Inside Out: Two Jointly Predictive Models for Word Representations and Phrase Representations

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and **Xueqi Cheng**
CAS Key Lab of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100190, China

Abstract

Distributional hypothesis lies in the root of most existing word representation models by inferring word meaning from its external contexts. However, distributional models cannot handle rare and morphologically complex words very well and fail to identify some fine-grained linguistic regularity as they are ignoring the word forms. On the contrary, morphology points out that words are built from some basic units, *i.e.*, morphemes. Therefore, the meaning and function of such rare words can be inferred from the words sharing the same morphemes, and many syntactic relations can be directly identified based on the word forms. However, the limitation of morphology is that it cannot infer the relationship between two words that do not share any morphemes. Considering the advantages and limitations of both approaches, we propose two novel models to build better word representations by modeling both external contexts and internal morphemes in a jointly predictive way, called BEING and SEING. These two models can also be extended to learn phrase representations according to the distributed morphology theory. We evaluate the proposed models on similarity tasks and analogy tasks. The results demonstrate that the proposed models can outperform state-of-the-art models significantly on both word and phrase representation learning.

Introduction

Representing words as dense, real-valued vectors in a relatively low-dimensional space, called distributed word representations, has attracted a huge spike of interest in recent years. These vectors have been widely used in various natural language processing tasks, *e.g.*, named entity recognition (Collobert et al. 2011), question answering (Zhou et al. 2015), and parsing (Socher et al. 2011). Building such representations follows the well-known linguistic principle—*Distributional Hypothesis* (Harris 1954; Firth 1957), which states that words occurring in similar contexts tend to have similar meanings.

However, reliable distributional vectors can be learned only for words that occur in plenty external contexts in the corpus. Therefore, the rare words, *e.g.*, morphologically complex words or new words, are often poorly represented. Even for the frequent words, distributional hypothesis may meet its own difficulty in identifying fine-grained linguistic regularity. For example, there are some words that are very difficult to differentiate from external contexts, like "*buy*" and "*sell*". This may further make distributional semantic models difficult to answer the similarity question, such as which word is more close to "*buy*", "*buys*" or "*sells*"?

In fact, this question is not difficult to answer if we take the internal structures of words into account. This involves another important research field in linguistics, *Morphology*. It studies how words are built from morphemes, the smallest grammatical (or meaningful) unit in a language, such as root, prefix, and suffix. In morphology, there also exists an underlying distributional principle of semantics, which states that words contain the same morphemes may convey similar meaning or function (Williams 1981; Bybee 1985). Therefore, one can infer the meaning of word "*breakable*" by its root "*break*" and suffix "*able*", and accomplish the analogy task "*breakable* to *break* as *doable* to *do*" simply based on their word forms. However, the limitation of morphology is that it cannot infer the relationship between two words that do not share any morphemes, like "*dog*" and "*husky*", even though they might be related.

In short, distributional hypothesis infers the word meaning from its external context, while morphology infers the word meaning from its internal forms. Both have their own advantages and limitations. It is natural to seek a way to integrate these two sources to obtain better word representations.

In this work, we propose two simple and general models to integrate both external contexts and internal morphemes to learn better word representations, called BEING and SEING. The proposed models are built on the basis of Continuous Bag-of-Words (CBOW) model and Skip-Gram (SG) model (Mikolov et al. 2013a) due to the efficiency concern. In a nutshell, we view the two sources, *i.e.*, internal morphemes and external contexts, equivalently in inferring word representations, and model them in a general predictive way. Comparing with the word representation models which rely on context information alone, *e.g.*, `word2vec` and GloVe, our models can capture the relationships among morphologically related words, rather than treating each word as an independent entity. As a result,

the proposed models can alleviate the sparsity problem and learn the rare words much better. Comparing with those compositional models which also leverage morphological and context information (Luong, Socher, and Manning 2013; Botha and Blunsom 2014), our simple predictive models can avoid the errors accumulated in the sophisticated usage of the morphological information.

Moreover, according to the theory of distributed morphology (Halle and Marantz 1993), there is no divide between the construction of complex words and complex phrases. Based on this idea, our models can be easily extended to the learning of phrase representations, by viewing constituting words in a phrase as its *morphemes*.

We evaluate our models on both word representations and phrase representations. For word representation learning, we evaluate the learned representations on two tasks, word similarity and word analogy. The results show that the proposed models outperform not only the widely-used state-of-the-art methods, but also other models which used morphological information. For phrase representation learning, we evaluated our models on the phrase analogy task introduced in (Mikolov et al. 2013b). The results show that our models significantly outperform all the baselines.

