dec})$ and f is a ReLU non-linear activation function. We set $d = 900$.

4 Experiments

We introduce the experiments taken to examine how well the representations embed word meanings incorporating distinct properties. First, we apply our proposed embedding method to a word similarity measure task and a hypernym prediction task to measure its overall quality. Then we conducted a series of experiments for analyzing the characteristics of word meanings.

4.1 Word Similarity Measure and Hypernym Prediction

To assess the overall quality of proposed embedding method, we examined its performance via the word similarity task on SimLex-999 (Hill et al., 2016), WordSim-353 (Agirre et al., 2009), and MEN (Bruni et al., 2014) datasets. The similarity of each word pair is computed through cosine proximity, and we use Spearman's rank correlation as an evaluation metric. We also measure the performance of the different ensemble methods described in subsection 3.2. The result is compared with three baselines: Word2Vec, GloVe, and

Meta-embedding(1toN) (Yin and Schütze, 2015). The result is shown on table 1.

Baseline	SL	WS	MEN
Word2Vec	.442	.698	.782
GLoVe	.453	.754	.816
MetaEmb	.464	.745	.816
Polymodal (CONC)	.533	.622	.778
Polymodal (SVD)	**.580**	**.775**	**.838**
Polymodal (AE)	.507	.599	.751
Polymodal (NMF)	.414	.509	.589
Avg. Human	.780	.791	.840

Table 1: Spearman's correlation score on SimLex-999 (SL), WordSim-353 (WS), and MEN datasets. "Avg. Human" score is an inter-agreement between human annotators.

Our proposed method clearly outperforms the baselines in all the datasets, with near-human performance in WordSim-353 and MEN. Among the ensemble methods, SVD gave the best result showing its strong capability of combining information from different modules for this task.

We also conducted a hypernym prediction experiment using HyperLex dataset (Vulić et al., 2016) to analyze the quality of proposed embedding from a different perspective. Given a pair of two words, the task is to predict the degree of the first word being a type of the second word, for example "To what degree is *chemistry* a type of *science*?". We build a 2-layer feedforward network of dimensions 1000 and 500 respectively with a ReLU activation function to predict the hypernyms. Then the network is trained to predict the degree of hypernymity of the scale from 0.0 to 10.0 to minimize categorial cross-entropy loss using AdaGrad optimizer on the training set. The final evaluation metrics are obtained by calculating Spearman's correlation between the predicted degrees and the test set.

As in Table 2, the proposed method shows the highest correlation to the test set among all the cases including the baselines. Among the ensemble method, SVD again shows the highest performance. For the hypernym prediction, NMF gives a slightly better result than the simple weighted concatenation.

4.2 Property Norms Analysis

While the corpus-driven word representations such as Word2vec and GLoVe have been shown to embed some word-to-word relations such as *men* is to *women* as *king* is to *queen*, but it is still uncertain that they are also able to capture the properties like *has_ four_ legs* or *is_ delicious*. To see how well the models capture such properties of words, we perform the property norms analysis. We utilize the CSLB concept property norms dataset (Devereux et al., 2014) which annotates the normalized feature labels to the set of concepts. This dataset provides the normalized features of five categories: visual perceptual, other perceptual, taxonomic, encyclopedic, and functional. C is the set of all concepts and F is the set of all normalized features in CSLB dataset where $|C| = 638$ and $|F| = 5929$. For $f \in F$ and $c \in C$, $c \in C_f$ if and only if c has the feature f where $C_f \subset C$. The valid feature set F_v is a subset of F such that $f \in F_v$ only if there exist more than three concepts that have f, or equivalently, $|C_f| > 3$. Then the $|F_v| = 1053$.

	Test correlation (ρ)
Word2Vec	.319
GloVe	.391
MetaEmb	.400
Polymodal (CONC)	.445
Polymodal (SVD)	**.463**
Polymodal (NMF)	.454
Polymodal (AE)	.434

Table 2: Spearman's correlation score of HyperLex test dataset and predictions. The proposed method shows the highest correlation with the test dataset.

To examine how well each representation captures the normalized feature $f_i \in F_v$, we calculate the cosine similarity between $R(c)$ for $c \in C_{f_i}$ and $R(\overline{C_{f_i}})$ where $R(\cdot)$ is a mapping from concept to its distributed representation and $\overline{C_{f_i}}$ is a centroid of all concepts in C_{f_i} or

$$\overline{C_{f_i}} = \frac{1}{|C_{f_i}|} \sum_{c \in C_{f_i}} R(c)$$

In other words, $\overline{C_{f_i}}$ is a centroid of concepts that share the feature f_i. We define the *feature density* as the cosine similarity between the concept and the centroid. That is,

$$\textit{feature density}(c, f) = R(c) \cdot \overline{C_f}$$

We calculate the feature density of all target concept-feature pairs assuming vectors that share

