Model-Free Context-Aware Word Composition

Bo An[2,3], **Xianpei Han**[1,2], **Le Sun**[1,2]
[1]Beijing Advanced Innovation Center for Language Resources, Beijing, China
[2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{anbo, xianpei, sunle}@iscas.ac.cn

Abstract

Word composition is a promising technique for representation learning of large linguistic units (e.g., phrases, sentences and documents). However, most of the current composition models do not take the ambiguity of words and the context outside of a linguistic unit into consideration for learning representations, and consequently suffer from the inaccurate representation of semantics. To address this issue, we propose a model-free context-aware word composition model, which employs the latent semantic information as global context for learning representations. The proposed model attempts to resolve the word sense disambiguation and word composition in a unified framework. Extensive evaluation shows consistent improvements over various strong word representation/composition models at different granularities (including word, phrase and sentence), demonstrating the effectiveness of our proposed method.

1 Introduction

Recent development in natural language processing (NLP) has seen a rise of continuous representation learning of linguistic units, which has been successfully applied to various tasks such as contextual word similarity (Iacobacci et al., 2015; Huang et al., 2012), machine translation (Devlin et al., 2014), paraphrase detection (Socher et al., 2011a) and sentiment classification (Tang et al., 2015). The continuous representation of linguistic units bring many benefits. For example, we can easily calculate the semantic relatedness between two words *'dog'* and *'cat'* using cosine distance of their continuous representations. Specifically, a representation learning method aims at mapping a linguistic unit with k words $S = [w_1, w_2, ..., w_k]$ into a continuous vector in a low dimensional feature space.

Currently, most of the representation learning methods focus on learning word embeddings, mostly based on the distributional hypothesis (Harris, 1954), i.e., *words in similar contexts tend to have similar meanings*. In most cases, a word is represented by summarizing all its contexts in a large text corpus, either by dimension reduction from co-occurrence matrix (Levy and Goldberg, 2014), or by neural network models (Collobert and Weston, 2008; Mikolov et al., 2013; Huang et al., 2012).

Intuitively, a word may express distinct meanings in different contexts, so its representations may be distinct in various linguistic units. For example, the word *'bank'* should have different representations in *'commercial bank'* and in *'river bank'*, because they correspond with two distinct senses of *'bank'*: the former with the meaning of *'financial institution'*, and the latter with *'riverside'* meanings. To address this issue, some multi-prototype word embeddings models are proposed to learn multiple representations for individual words. For instance, MSSG (Neelakantan et al., 2015) utilizes local context to infer the accurate semantic of a word, and the global context is incorporated by TWE (Liu et al., 2015) to learn accurate representations of a word.

The representations of linguistic units larger than words (e.g., phrases, sentences), unfortunately, usually cannot be learned using the method as word representations. The main reason is the context sparsity problem, i.e., most of the large linguistic units will not occur frequently even in a large text corpus, which leads to insufficient context statistics for accurate representations learning.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2834–2845
Santa Fe, New Mexico, USA, August 20-26, 2018.

To effectively learn the representations of large linguistic units, a promising approach is word composition, which is based on the Frege's principle of composition: (Pelletier, 2001) – *the meaning of a complex expression is determined by the meanings of their constituent expressions and the rules used to combine them.* Specifically, given a linguistic unit S with $[w_1, w_2, ..., w_k]$ words, a composition method first infers the vectors of all words in S based on embedding matrix. After that, a composition model is utilized to transform the word vectors into a single representation of unit *S*. For example, to represent the phrase *'the dog outside'*, a composition method will first infer the representations of *'the'*, *'dog'* and *'outside'*, and then feed those word vectors into a composition model, e.g., element-wise addition, element-wise multiplication (Mikolov et al., 2013), recursive neural network (Socher et al., 2012), gated recurrent neural network, Long-Short Term Memory network (Le and Zuidema, 2015) or convolutional neural network (Xu et al., 2015). There are also approaches that jointly learn the representations of words and larger units, e.g Le and Mikolov (2014).

However, the meaning of a large linguistic unit not only depends on its constituent words but also affected by the context around it. For example, the meaning of word *'going'* in sentence *'I am going to the bank.'* can be disambiguated based on the local context (the other words in the sentence). But the meaning of word *'bank'* in this sentence can not be inferred without context outside this sentence. Unfortunately, deriving contextual information outside of a given sentence is not trivial due to the data sparsity problem. What's worse, in some practical NLP applications, there is no context for a given sentence, such as question answering or information retrieval.