Related Work

Distributional Hypothesis

Representing words as continuous vectors in a low-dimensional space can go back decades ago (Hinton, McClelland, and Rumelhart 1986). Building such representations follows the distributional hypothesis, stating that words in similar contexts have similar meanings. Based on this hypothesis, various methods have been developed in the NLP community, including clustering (Brown et al. 1992), matrix factorization (Deerwester et al. 1990; Pennington, Socher, and Manning 2014), probabilistic models (Blei, Ng, and Jordan 2003), and neural networks (Bengio et al. 2003; Collobert and Weston 2008). Inspired by the success in neural network language modeling, there has been a flurry of subsequent work, which explored various neural network structures and optimization methods to learn word representations, including (Collobert and Weston 2008; Mikolov et al. 2013a; Mnih and Kavukcuoglu 2013; Mikolov et al. 2013b). Among these methods, the state-of-the-art methods are continuous bag-of-words model (CBOW) and Skip-Gram (SG) model introduced by Mikolov et al. (2013a), because of their simplicity, efficiency, and scalability.

Morphology

There is another line of research attempts to leverage subword units information for better word representations.

Alexandrescu and Kirchhoff (2006) proposed a factored neural language model. In that model, each word is viewed as a vector of fixed number of features like stems, morphological tags, and capitalization. Collobert et al. (2011) tried to enhance their word vectors using extra character-level features such as capitalization and part-of-speech (POS).

Some work tries to uncover morphological compositionality. Lazaridou et al. (2013) explored compositional Distributional Semantic Models (cDSMs) with different compositional methods to derive the representations of morphologically complex words. However, their models can only combine a stem with an affix. Luong, Socher, and Manning (2013) proposed a context-sensitive morphological Recursive Neural Network (csmRNN) to model morphological structure of words in the neural language model training approach proposed by (Collobert et al. 2011). Botha and Blunsom (2014) integrated compositional morphological representations into a log-bilinear language model (CLBL++). Generally, these works make use of the morphological information in a sophisticated way in order to build a neural language model. However, several recent works have shown that simple and straightforward models can acquire better word representations (Mikolov et al. 2013a; Pennington, Socher, and Manning 2014).

To the best of our knowledge, the most closest work to ours are (Qiu et al. 2014; Chen et al. 2015). Qiu et al. (2014) enhanced CBOW with the contexts' morphemes, named MorphemeCBOW. Chen et al. (2015) adopted similar models to learn Chinese character and word representations. However, they did not capture the interaction between the words and their morphemes. On the contrary, our proposed models directly capture such interaction in a predictive way.

Our Models

In this section, we will introduce two simple and general model integrating both external and internal information. Without loss of generality, we will take word representation learning for an example to introduce our models. Besides, we also show how these models can be employed to learn phrase representations.

Notation

We first list the notations used in this paper. Let $\mathcal{C}=\{w_1,\ldots,w_N\}$ denote a corpus of N word sequence over the word vocabulary W. The external contexts for word $w_i \in W$ (*i.e.*, i-th word in corpus) are the words surrounding it in an l-sized window $(c_{i-l},\ldots,c_{i-1},c_{i+1},\ldots,c_{i+l})$, where $c_j \in C$, $j \in [i-l,\ i+l]$. The internal morphemes for word w_i are denoted by $(m_i^{(1)},\ldots,m_i^{(s(w_i))})$, where $s(w_i)$ is the number of morphemes for w_i. Each word $w \in W$, context $c \in C$, and morpheme $m \in M$ are associated with vectors $\vec{w} \in \mathbb{R}^d$, $\vec{c} \in \mathbb{R}^d$, and $\vec{m} \in \mathbb{R}^d$ respectively, where d is the representation dimensionality. In this paper, $\vec{x}$ denotes the vector of the variable x unless otherwise specified. The entries in the vectors are parameters to be learned.