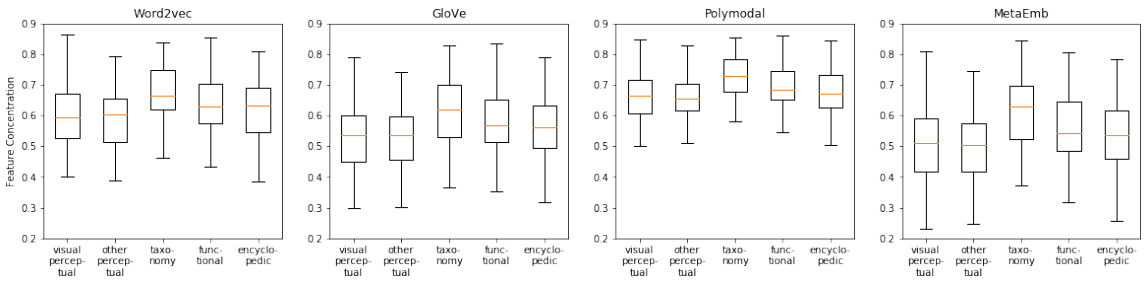

Figure 1: The feature density of different types of embedding on CSLB Concept Property Norms dataset.

	Word2Vec	GloVe	MetaEmb	Proposed
All features	.308	.262	.283	.343
Visual Perceptual	.204	.162	.188	.241
Other Perceptual	.122	.143	.148	.148
Taxonomic	.271	.236	.244	.263
Encyclopedic	.138	.129	.145	.136
Functional	.314	.288	.280	.310

Table 3: Spearman's correlation between the CSLB normalized feature representation and the target distributed representation.

the same features will also be distributionally similar (Erk, 2016).

In Figure 1 that summarizes the result, the proposed embedding method shows higher averages and lower deviations of feature densities across all the categories. It shows that our proposed embedding method is more capable of capturing normalized features than the baselines.

To further cement the observations, we calculate Spearman's correlation of word similarity measures between the normalized feature representation and the target distributed representation. The normalized feature representation of a concept is constructed as an one-hot vector which assigns 1 if the concept has the feature and 0 otherwise, and then L2-normalized to have length 1. Then we calculate the correlations of similarity measures by the feature categories. The results are shown in Table 3. While the proposed embedding method shows the highest correlation to the case of using all normalized features, it also shows a noticeable improvement in the visual perceptual category.

4.3 Positive vs Negative

One of the critical weakness of context-based word representation is that it cannot differentiate the sentiment polarity correctly. So we examine the ratio of neighbors that have same/opposite/neutral sentiment polarities with a

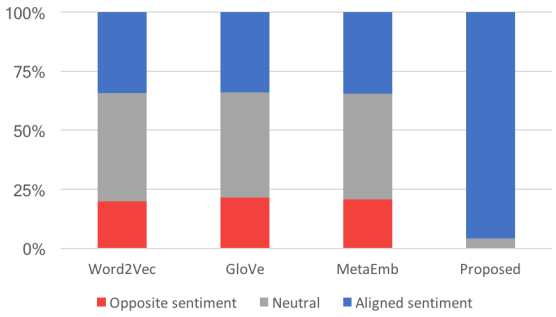

Figure 2: The ratio of 10 nearest neighbors that have same/opposite/neutral sentiment polarities of 15010 words.

target word among 15010 words and see how this problem can be mitigated. Figure 2 illustrates the result. The three context-based approaches show roughly 20% of incorrect sentiment differentiation. This can be benefited greatly from the sentiment module of the proposed approach as this issue is almost perfectly resolved by simply attaching sentiment values to the embedding. The result might be straightforward but this can improve the quality of embedding greatly.

4.4 Concrete vs Abstract

We hypothesize that the role of a certain module would be different depending on word characteristics such as the degree of concreteness. To validate this idea, we divided the Simlex-999 dataset into

two groups for different degrees of concept concreteness. This corresponds to 500 pairs of concrete words vs. 499 pairs of abstract words. Then we examine the relative importance of the different modules to each group via an ablation test. The result is reported in Table 4.

Modules	All	Concrete	Abstract
L (linear)	.442	.462	.449
T (syntactic)	.446	.439	.459
C (cognition)	.464	.451	.456
P (perception)	.157	.355	.010
S (sentiment)	.221	-.100	.293
E (emotion)	.376	.350	.385
All-but-L	.527▽	.464▽	.538▽
All-but-T	.524▽	.478▽	.501▽
All-but-C	.514▽	.466▽	.492▽
All-but-P	.531▽	.476▽	.570▲
All-but-S	.503▽	.491▲	.484▽
All-but-E	.526▽	.488▲	.540▽
All	.533	.483	.545

Table 4: Ablation tests for different word groups in Simlex-999. The metric is Spearman's correlation. Embeddings here are ensembled via weighted concatenation.