To address this issue, this paper proposes a model-free context-aware word composition model to learn proper representations for linguistic units at different granularity levels by incorporating latent semantic information. The proposed model attempts to address the word sense disambiguation and word composition in a unified architecture. Specifically, inspired by TWE (Liu et al., 2015), this paper utilizes topic distribution as the global context of a linguistic unit. Each topic is utilized to derive accurate meanings for all the word occurrences in the linguistic unit and to learn the topic-specific representation of the unit. After that, the context-aware representation of the linguistic unit is inferred by summarizing all its representations under different topics based on the topic distribution. In this way, our method can make use of the topic information of a word to learn its accurate topic-specific representation, and the topic distribution of the a unit is employed as a cue to guide the process of word composition to learn meaningful representation.

Note that, our context-aware composition framework is model-free in that it can employ any existing composition model as the base composition model. We assess our method on multiple text similarity tasks based on various base composition models. Experimental results verify the effectiveness of our model on learning the representations of linguistic units at different granularity level, and show consistently improvements over various base composition models. The main contribution are threefold: (1) It regards the linguistic unit as a whole, and utilizes the latent semantic information to learn the accurate representations for both the linguistic unit and the word occurrences in it; (2) This paper verifies that the topic information is benefit for both the accurate word representation and word composition. (3) We proposed a model-free framework that can enhance various kinds of methods.

2 Related Work

This section briefly reviews related work, including context representation models, word/phrase embeddings leanings methods and word composition methods.

2.1 Context Representation Model

How to represent contextual information is a key topic in natural language processing research area. Currently, a main trend is to represent context via latent variable models (Szpektor and Dagan, 2008), e.g., the Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Using LDA models, a context will be represented as a topic distribution. Frank et al. (2014) used topic models to provide context for a vowel categorisation task in child directed speech.

2.2 Word/Phrase Embeddings

Currently, most of the word embeddings methods use either neural network or co-occurrence matrix. Several well-known models include C&W (Collobert and Weston, 2008), word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The main drawback of these methods is that they represent a word using a single vector, such a uniform representation method cannot accurately represent polysemy.

There are also various methods which try to learn multi-prototype word embeddings to solve the polysemy problem, with each vector corresponding to a sense instead of word, such as the Multi-Prototype model (Huang et al., 2012), the MSSG (Neelakantan et al., 2015), the EHModel (Tian et al., 2014), the SenseEmb (Guo et al., 2014) and the SAMS model (Cheng and Kartsaklis, 2015). The main drawback of these models is that they need to induce the senses of a word, which is a very challenging task.

Salant and Berant (2017) verified the importance of the contextual information for word representation. Recently, a number of methods are proposed to learn contextual word representations based on neural network (Melamud et al., 2016; Peters et al., 2018) without considering word composition.

(Liu et al., 2015) proposed a TWE model which employs the topic assignments to produce topic-specific word embeddings to avoid the challenging task of word sense induction, and generates the representation of a larger unit by adding the vectors of topical word embeddings weighted by TFIDF of them. Fadaee et al. (2017) labels a word with topic distribution in both hard and soft ways, and learns topic-specific representations and general representation for a word simultaneously. Shi et al. (2017) introduces a method to enhance the word representation and topic model with each other. By contrast, our proposed model aims at utilizing topic distribution to enhance the word composition models for learning representations of linguistic units of different granularity levels, including word, phrase and sentence.

2.3 Word Composition

Due to the context sparsity problem, word composition is a promising technique to learn representations of linguistic units larger than words. A lot of models have been developed for word composition, e.g., simple models such as element-wise addition/multiplication (Mitchell and Lapata, 2010; Mikolov et al., 2013) and neural network based models such as recurrent neural network (Paulus et al., 2014), gated recurrent neural network, Long-Short Term Memory network (Le and Zuidema, 2015), convolution neural network (Xu et al., 2015) and attention-based model (Ling et al., 2015). These composition models could be used as the base composition model in our model-free framework.