Continuous Bag of External and Internal Gram Model

The architecture of the first proposed model is shown in Figure 1. In this model, a target word is predicted by its surrounding context, as well as the morphemes it contains. To illustrate it, we take a word sequence "*... glass is breakable, take care ...*" for an example. For the target word "*breakable*", the external context words ("*glass*", "*is*", "*take*", and "*care*") are used to predict it. Such prediction task captures

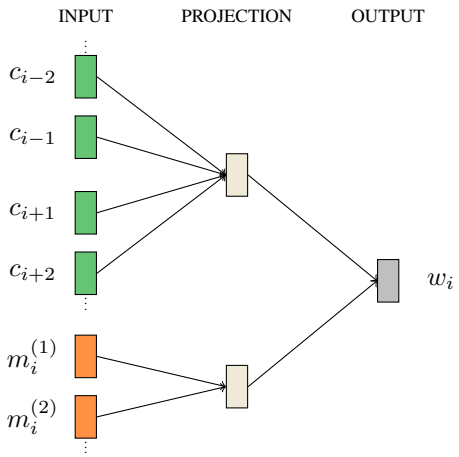

Figure 1: The framework for BEING model.

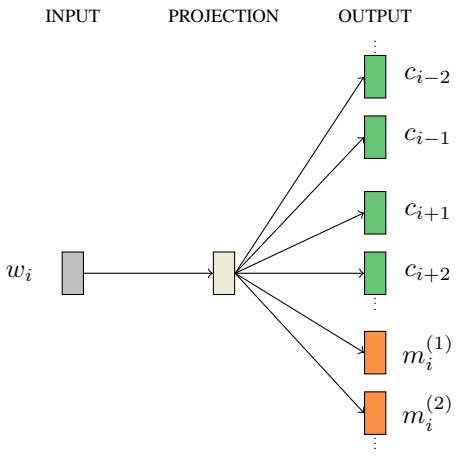

Figure 2: The framework for SEING model.

the distributional hypothesis, since words with similar context tend to have similar vectors. Besides, the morphemes for word "*breakable*" (*i.e.*, "*break*" and "*able*") are also used to predict it. This prediction task captures the morphological relation, since words sharing the same morphemes tend to have similar vectors. We call this model the Continuous **B**ag of **E**xternal and **IN**ternal **G**ram (BEING) model.

Formally, given a corpus $\mathcal{C}$, the goal of BEING model is to maximize the following objective function:

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\log p(w_i|\mathcal{P}_i^c) + \log p(w_i|\mathcal{P}_i^m) \Big)$$

where $\mathcal{P}_i^c$ and $\mathcal{P}_i^m$ denote the projection of w_i's external contexts and internal morphemes, respectively.

We use softmax function to define the probabilities $p(w_i|\mathcal{P}_i^c)$ and $p(w_i|\mathcal{P}_i^m)$ as follows:

$$p(w_i|\mathcal{P}_i^c) = \frac{\exp(\overrightarrow{w_i} \cdot \overrightarrow{\mathcal{P}_i^c})}{\sum_{w \in W} \exp(\overrightarrow{w} \cdot \overrightarrow{\mathcal{P}_i^c})}$$

$$p(w_i|\mathcal{P}_i^m) = \frac{\exp(\overrightarrow{w_i} \cdot \overrightarrow{\mathcal{P}_i^m})}{\sum_{w \in W} \exp(\overrightarrow{w} \cdot \overrightarrow{\mathcal{P}_i^m})}$$

where $\overrightarrow{\mathcal{P}_i^c}$ and $\overrightarrow{\mathcal{P}_i^m}$ denote the projected vectors of w_i's external contexts and internal morphemes, respectively. They are defined as:

$$\overrightarrow{\mathcal{P}_i^c} = h_c(\overrightarrow{c_{i-l}}, \ldots, \overrightarrow{c_{i-1}}, \overrightarrow{c_{i+1}}, \ldots, \overrightarrow{c_{i+l}})$$

$$\overrightarrow{\mathcal{P}_i^m} = h_m\Big(\overrightarrow{m_i^{(1)}}, \ldots, \overrightarrow{m_i^{(s(w_i))}}\Big)$$

where $h_c(\cdot)$ and $h_m(\cdot)$ can be sum, average, concatenate or max pooling of context vectors. In this paper, we use average for both of them, as that in `word2vec` tool.

We adopt the negative sampling technique (Mikolov et al. 2013b) to learn the model, due to the high computational complexity of the original objective function. The negative sampling actually defines an alternate training objective function as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\log \sigma(\overrightarrow{w_i} \cdot \overrightarrow{\mathcal{P}_i^c}) + k \cdot \mathbf{E}_{\tilde{w} \sim P_{\tilde{W}}} \log \sigma(-\overrightarrow{\tilde{w}} \cdot \overrightarrow{\mathcal{P}_i^c}) + \log \sigma(\overrightarrow{w_i} \cdot \overrightarrow{\mathcal{P}_i^m}) + k \cdot \mathbf{E}_{\tilde{w} \sim P_{\tilde{W}}} \log \sigma(-\overrightarrow{\tilde{w}} \cdot \overrightarrow{\mathcal{P}_i^m}) \Big)$$

where $\sigma(x) = 1/(1+\exp(-x))$, k is the number of *negative* samples, $\tilde{w}$ denotes the sampled negative word (*i.e.*, random sampled word which is not relevant with current contexts), and $P_{\tilde{W}}$ denotes the distribution of negative word samples. The noise distribution is set the same as that of (Mikolov et al. 2013a), $p_{\tilde{W}}(w) \propto \#(w)^{0.75}$, where $\#(w)$ means the number of word w appearing in corpus $\mathcal{C}$.