Interesting properties are revealed through the ablation test. By comparing the results between the different word groups, we can observe that the importance of a certain word aspect varies depending on the word characteristics. While concrete words profit from perception embeddings, the sentiment and emotion aspects are somewhat disturbing. We can observe an opposite result for abstract words. This result is quite intuitive since we can easily imagine the perceptual image from a concrete concept but not from an abstract one like *love*.

For a deeper analysis, we further investigate the role of each module in different word groups. For instance, since concrete concepts are perception-revealing, they would benefit from a strong emphasis on the perception embedding. On the other hand, emotion-revealing word groups such as abstract concepts would be opposite. Noting that the different types of words may have different sensitivity toward the modules, we adjusted the relative weights for a particular aspect of interest to be from 0.1 to 3.5 while maintaining others to 1.0. Then we observed the changes of the performance in word similarity task. The result is shown in Figure 3.

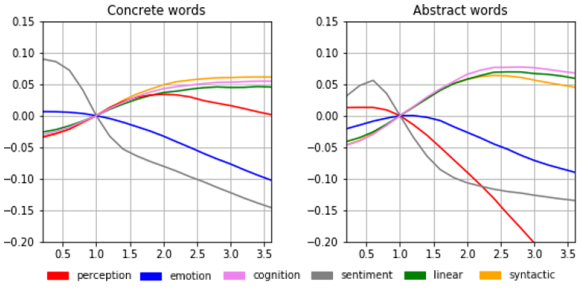

Figure 3: The result of sensitivity analysis. The weight of aspect-of-interest is adjusted while others are fixed to 1. These graphs reveal the distinct profiles of different word groups. Gradual patterns of emotion and perception are opposite for the concrete and abstract word groups.

The result of sensitivity analysis supports the idea that different word groups are influenced by each module with varying degrees. The x-axis refers to the relative weight of a particular aspect while setting the others to 1.0. The y-axis indicates the changes of Spearman's correlation score ρ on Simlex-999. The results in Figure 3 illustrate the different preferences among different word groups, which show the distinct nature between the two groups. In particular, the gradual patterns revealed by increasing relative weights of perception and emotion are contrary to concrete and abstract words. Increasing the weight of perception is beneficial for concrete word groups but detrimental to abstract word groups. However an exactly reverse pattern can be observed for the emotion. Increasing the weight of emotion is advantageous for abstract words but adverse for concrete words.

4.5 Similarity vs Relatedness

The "similarity" between two words is more strict term than the "relatedness". While the relatedness measures how much the two words are related to each other in some senses, the similarity measures how much the two words can be regarded as "similar" than just simply related. For example, consider the three word pairs: (bread, butter), (bread, toast), and (bread, stale). All of them can be regarded as "related" but only the (bread, toast) pairs can be regarded as "similar" because the other two words (butter and stale) are related but not similar to the "bread".

The two data sets SimLex-999 and WordSim-

353 capture this difference of similarity and relatedness. While the scores of WordSim-353 focus on the relatedness, those of the Simlex-999 deliberately try to distinguish between them. For example, a word pair (cloth, closet) is scored 8.00 in WordSim-353 dataset whereas 1.96 in the SimLex-999 dataset. To capture the difference between relatedness and similarity and see what modules contributes most to capture the similarity or the relatedness, we conduct a sensitivity analysis on WordSim-353 and SimLex-999 dataset.

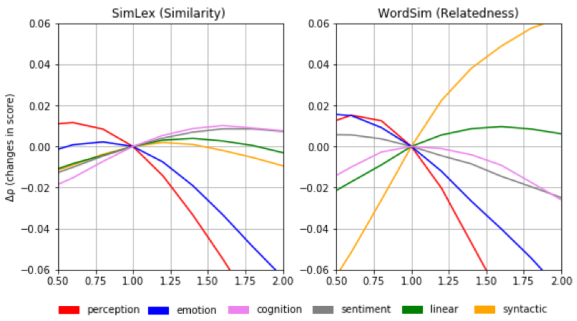

Figure 4: The result of sensitivity analysis on word similarity and word relatedness. While context information is important to the relatedness, sentiment polarity and lexical relations are important to the similarity.

Figure 4 shows the result of sensitivity analysis. In the SimLex-999 dataset which focuses on the word similarity, the cognition (lexical relation) and the sentiment modules turned out to be important. On the other hand, in the WordSim-353 dataset which focuses on the word relatedness, both linear context and syntactic context are turned out to be critical. This difference can be interpreted that the word properties extracted from the contexts are of the word relatedness, and in order to differentiate the similarity from the relatedness, additional properties such as lexical relations and sentiment polarities need to be introduced.