Skip-Thought (Kiros et al., 2015) learns word and sentence representations based on sentences around it simultaneously. Doc2vec (Le and Mikolov, 2014) jointly trains the word sentence vectors by predicting the words in it. Recently, (Mao et al., 2017) introduced a topic-aware model to enhance the word composition model based on topic embeddings, which verifies the benefit of topic information for word composition. But despite its apparent success, there remains a major drawback: this method suffers from the limitations that learning topic embeddings and the uniform representation of words. By contrast, our method generates different vectors for a word in various context and only depends on the topic distributions of linguistic units.

3 Model-Free Context-Aware Word Composition Framework

In this section, we present our model-free context-aware word composition framework, which contains two main components: context-aware word representation and context-aware word composition. We first describe the method of learning a context-aware word embeddings. After that, we propose a model-free context-aware word composition framework to learn contextual representation of a linguistic unit.

3.1 Context-Aware Word Embeddings

This section describes how to learn the context-aware word representation of a contextual word based on contextual Information.

Contextual Information. In this paper, we represent contextual information using the most prevalent topic model (Blei et al., 2003), i.e., the context of a linguistic unit is represented as a topic distribution $C = \{p(Topic_1), p(Topic_2), ..., p(Topic_K)\}$. For example, the phrase context *'river bank'* of the word *'bank'* may be represented as $\{Geography^{0.8}, Financial^{0.01}, ..., Sport^{0.1}\}$, by contrast the phrase context *'commercial bank'* of the word *'bank'* may be represented as $\{Geography^{0.02}, Financial^{0.85}, ..., Sport^{0.03}\}$. Note that, the proposed method can also use other context representation model, such as Melamud et al. (2016; Peters et al. (2018).

Using the above context representation model, we estimate the context-aware representation as follows: (1) We train a topic model on a large-scale text corpus using topic model; (2) Given the context of a word occurrence $S = [w_1, w_2, ..., w_k]$, where w_i represents a word, we employ collapsed Gibbs sampling algorithm (Griffiths and Steyvers, 2004) to infer the topic assignments for all words in S, and the topic distribution of S is utilized as contextual information.

Context-Aware Word Embedding Learning: Based on the above contextual information, a word can express distinct meanings under different topic assignments, and the unit representation can be learned separately for each individual topic, before being combined to constitute the final context-aware unit representation. For example, the word *'bank'* with topic assignments *'Geography'* and *'Financial'* respectively express two distinct senses of *'bank'*: *'riverside'* and *'financial institution'*. We employ the topic assignment as the contextual cue of a word in a specific context. In this way, we need to learn the embeddings of all $\langle word : topic \rangle$ pairs – we refer them topic-specific word embeddings. Formally, we learn a set of vectors for each word, with each vector corresponding to a specific topic. For instance, the word *'bank'* will be represented as:

$$\vec{V}_{Geo}(bank) = [0.628, 0.093, 0.051, ..], \vec{V}_{Fin}(bank) = [0.034, 0.016, 0.320, ..]$$

where $\vec{V}_{Geo}(bank)$ and $\vec{V}_{Fin}(bank)$ correspond to two distinct senses of *'bank'* under topics of *'Geography'* and *'Financial'*, respectively.

To learn the above topic-specific word embeddings from a text corpus, we first assign each word in the text corpus with a topic using the topic models; secondly we treat each $\langle word : topic \rangle$ pair as an individual unit; finally we learn the embeddings of all $\langle word : topic \rangle$ pairs using word embedding model, such as Word2Vec and Glove.

3.2 Context-aware Word Composition

In this section, we propose a model-free context-aware word composition framework, which represents a linguistic unit based on contextual information and context-aware word representation. Our method employs topic distribution of a linguistic unit as its contextual information. Specifically, given a linguistic unit $S = [w_1, w_2, ..., w_n]$, its topic distribution $C = \{p(topic_1|S), p(topic_2|S), ...\}$ and the topic-specific embeddings of all words, we construct the representation of S as follows:

1) We calculate the topic-biased representation of S under topic *t* by composing the topic-specific word embeddings as follows:

$$\vec{V}_{TopicBiased}(S, t) = f(\vec{V}_t(w_1), ..., \vec{V}_t(w_n)) \tag{1}$$

where *f* is a base composition model, which composes a sequence of vectors into a single representation. Notice that we can utilize any context-unaware word composition model as our base composition model, such as recurrent neural network, convolutional neural network, element-wise addition/multiplication, etc. For example, given the sentence *'we run along the bank'*, the topic-biased representation of the sentence under topic *'Geography'* is learned as:

$$\vec{V}_{TopicBiased}(S, Geo) = f(\vec{V}_{Geo}(we), ..., \vec{V}_{Geo}(bank))$$