Continuous Skip External and Internal Gram Model

We also extend the Skip-Gram model to integrate both external contexts and internal morphemes, as shown in Figure 2. We name it the Continuous **S**kip **E**xternal and **IN**ternal **G**ram (SEING) model. In this model, the target word is used to predict its surrounding external context words, as well as the morphemes it contains. For the word sequence "*... glass is breakable, take care ...*", the center word "*breakable*" needs to predict not only external context words ("*glass*", "*is*", "*take*", and "*care*"), but also its morphemes (*i.e.*, "*break*" and "*able*").

Formally, the goal of SEING model is to maximize the following objective function:

$$\mathcal{L} = \sum_{i=1}^{N} \Bigg(\sum_{\substack{j=i-l \\ j \neq i}}^{i+l} \log p(c_j|w_i) + \sum_{z=1}^{s(w_i)} \log p(m_i^{(z)}|w_i) \Bigg)$$

where $p(c_j|w_i)$ and $p(m_i^{(z)}|w_i)$ are defined by softmax function as follows:

$$p(c_j|w_i) = \frac{\exp(\overrightarrow{c_j} \cdot \overrightarrow{w_i})}{\sum_{c \in C} \exp(\overrightarrow{c} \cdot \overrightarrow{w_i})}$$

$$p(m_i^{(z)}|w_i) = \frac{\exp(\overrightarrow{m_i^{(z)}} \cdot \overrightarrow{w_i})}{\sum_{m \in M} \exp(\overrightarrow{m} \cdot \overrightarrow{w_i})}$$

We also adopt negative sampling to learn this model. As a result, it defines an alternate objective function as follows:

$$\mathcal{L}=\sum_{i=1}^{N}\Bigg(\sum_{\substack{j=i-l\\ j\neq i}}^{i+l}\Big(\log\sigma(\vec{c_j}\cdot\vec{w_i})+k\cdot\mathbf{E}_{\tilde{c}\sim P_{\tilde{C}}}\log\sigma(-\vec{\tilde{c}}\cdot\vec{w_i})\Big)+\sum_{z=1}^{s(w_i)}\Big(\log\sigma(\overrightarrow{m_i^{(z)}}\cdot\vec{w_i})+k\cdot\mathbf{E}_{\tilde{m}\sim P_{\tilde{M}}}\log\sigma(-\vec{\tilde{m}}\cdot\vec{w_i})\Big)\Bigg)$$

where $\tilde{c}$ and $\tilde{m}$ denote the sampled negative context and morpheme, respectively; $P_{\tilde{C}}$ and $P_{\tilde{M}}$ denote the noise distribution of contexts and morphemes, respectively.

Optimization

Following the optimization scheme used in (Mikolov et al. 2013b), we use stochastic gradient descent (SGD) for optimization, and adopt the same linear learning rate schedule described in (Mikolov et al. 2013a). Gradients are calculated via back-propagation algorithm. Both word and morpheme vectors are initialized randomly using the same scheme as in `Word2Vec` and GloVe.

Phrase Representation

The basic principle of distributed morphology is that both phrases and words are assembled by the single generative engine (the syntax). This means that there is no difference between the construction of words and phrases. Therefore, we can easily apply our models to phrase representations through viewing constituting words in a phrase as its *morphemes*.