5 Conclusion

In this paper, we raise a question if the current distributed word representations sufficiently capture different aspects of word meanings. To address the question, we proposed a novel method for composing word embeddings, inspired by a human cognitive model. We compared our proposed embedding to the current state-of-the-art distributed word embedding methods such as Word2Vec, GloVe, and Meta-embedding from the perspective of capturing diverse aspects of word meanings.

Our proposed embedding performs better in the word similarity and hypernym prediction tasks than the baselines. We further conducted a series of experiments to study how well the word meanings are reflected by the representations and analyze the relationships between the modules and the word properties. From the property norms analysis, our findings show that the proposed method can capture the visual properties of words better than the baselines. Also, harnessing sentiment values helps the embedding greatly to resolve the sentiment polarity issue which is a limitation of current context-driven approaches. Based on the experimental results, we can conclude that some aspects of word meanings are not captured enough from the corpus and we can further improve the word embedding by referring to additional data related to a human mind model.

Finally, using our proposed method we show the different characteristics of concrete and abstract word groups and the difference between the concept relatedness and the concept similarity. We observe that emotional information is more important than the perceptual information for the abstract words whereas the opposite result is observed for the concrete words. Also, we see that the context-driven embeddings mostly capture the word relatedness and therefore lexical relation and sentiment polarities would be beneficial when considering the word similarity.

In conclusion, we concentrate on analyzing the relationships between the diverse aspects of word meanings and their distributed representations and propose a way to improve them by harnessing additional information based on the human cognitive model. Since our proposed method largely relies on the labeled extra data, this work has a limitation in terms of the scalability. For future research, we need to explore unsupervised ways of introducing perceptual properties and lexical relationships of words and annotating their sentiment and emotional properties. It will make our method more scalable.

Acknowledgements

This work was supported by Institute for Information communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2013-0-00179, Development of Core Technology for Context-aware Deep-

Symbolic Hybrid Learning and Construction of Language Resources)

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 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*, pages 19–27. Association for Computational Linguistics, 2009.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288, 2013.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

LW Barsalou. Perceptual symbol system. *Behavioral and Brain Science*, 22(4):577–609, 1999.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(1-47), 2014.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (CSLB) concept property norms. *Behavior research methods*, 46(4):1119, 2014.

Katrin Erk. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9:17–1, 2016.

Manaal Faruqui and Chris Dyer. Non-distributional word vector representations. *arXiv preprint arXiv:1506.05230*, 2015.

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

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Aron Henriksson. *Ensembles of semantic spaces: On combining models of distributional semantics with applications in healthcare*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University, 2015.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2016.

Geoffrey E Hinton and Terrence J Sejnowski. Learning and releaming in Boltzmann machines. *Parallel Distrilmted Processing*, 1, 1986.

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

Douwe Kiela. Deep embodiment: grounding semantics in perceptual modalities. Technical report, University of Cambridge, Computer Laboratory, 2017.

Douwe Kiela and Stephen Clark. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Douwe Kiela, Luana Bulat, and Stephen Clark. Grounding semantics in olfactory perception. In *ACL (2)*, pages 231–236, 2015.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *Icml*, volume 14, pages 595–603, 2014a.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014b.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL 2014*, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

Thang Luong, Richard Socher, and Christopher D Manning. Better word representations with recursive neural networks for morphology. In *CoNLL*, pages 104–113, 2013.

Mark E Maruish and James A Moses. *Clinical neuropsychology: Theoretical foundations for practitioners*. Psychology Press, 2013.

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

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

David C Plaut and James R Booth. Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological review*, 107(4):786, 2000.

Friedemann Pulvermüller. Words in the brain's language. *Behavioral and brain sciences*, 22(02):253–279, 1999.

Stephen Roller and Sabine Schulte Im Walde. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, 2013.

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

Md Sahidullah and Tomi Kinnunen. Local spectral variability features for speaker verification. *Digital Signal Processing*, 50:1–11, 2016.

Mikkel N Schmidt, Jan Larsen, and Fu-Tien Hsiao. Wind noise reduction using non-negative sparse coding. In *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, pages 431–436. IEEE, 2007.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pages 721–732, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Luc Steels. The symbol grounding problem has been solved. so whats next. *Symbols and embodiment: Debates on meaning and cognition*, pages 223–244, 2008.

Leo Taslaman and Björn Nilsson. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PloS one*, 7(11):e46331, 2012.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. Hyperlex: A large-scale evaluation of graded lexical entailment. *arXiv preprint arXiv:1608.02117*, 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.

Wenpeng Yin and Hinrich Schütze. Learning meta-embeddings by using ensembles of embedding sets. *arXiv preprint arXiv:1508.04257*, 2015.