2) We construct the context-aware representation of S by the weighted average of all topic-biased representations of S based on its topic distribution:

$$\vec{V}_{Contextual}(S) = \sum_{t \in T} p(t|S) \cdot \vec{V}_{TopicBiased}(S, t) \tag{2}$$

where $p(t|S)$ is the topic probability of topic *t* of S. For instance, using the inferred topic distribution of the sentence S as $\{Geography^{0.03}, Financial^{0.01}, .., Sport^{0.73}\}$, we generate the context-aware representation of S by averaging all the topic-biased vectors of S according to its topic distribution as:

$$\vec{V}_{Contextual}(S) = 0.03 \cdot \vec{V}_{TopicBiased}(S, Geo) + ... + 0.73 \cdot \vec{V}_{TopicBiased}(S, Sport)$$

To learn the context-aware representation for a word in a specific sentence,the latent topic distribution of the sentence and the topic-specific word embeddings are needed. Specifically, a word in a specific context can be learned by averaging the topic-specific word embeddings based on the topic distribution:

$$\vec{V}_{Contextual}(w|S) = \sum_{t \in T} p(t|S) \cdot \vec{V}_t(w) \tag{3}$$

where $\vec{V}_{contextual}(w|S)$ represents the context-aware representation of word w in context S, $p(t|S)$ is the topic probability of S under topic t, $\vec{V}_t(w)$ represents the topic-specific embedding of word w and T represents the set of topics. For example, given a sentence *S='Bank is a financial institution that accepts deposits'*, we first infer its topic distribution of the sentence as $C = \{Geography^{0.01}, Financial^{0.8}, ..., Sport^{0.03}\}$. Secondly we learn the context-aware representation of *'bank'* as follows:

$$\vec{V}_{Contextual}(bank|S) = 0.01 \cdot \vec{V}_{Geo}(bank) + 0.8 \cdot \vec{V}_{Fin}(bank) + 0.03 \cdot \vec{V}_{Sport}(bank) \tag{4}$$

where $\vec{V}_{sport}(bank)$ represents the topic-specific representation of *'bank'* under topic *'Sport'*.

4 Experiments

In this section, we assess our method on text similarity tasks at different granularities. And we conduct experiments on paraphrase detection task to evaluate the improvements of our method over various base word composition models.

4.1 Model Pre-training

In our experiments, we employ two different corpora for a comprehensive comparison to the various baseline models. We train the LDA model and the topic-specific word embeddings on the British National Corpus (BNC) (Consortium and others, 2007), which contains more than 93 million terms, and a bigger corpus – a snapshot of the English Wikipedia corpus[1] with about 990 million tokens.

Specifically, we use GibbsLDA++ (Phan and Nguyen, 2007) to estimate the topic distribution and infer the topic assignment for each word in the corpus. The parameters of GibbsLDA++ are empirically tuned as follows: $\alpha = 0.5, \beta = 0.1$, the number of topics 50, the number of iterations 400. For the topic-specific word embeddings, we use the SkipGram algorithm, with the dimension of word vector 300, the windows size 5, the number of iterations 5 and 10 negative samples per occurrence.

4.2 Overall Results

The proposed model focuses on learning the context-aware representations of large linguistic units. To evaluate the learned representations of linguistic units, we conduct experiments on text semantic similarity tasks at different granularities, including contextual word similarity, phrase similarity and sentence similarity. In this section, we calculate the similarity of two linguistic units based on cosine similarity of their representations, and we assess different systems using the Spearman's correlation between system outputs and gold standards manually labelled.

Contextual Word Similarity Results

To assess the quality of our context-aware word representations, we conduct experiments on the Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012), which contains 2003 manually labelled word pair similarities, with each word is paired with a sentence context.