More specifically, in phrase representation learning, the objective function of BEING is defined as follows:

$$\mathcal{L}=\sum_{i=1}^{N}\Big(\log\sigma(\vec{g_i}\cdot\overrightarrow{\mathcal{P}_i^{\tilde{c}}})+k\cdot\mathbf{E}_{\tilde{g}\sim P_{\tilde{G}}}\log\sigma(-\vec{\tilde{g}}\cdot\overrightarrow{\mathcal{P}_i^{\tilde{c}}})+\log\sigma(\vec{g_i}\cdot\overrightarrow{\mathcal{P}_i^{\tilde{w}}})+k\cdot\mathbf{E}_{\tilde{g}\sim P_{\tilde{G}}}\log\sigma(-\vec{\tilde{g}}\cdot\overrightarrow{\mathcal{P}_i^{\tilde{w}}})\Big)$$

where $\vec{g_i}$ denotes the vector of target phrase g_i, $\tilde{g}$ denotes the sampled negative phrase, $\overrightarrow{\mathcal{P}_i^{\tilde{c}}}$ and $\overrightarrow{\mathcal{P}_i^{\tilde{w}}}$ denote the projection vectors of g_i's external contexts and internal words respectively, and $P_{\tilde{G}}$ denotes the distribution of negative phrase samples.

The objective function of SEING in phrase representation learning is defined as follows:

$$\mathcal{L}=\sum_{i=1}^{N}\Bigg(\sum_{\substack{j=i-l\\ j\neq i}}^{i+l}\Big(\log\sigma(\vec{c_j}\cdot\vec{g_i})+k\cdot\mathbf{E}_{\tilde{c}\sim P_{\tilde{C}}}\log\sigma(-\vec{\tilde{c}}\cdot\vec{g_i})\Big)+\sum_{z=1}^{s(g_i)}\Big(\log\sigma(\overrightarrow{w_i^{(z)}}\cdot\vec{g_i})+k\cdot\mathbf{E}_{\tilde{w}\sim P_{\tilde{W}}}\log\sigma(-\vec{\tilde{w}}\cdot\vec{g_i})\Big)\Bigg)$$

where c_j and $w_i^{(z)}$ denote external context and internal word of phrase g_i, $s(g_i)$ is the number of words that g_i contains, $\tilde{c}$ and $\tilde{w}$ denote the sampled negative external context and internal word respectively, $P_{\tilde{C}}$ and $P_{\tilde{W}}$ denote the noise distribution of external contexts and internal words, respectively.

Experiments

Experimental Settings

We choose the Wikipedia April 2010 dump[1] (Shaoul and Westbury 2010), which has been widely used by (Huang et al. 2012; Luong, Socher, and Manning 2013; Neelakantan et al. 2014), as the corpus to train all the models[2]. The corpus contains 3,035,070 articles and about 1 billion tokens. In preprocessing, we lowercase the corpus, remove pure digit words and non-English characters. During training, the words occurring less than 20 times are ignored, resulting in a vocabulary of 388,723 words. We obtain morphemes for words in the vocabulary by an unsupervised morphological segmentation toolkit, named `Morfessor` (Creutz and Lagus 2007), which was also used in (Luong, Socher, and Manning 2013; Botha and Blunsom 2014; Qiu et al. 2014). Following the practice in (Mikolov et al. 2013b; Pennington, Socher, and Manning 2014), we set context window size as 10 and use 10 negative samples.

We compare our models[3] with two classes of baselines:

- Models also using morphological information including csmRNN (Luong, Socher, and Manning 2013), MorphemeCBOW (Qiu et al. 2014), and CLBL++ (Botha and Blunsom 2014).
- State-of-the-art word representation models including CBOW, SG (Mikolov et al. 2013a), and GloVe (Pennington, Socher, and Manning 2014).

For csmRNN, we use the word vectors provided by the authors, including HSMN+csmRNN and C&W+csmRNN[4] that use the HSMN and C&W vectors as the initialization. For MorphemeCBOW and CLBL++, we just take the results reported in (Botha and Blunsom 2014; Qiu et al. 2014) since they do not release the code or word vectors. For CBOW, SG[5], and GloVe[6], we use the tools released by the authors. They are trained on the same corpus with the same setting as our models for fair comparison.

Word Similarity

To see whether integrating external contexts and internal morphemes improves the quality of rare word vectors, we first evaluate the proposed models on the English rare word (RW) testset (Luong, Socher, and Manning 2013). It consists of 2034 word pairs together with human assigned similarity scores, and contains more morphological complex words than other word similarity testsets. We also evaluate our models on a variety of well-known standard testsets including WordSim-353 (WS-353) (Finkelstein et al. 2002) and SimLex-999 (SL-999) (Hill, Reichart, and Korhonen 2014).

[1] `http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html`

[2] `Morfessor` is also trained on this corpus.

[3] The source code and word vectors can be downloaded at `http://ofey.me/projects/InsideOut/`.

[4] `http://stanford.edu/~lmthang/morphoNLM/`

[5] `https://code.google.com/p/word2vec/`

[6] `http://nlp.stanford.edu/projects/glove/`

Table 1: Spearman rank correlation $\rho \times 100$ on word similarity tasks. Bold scores are the best within groups of the same dimensionality.

Model	Dim	RW	WS-353	SL-999
HSMN+csmRNN	50	22.31	64.48	17.29
C&W+csmRNN	50	34.36	58.81	24.0
CLBL++	50	30	39	——
MorphemeCBOW	50	32.88	65.19	——
GloVe	50	30.57	55.48	25.28
CBOW	50	40.65	64.47	30.87
SG	50	39.57	65.30	27.66
BEING	50	**45.92**	67.57	**32.14**
SEING	50	42.08	**67.85**	29.36
GloVe	300	34.13	59.18	32.35
CBOW	300	45.19	67.21	38.82
SG	300	45.55	70.74	36.07
BEING	300	**50.41**	**73.36**	**44.91**
SEING	300	46.86	72.09	38.93

We employ the spearman rank correlation between the human judgements and the similarity scores based on learned word vectors as evaluation metric.

Result Table 1 shows the results on three different word similarity testsets. As we can see that CBOW and SG are much stronger baselines, comparing with compositional language models using morphological information like csmRNN and CLBL++. This once again confirms that simple model directly learning word representations can derive better word representations.

Moreover, our models, especially BEING, outperform these state-of-the-art methods on all three testsets. On WS-353 and SL-999, the proposed models perform consistently better than other baselines, showing that they can also learn better representations for common words. All these results suggest that modeling both external contexts and internal morphemes in a jointly predictive way can derive better word representations.

Word Analogy

Besides the word similarity task, we also evaluate our models on word analogy task. This task, introduced by (Mikolov et al. 2013a), is to evaluate the linguistic regularities between pairs of word vectors. The task consists of questions like "*a* is to *b* as *c* is to __", where __ is missing and must be inferred from the entire vocabulary. The testset contains 5 types of semantic analogies and 9 types of syntactic analogies[7]. The semantic analogy contains 8869 questions, typically about people and place like "*Athens* is to *Greece* as *Paris* is to *France*", while the syntactic analogy contains 10,675 questions, mostly on forms of adjectives or verb tense, such as "*calm* is to *calmly* as *quiet* to *quietly*".

To answer such questions, one needs to find a word vector $\vec{x}$, which is the closest to $\vec{b} - \vec{a} + \vec{c}$ according to the cosine similarity:

$$\arg \max_{\substack{x \in W, x \neq a \\ x \neq b,\ x \neq c}} \cos(\vec{b} - \vec{a} + \vec{c},\ \vec{x})$$

The question is regarded as answered correctly only if x is exactly the answer word in the evaluation set. We use the percentage of questions answered correctly as the evaluation metric for this task.

Table 2: Results on the word analogy task. Bold scores are the best within groups of the same dimension.

Model	Dim	Semantic	Syntactic	Total
HSMN+csmRNN	50	5.06	9.36	7.41
C&W+csmRNN	50	9.21	12.34	10.93
GloVe	50	56.6	43.53	49.46
CBOW	50	60.86	50.55	55.23
SG	50	50.27	43.93	46.81
BEING	50	**63.67**	**56.76**	**59.90**
SEING	50	50.92	49.06	49.90
GloVe	300	79.85	61.15	69.64
CBOW	300	79.65	68.54	73.58
SG	300	77.16	65.31	70.69
BEING	300	**81.95**	**73.17**	**77.15**
SEING	300	79.07	70.26	74.26

Result Table 2 shows the results on word analogy task including semantic, syntactic, and total precision. Firstly, we can observe that csmRNN performs much poorer than other models, while CBOW, SG, and GloVe solve the analogy task pretty well. This is consistent with the explanation of Arora et al. (2015), that simple loglinear models like Skip Gram or GloVe can capture linear linguistic regularities. Moreover, our BEING model performs significantly better than these state-of-the-art methods.

The results of MorphemeCBOW are absent since the authors reported the word analogy results on enwiki9 corpus[8] instead of entire Wikipedia corpus. Therefore, we test BEING on enwiki9 corpus using the same setting as MorphemeCBOW, and get a total precision of 60.23% on 300-dimensional representations, while the best precision of MorphemeCBOW is only 41.96%.

Besides, the results show our models can gain more improvement on syntactic subtask than semantic subtask. This is because morphemes can strengthen the inference on syntactic analogy task, such as "*great* to *greatest* as *cool* to *coolest*".