We compare our context-aware word representation model with several baseline word embeddings models, including C&W (Collobert and Weston, 2008), CBOW and SkipGram (SG) (Mikolov et al., 2013). We also compare with two multi-prototype word embeddings models, including MSSG model (Neelakantan et al., 2015) and SAMS model (Cheng and Kartsaklis, 2015), both of which predict the

[1] https://www.wikipedia.org/

sense of a word based on its local context. All of the above models are trained on BNC. In addition, we compare with CBOW and SkipGram (SG) (Mikolov et al., 2013), Huang (Huang et al., 2012), Tian (Tian et al., 2014), TWE (Liu et al., 2015), STD-dif (Shi et al., 2017) and HTLE (Fadaee et al., 2017) models based on Wikipedia corpus for fair comparison, which have achieved state-of-the-art performances. And we report their best published result. The overall results are presented in Table 1.

Corpus	CBOW	SK	C&W	MSSG	SAMS	Tian	Huang	TWE(best)	STE-Dif	HTLE	Our
BNC	59.0	61.0	55.0	56.0	58.0	-	-	-	-	-	**63.2**
Wikipedia	65.3	65.7	-	-	-	65.4	65.3	68.1	68.0	63.0	**68.3**

Table 1: The Spearman's correlation ρ of different methods on SCWS dataset.

From Table 1, we can see that: (1) Our context-aware word representation had achieved the best performance on both the BNC and Wikipedia corpora. Compared with the three context-unaware word representation baselines CBOW, SkipGram and C&W, our method correspondingly achieves 6.8%, 3.3% and 14.5% ρ improvements. (2) The latent topic model based context representation can effectively capture the proper meanings of a contextual word. Compared with the multi-prototyped word embeddings baselines MSSG and SAMS, our method achieved 12.5% and 8.6% ρ improvements, respectively. We believe the topic distribution of the entire sentence is a beneficial cue for accurately representing polysemy words. (3) The results verify that the context-aware way of learning word representation as Formula (3) may provide a better way of utilizing topic distribution than hard label schema as in TWE, because both of the models implemented on the dataset.

Phrase Similarity Results

To assess the performance of our method on phrase representation learning, we conduct phrase similarity experiments on the ML2010 dataset (Mitchell and Lapata, 2010), which contains 108 pairs of phrases for adjective-noun (AN), verb-object (VO) and compound-noun (NN) respectively.

In our experiments, we employ the element-wise addition as base composition model, which turns out to be both robust and effective in many tasks. We compare our system with two baselines: the element-wise addition models correspondingly using the context-unaware word embeddings learned by CBOW and SkipGram. We also compare our system with state-of-the-art results from ML Original (Mitchell and Lapata, 2010), PAS(Hashimoto et al., 2014), PARAGRAM (Wieting et al., 2015b), DeepCCA (Lu et al., 2015) and vecDCS (Tian et al., 2016). The results of the models are taken from the original papers. The overall results are shown in Table 2.

	ML Original	PAS	PARAGRAM	DeepCCA	vecDCS	SkipGram	CBOW	Our
AN	0.46	0.46	0.50	0.45	0.41	0.47	0.45	**0.60**
NN	0.37	0.49	0.51	0.45	0.51	0.51	0.49	**0.54**
VO	0.45	0.45	0.40	0.47	0.49	0.39	0.40	**0.47**
ALL	0.44	0.47	0.47	0.46	0.47	0.41	0.43	**0.54**

Table 2: The Spearman's correlation ρ of different methods on ML2010 datasets.

From Table 2, we can see that: (1) By incorporating the topic distributions of phrases, our context-aware word composition model achieves the best results on all three types of phrases: compared with SkipGram and CBOW baselines, our method achieves 22.2% and 20.8% ρ improvements on average. (2) Compared with state-of-the-art systems, our method achieves the best performances on all types of phrases. The result verified the effectiveness of contextual information in phrase representation.

Sentence Similarity Results

To assess our method on sentence representations, we conduct experiments on the semantic textual similarity dataset (STS2015) from SemEval 2015, which contains five datasets of different genres: answers-students, answers-forums, belief, headlines and images.

In our experiments, the element-wise addition method is employed as the base composition model for our context-aware word composition model. We compare our method with two baselines: the element-wise addition composition model using word embeddings learned from CBOW and SkipGram. All of the models in this experiment is trained based on Wikipedia corpus. We also compare to several strong baselines: skip-thought vector (ST) (Kiros et al., 2015) and average Glove vector (Pennington et al., 2014). Because our proposed model aims at improving word composition model based on topic distribution, we don't compare with the methods which incorporate syntax information or other resources (Wieting et al., 2015a). The overall results are showed in Table 3.