We also present the precision of syntactic analogies discovered in 300-dimensional vectors of each model in Table 3, which is broken down by relation type. Clearly, BEING and SEING perform significantly better than CBOW and SG on almost all subtasks (except adjective-to-adverb). For *adjectives-to-adverbs* relation, it contains lots of words wrongly segmented by `Morfessor`. For example, "*luckily*" is segmented to "*lucki*" + "*ly*", however its root should be

[7] `http://code.google.com/p/word2vec/source/browse/trunk/questions-words.txt`

[8] `http://mattmahoney.net/dc/textdata.html`

Table 3: Breakdown of syntactic analogy in each representation by relation type.

Syntactic Subtask	CBOW	BEING	SG	SEING
adjective-to-adverb	**31.85**	26.51	**38.10**	37.20
opposite	34.73	**45.07**	30.79	**39.16**
comparative	88.14	**91.82**	79.58	**83.93**
superlative	61.14	**71.30**	48.31	**61.94**
present-participle	67.23	**67.42**	62.59	**66.67**
nationality-adjective	90.18	**91.56**	90.24	**90.68**
past-tense	66.86	**69.17**	61.28	**64.94**
plural	81.91	**86.86**	82.21	**84.53**
plural-verbs	65.86	**85.86**	67.47	**81.26**

"*lucky*". Our models failed in this subtask, since they cannot correctly connect the word "*luckily*" and "*lucky*" with such wrong morphemes. In the future, we will tackle this problem by segmenting words into morphemes via a dictionary, like *root* method in MorphemeCBOW.

Phrase Analogy

We also conduct experiments to test our models' ability on learning phrase representations. To learn vector representation for phrases, we first identify phrases (1–4 grams) in the Wikipedia April 2010 corpus following the idea introduced by (Mikolov et al. 2013b), and then train our models and baselines on this new corpus with the same setting as in word tasks. As a result, we obtain the representations of 799,805 phrases for each model.

We evaluate the quality of the phrase representations using the phrase analogy task introduced in (Mikolov et al. 2013b). The testset contains 3218 questions like "*boston* is to *boston bruins* as *los angeles* is to *los angeles kings*".

Result Figure 3 shows the results on phrase analogy task with different dimensions from 50 to 400. The figure shows that our models, especially SEING model, perform significantly better than all the other baselines on all the dimensions. The best result reported earlier was 72% achieved by 1000-dimensional vectors of skip-gram model trained on a dataset with about 33 billion words using the hierarchical softmax (Mikolov et al. 2013b), while our SEING model outperforms it using 300-dimensional representations trained on 1 billion words. This is a promising result, indicating that those models based solely on external contexts can benefit a lot from internal information.

Discussion

Comprehensively observing the results of word and phrase tasks, some trends emerge: 1) BEING and CBOW are superior in the word tasks, while SEING and SG are superior in the phrase tasks; 2) Compared with their sub-models, BEING gains more in word tasks, while SEING has a better margin on phrase task. Next, we try to explain these phenomena from the perspective of matrix factorization.

Following the idea introduced in (Levy and Goldberg 2014), CBOW and SG using negative sampling can be seen as factorizing the terms-by-n-terms and terms-by-terms co-occurrence matrix respectively, where term can be word or

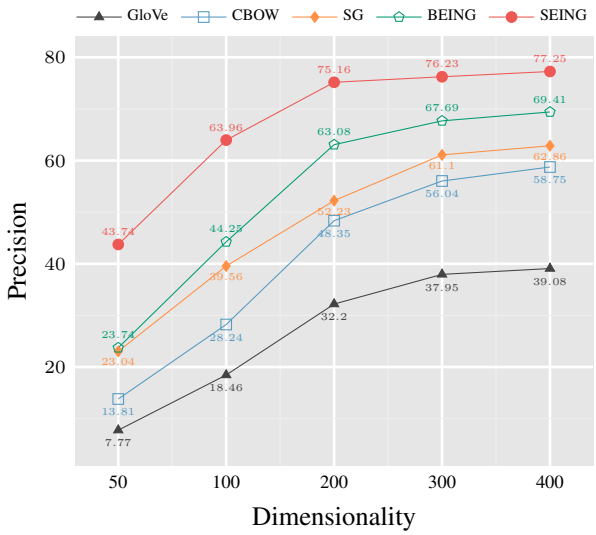

Figure 3: Results on the phrase analogy task.

phrase. Meanwhile, SEING (or BEING) with negative sampling can be seen as co-factorizing both a terms-by-terms (or terms-by-n-terms) co-occurrence matrix and a terms-by-morphemes (or terms-by-n-morphemes) co-occurrence matrix simultaneously.

From the perspective of matrix factorization, given the vocabulary, CBOW will outperform SG when the corpus is sufficient, since n-terms can provide more expressive power. And yet, CBOW will also confront with more serious sparsity problem than SG when the corpus is not big enough for the vocabulary. This is the reason why CBOW can outperform SG on word tasks but fail on phrase task, considering we use the same corpus but the vocabulary with the twice size on phrase task. It also explains why SEING improves SG with a larger margin on phrase task but improves not much on word tasks comparing with BEING.

Conclusion and Future Work

In this paper, we propose two novel models to build better word representations by modeling both external contexts and internal morphemes in a jointly predictive way. The experimental results on both word similarity tasks and word analogy tasks show that our models perform not only significantly better than state-of-the-art models that do not integrate morphological information, but also much better than other models also using morphological information.

Several directions remain to be explored. Although this paper focuses on English, our models deserve to be applied to other morphologically rich languages such as French and Turkish. Considering morphemes (words) are used as sub-unit for words (phrases) in this work, other more general sub-unit will be an interesting way to explore, like letter n-gram (Huang et al. 2013).

Acknowledgments

This work was funded by 973 Program of China under Grants No. 2014CB340401 and 2012CB316303, 863 Program of China under Grants No. 2014AA015204, the National Natural Science Foundation of China (NSFC) under Grants No. 61472401, 61433014, 61425016, and 61203298, Key Research Program of the Chinese Academy of Sciences under Grant No. KGZD-EW-T03-2, and Youth Innovation Promotion Association CAS under Grants No. 20144310. We thank Minh-Thang Luong, Jeffrey Pennington and Tomas Mikolov for their kindness in sharing codes and word vectors. We also thank Yanran Li for valuable comments and discussion.

References

Alexandrescu, A., and Kirchhoff, K. 2006. Factored neural language models. In *Proceedings of NAACL*, 1–4.

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR* abs/1502.03520.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Botha, J. A., and Blunsom, P. 2014. Compositional morphology for word representations and language modelling. In *Proceedings of ICML*, 1899–1907.

Brown, P. F.; deSouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based n-gram models of natural language. *Comput. Linguist.* 18(4):467–479.

Bybee, J. L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Chen, X.; Xu, L.; Liu, Z.; Sun, M.; and Luan, H. 2015. Joint learning of character and word embeddings. In *Proceedings of IJCAI*, 1236–1242.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the ICML*, 160–167.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12:2493–2537.

Creutz, M., and Lagus, K. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.* 4(1):3:1–3:34.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci. Technol.* 41(6):391–407.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; andGadi Wolfman, Z. S.; and Ruppin, E. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.* 20(1):116–131.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)* 1952-59:1–32.

Halle, M., and Marantz, A. 1993. Distributed morphology and the pieces of inflection. *The View from Building 20, ed.* 111–176.

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

Hill, F.; Reichart, R.; and Korhonen, A. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR* abs/1408.3456.

Hinton, G. E.; McClelland, J. L.; and Rumelhart, D. E. 1986. Distributed representations. In Rumelhart, D. E.; McClelland, J. L.; and PDP Research Group, C., eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*. Cambridge, MA, USA: MIT Press. 77–109.

Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, 873–882.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of CIKM*, 2333–2338.

Lazaridou, A.; Marelli, M.; Zamparelli, R.; and Baroni, M. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*, 1517–1526.

Levy, O., and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*. 2177–2185.

Luong, M.-T.; Socher, R.; and Manning, C. D. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, 104–113.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop of ICLR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. Curran Associates, Inc. 3111–3119.

Mnih, A., and Kavukcuoglu, K. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of NIPS*. 2265–2273.

Neelakantan, A.; Shankar, J.; Passos, A.; and McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, 1059–1069.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 1532–1543.

Qiu, S.; Cui, Q.; Bian, J.; Gao, B.; and Liu, T. 2014. Co-learning of word representations and morpheme representations. In *Proceedings of COLING*, 141–150.

Shaoul, C., and Westbury, C. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta*.

Socher, R.; Lin, C. C.; Manning, C.; and Ng, A. Y. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of ICML*, 129–136. ACM.

Williams, E. 1981. On the notions "lexically related" and "head of a word". *Linguistic inquiry* 245–274.

Zhou, G.; He, T.; Zhao, J.; and Hu, P. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of ACL*, 250–259.