From Table 3, we can note that: our context-aware word composition method significantly improved the performances on all of the datasets by incorporating the topic distribution information. Compared with the baselines CBOW and SkipGram, our method achieved 35.5% and 19.2% ρ improvements on average. And the topic distribution is a beneficial resource for modelling the context of a sentence for learning better representation of the sentence.

Model	Answers students	Answers forums	Beliefs	Headlines	Images	Overall
ST	0.361	0.330	0.246	0.436	0.177	0.310
GloVe	0.305	0.630	0.405	0.618	0.675	0.527
SkipGram	0.413	0.475	0.392	0.581	0.621	0.496
CBOW	0.477	0.620	0.467	0.573	0.684	0.564
Our	**0.632**	**0.631**	**0.657**	**0.681**	**0.761**	**0.672**

Table 3: The Spearson's correlation ρ for different methods on STS2015 datasets.

All of the previous experiments have verified that the context-aware combination model is effective and can achieve better results than the baseline models. Furthermore, the proposed context-aware model achieved the greatest improvements over base models at sentence level, and made the least progress on word level. We suspect that because the local contextual information is utilized in learning word representation, which weakens the value of global context. By contrast, the contextual information is rarely taken into consideration by word composition models. As a result, we believe the global contextual information is beneficial for learning representations of larger linguistic units, such as sentence.

4.3 The Effect of Using Different Base Composition Models

Our method is model free that it can employ any context-unware composition models as its base word composition models. In this section, we intend to clarify whether the context-aware method could enhance the representations on different base composition models. Concretely, we conduct experiments on four of the most commonly used word composition models as base models: element-wise addition (Mikolov et al., 2013), recurrent neural network (RNN) (Mikolov et al., 2010; Socher et al., 2011b), gated recurrent neural network (GNN) (Tang et al., 2015) and Long-Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997; Le and Zuidema, 2015; Ghosh et al., 2016) on paraphrase detection task.

Paraphrase Detection

Paraphrase detection is a binary classification task which is used to decide whether two sentences express the same meanings, which is an appropriate task for evaluating our method. In this paper, we employ MSRPC datasets (Dolan et al., 2004).

Model	**C-unware**		**C-aware**	
	Acc	F	Acc	F
Addition	0.658	0.765	0.699	0.801
RNN	0.682	0.80	0.732	0.822
GNN	0.735	0.832	0.763	0.841
LSTM	0.739	0.838	**0.776**	**0.852**
ST	0.642	0.739	-	-
MSSG	-	-	0.692	0.793
SAMS	-	-	**0.786**	**0.853**

Table 4: Accuracies and F-Scores of paraphrase detection on MSRPC datasets.

For each base composition model, we compare our method with its context-unaware version. In addition, we compare with Skip-Thought (ST) vector as a strong baseline, MSSG and SAMS are used

Topic	Top 4 Frequent Words of Topic
Financial	money, million, cost, tax
Geography	river, lake, mountain, island
Information	user, systems, ibm, software
Sport	play, team, season, ball

Table 5: Top 4 topics of the word *‘bank’*.

⟨ word:topic ⟩	Neighbors
bank:Geography	longitude:Geography, inland:Geography, coasts:Geography, southward:Geography
bank:Financial	bankers: Financial, debt: Financial, loans: Financial, finance: Financial

Table 6: The top 4 nearest neighbors of word *‘bank’* with topics *‘Geography’* and *‘Financial’*.

as the multi-prototype word embeddings models for comparison. In this experiment, we implement a two layers RNN/GNN/LSTM with a max pooling layer as base word composition model. We randomly select 500 sentence pairs in training data as validation data to find the best parameters. We learn the best hyper-parameters based on validation data as follows: the hidden units for RNN/GNN/LSTM is 300, the mini-batch size as 100, the dropout rate is 0.4, the learning rate of SGD is 0.01. All of the word embeddings are trained on BNC (Consortium and others, 2007) for fair composition with SAMS. Accuracy and F-Score are used as the metrics. The results are shown in Table 4.

From Table 4, we can see that:(1) The context-aware representation consistently improved the performances on all the base word composition models. (2) Compared with Skip-thought vector model and MSSG model, our proposed model had achieved better performance, we believe the latent semantic information of a sentence may be the main reason that leads to the results. (3) SAMS achieved better results than our model, however, it utilized external knowledge resource (PPDB), which is a larger dataset for enhancing the performance on paraphrase detection task. The paraphrase detection results demonstrate that our context-aware representation brings beneficial information for different base word composition models and improves the performances on this task. The proposed model may achieve better results by utilizing more advanced base word composition models, we leave it for future work.

5 Detailed Analysis

To better understand the way of the proposed method works, this section provides detailed analysis on the quality of the topic-specific word embeddings and context-aware representations.

5.1 The Quality of Topic-Specific Word Representations

Topic-specific word representations are essential to our method, which is the basic unit for composing representations of larger linguistic units. In this section, we analyze the quality of topic distributions and topic-specific word embeddings.

For the reason that the proposed context-aware model is based on an assumption that a word can express distinct meanings with different topic assignments, so it is important to estimate whether topics can distinguish different senses of a word. Table 5 shows the top 4 topics of word *‘bank’* and the most frequent words of these topics. From Table 5, we can infer that, *‘bank’* with topic of *‘Financial’* associates with the meanings of *‘money’* or *‘tax’*, and *‘bank’* with *‘Geography’* topic associates with the meanings of *‘river’* or *‘lake’*. In addition, We list the top 4 $\langle word : topic\rangle$ neighbors for *‘bank:Geography’* and *‘bank:Financial’* in Table 6. The conclusion can be drawn that the topic information can effectively distinguish distinct senses of a word.

5.2 The Quality of Context-Sensitive Word Composition

We analyze the quality of linguistic unit representations using the *‘river bank’* and *‘commercial bank’* as examples. We first present the top 3 most frequent topics of their words in Table 7.

Word	Topic Probabilities
bank	Financial(62.7%), Geography(23.4%), Information(7.3%)
river	Geography(89.7%), Information(0.07%) , Sport(0.01%)
commercial	Financial(86.3%), Information (6.5%), Law(5.0%)

Table 7: The top 3 most frequent topics of *‘river’*, *‘bank’* and *‘commercial’*.

From Table 7, we can see that, both the word *'commercial'* and *'bank'* have high probability on topic *'Financial'*, so *'commercial bank'* will have a high probability belonging to topic *'Financial'*. In this way, the representation of *'commercial bank'* will mainly depend on the topic-biased representation under *'Financial'* topic, which is composed from topic-specific embeddings of *'commercial:Finanial'* and *'bank:Financial'*. By contrast, the representation of *'river bank'* will mainly depend on the embeddings of *'river:Geography'* and *'bank:Geography'*. In a word, the proposed model is capable of correctly identifying the important topics for a specific phrase, and generate proper representation for it.

In addition, we list top 5 nearest $\langle word : topic \rangle$ pairs of the phrases in Table 8. As shown in Table 8, all of the nearest neighbors of *'river bank'* are labelled with *'Geography'* topic and all of the nearest neighbors of *'commercial bank'* are related to its *'Financial'* topic. This result demonstrates the effectiveness of our method on learning accurate representations of phrases.

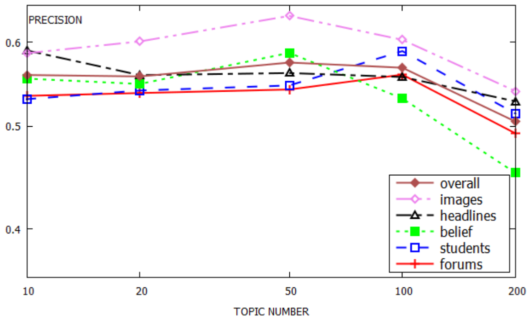

Figure 1: The precisions on STS2015 datasets with different numbers of topics in LDA model.

Phrase	$\langle word : topic \rangle$ **neighbors**
river bank	river:Geography, valley:Geography, park:Geography, north:Geography, boat: Geography
commercial bank	bank:Financial, leasing:Financial, merchandising: Financial, retailing: Financial, financing: Financial

Table 8: The top 5 nearest $\langle word : topic \rangle$ pairs for *'river bank'* and *'commercial bank'*.

The experimental results reveal that the topic distribution information could be used as the cue for learning proper representations of a linguistic unit in different contexts. And the proposed context-aware model can generate accurate representations for phrases.

5.3 The Number of Topics

One important parameter of our method is the number of topics of the LDA model. In order to detect the impact of the number of topics on our model, we conduct experiments using different topic numbers (including 10, 20, 50, 100 and 200) on STS2015 datasets with element-wise addition as base composition model, the results are shown in Figure 1.

From Figure 1, we can see that: (1) Our method is not very sensitive to the topic number, which achieved stable performances when using topic numbers range from 10 to 100. But the precision drops quickly when the number of topics is set as 200, we believe such large number of topics inevitably causes insufficient occurrences for learning topic-specific word embeddings, which is crucial basic units for our compositional method. (2) Our method achieved the best precision when the topic number is 50. We believe this is actually a trade-off between distinguishability and data sparsity.

6 Conclusions

In this paper, we have proposed a model-free context-aware word composition framework for text representation learning, which can effectively utilize the latent semantic information of the whole linguistic unit for learning context-aware representations for linguistic units. The proposed model is model-free in that our method can employ various context-unaware word composition models as the base model within our proposed framework. Experimental results demonstrated the effectiveness of our method on text representation learning at different granularities, including word, phrase and sentence, and the improvements on various base word composition models. In future work, to further improve the learned representations of linguistic units, we want to also take the non-compositional units into consideration, i.e., to solve the representations of multiword expressions like idioms or name entities, whose meanings cannot be composed from the meanings of its constituent parts, such as *kill the goose that lays the golden eggs* and *speak of the devil*.

Acknowledgments

This work is supported by the Project of Beijing Advanced Innovation Center for Language Resources (451122512), the National Natural Science Foundation of China under Grants no. 61433015, 61772505 and 61572477, and the Young Elite Scientists Sponsorship Program no. YESS20160177. Moreover, we sincerely thank the reviewers for their valuable comments.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jianpeng Cheng and Dimitri Kartsaklis. 2015. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. *arXiv preprint arXiv:1508.02354*.

Ronan Collobert and Jason Weston. 2008. 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.

British National Corpus Consortium et al. 2007. British national corpus version 3 (bnc xml edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved February*, 13:2012.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *ACL (1)*, pages 1370–1380. Citeseer.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *International Conference on Computational Linguistics*, page 350.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Learning topic-sensitive word representations.

Stella Frank, Naomi H Feldman, and Sharon Goldwater. 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1073–1083.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*, pages 497–507.

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

Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. 2014. Jointly learning word representations and composition functions using predicate-argument structures. In *Conference on Empirical Methods in Natural Language Processing*, pages 1544–1555.

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

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *ACL (1)*, pages 95–105.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *Computer Science*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Computer Science*, 4:1188–1196.

Phong Le and Willem Zuidema. 2015. Compositional distributional semantics with long short term memory. *arXiv preprint arXiv:1503.02510*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu Cheng Lin. 2015. Not all contexts are created equal: Better word representations with variable attention. In *Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *AAAI*, pages 2418–2424.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Kezhi Mao, Kezhi Mao, Kezhi Mao, and Kezhi Mao. 2017. Topic-aware deep compositional models for sentence classification. *IEEE/ACM Transactions on Audio Speech & Language Processing*, 25(2):248–260.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Signll Conference on Computational Natural Language Learning*, pages 51–61.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

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

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Romain Paulus, Richard Socher, and Christopher D Manning. 2014. Global belief recursive neural networks. In *Advances in Neural Information Processing Systems*, pages 2888–2896.

Francis Jeffry Pelletier. 2001. Did frege believe frege's principle? *Journal of Logic, Language and information*, 10(1):87–114.

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

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda). *URL: http://gibbslda. sourceforge. net*.

Shimi Salant and Jonathan Berant. 2017. Contextualized word representations for reading comprehension. *arXiv preprint arXiv:1712.03609*.

Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. 2017. Jointly learning word embeddings and latent topics.

Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, volume 24, pages 801–809.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011b. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 849–856. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pages 151–160.

Ran Tian, Naoaki Okazaki, and Kentaro Inui. 2016. Learning semantically and additively compositional distributional representations. In *Meeting of the Association for Computational Linguistics*, pages 1277–1287.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015a. Towards universal paraphrastic sentence embeddings. *Computer Science*.

John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015b. From paraphrase database to compositional paraphrase model and back. *Computer Science*, pages 98–104.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of NAACL-HLT*, pages 62–69.