Idiom-Aware Compositional Distributed Semantics

Pengfei Liu Kaiyu Qian Xipeng Qiu[*] Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{pfliu14,kyqian, xpqiu,xjhuang}@fudan.edu.cn

Abstract

Idioms are peculiar linguistic constructions that impose great challenges for representing the semantics of language, especially in current prevailing end-to-end neural models, which assume that the semantics of a phrase or sentence can be literally composed from its constitutive words. In this paper, we propose an idiom-aware distributed semantic model to build representation of sentences on the basis of understanding their contained idioms. Our models are grounded in the literal-first psycholinguistic hypothesis, which can adaptively learn semantic compositionality of a phrase literally or idiomatically. To better evaluate our models, we also construct an idiom-enriched sentiment classification dataset with considerable scale and abundant peculiarities of idioms. The qualitative and quantitative experimental analyses demonstrate the efficacy of our models. The newly-introduced datasets are publicly available at `http://nlp.fudan.edu.cn/data/`

1 Introduction

Currently, neural network models have achieved great success for many natural language processing (NLP) tasks , such as text classification (Zhao et al., 2015; Liu et al., 2017), semantic matching (Liu et al., 2016a,b), and machine translation (Cho et al., 2014). The key factor of these neural models is how to compose a phrase or sentence representation from its constitutive words. Typically, a shared compositional function is used to compose word vectors recursively until obtaining the representation of the phrase or sentence. The form of compositional function involves many kinds of neural networks, such as recurrent neural networks (Hochreiter and Schmidhuber, 1997; Chung et al., 2014), convolutional neural networks (Collobert et al., 2011; Kalchbrenner et al., 2014), and recursive neural networks (Socher et al., 2013; Tai et al., 2015; Zhu et al., 2015).

However, these methods show an obvious defect in representing idiomatic phrases, whose semantics are not literal compositions of the individual words. For example, "`pulling my leg`" is idiomatic, and its meaning cannot be directly derived from a literal combination of its contained words. Due to its importance, some previous work focuses on automatic identification of idioms (Katz and Giesbrecht, 2006; Li and Sporleder, 2009; Fazly et al., 2009; Peng et al., 2014; Salton et al., 2016). However, challenge remains to take idioms into account to improve neural based semantic representations of phrases or sentences.

Motivated by the **literal-first** psycholinguistic hypothesis proposed by Bobrow and Bell (1973), in this paper, we propose an end-to-end neural model for idiom-aware distributed semantic representation, in which we adopt a neural architecture of recursive network (Socher et al., 2013; Tai et al., 2015; Zhu et al., 2015) to learn the compositional semantics over a constituent tree. More concretely, we introduce a neural idiom detector for each phrase in a sentence to adaptively determine their compositionality: literal or idiomatic manner. For the literal phrase, we compute its semantics from its constituents while for the idiomatic phrase, we design two different ways to learn representations of idioms grounded in two different linguistic views of idioms (Katz, 1963; Fraser, 1970; Nunberg et al., 1994).

To evaluate our models towards the ability to understand sentences with idioms, we conduct our

[*] Corresponding author.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1204–1213
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

experiments on sentiment classification task due to the following reasons: 1) Idioms typically imply an affective stance toward something and they are common in reviews and comments (Williams et al., 2015). 2) The error analysis of sentiment classification results reveals that a large number of errors are caused by idioms (Balahur et al., 2013).

The contributions of this work are summarized as follows:

- We grow the capacity of recursive neural network, enabling it to model idiomatic phrases and handle ubiquitous phenomenon of idiomatic variations when learning a sentential representation.
- We integrate idioms understanding into a real-world NLP task instead of evaluating idiom detection as a standalone task.
- We construct a new real-world dataset covering abundant idioms with original and variational forms. The elaborate qualitative and quantitative experimental analyses show the effectiveness of our models.

2 Linguistic Interpretation of Idioms

Recently, idioms have raised eyebrows among linguists, psycholinguists, and lexicographers due to their pervasiveness in daily discourse and their fascinating properties in linguistics literature (Villavicencio et al., 2005; Salton et al., 2014). As peculiar linguistic constructions (Villavicencio et al., 2005; Salton et al., 2014), idioms have three following properties:

Invisibility Idioms always disguise themselves as normal multi-words in sentences. It makes end-to-end training hard since we should detect idioms first, and then understand them.

Idiomaticity Idioms are semantically opaque, whose meanings cannot be derived from their constituent words. Existing compositional distributed approaches fail due to the hypothesis that the meaning of any phrase can be composed of the meanings of its constituents.

Flexibility While structurally fixed, idioms allow variations. The words of some idioms can be removed or substituted by other words.

Table 1 shows the three properties of idioms and the resulting challenges for distributed compositional semantics. To address these challenges, two different perspectives have been held for idiom comprehension.

Property	**Challenges**	**V1**	**V2**
Invisibility	End-to-end training is difficult	✓	✓
Idiomaticity	Difficult to predict meanings of idioms	×	✓
Flexibility	Hard to handle variation and generalize	✓	×

Table 1: The main properties of idioms and corresponding challenges. **V1** and **V2** represent two different linguistic perspectives towards idiom comprehension: arbitrary and compositional perspectives. ✓ indicates the perspective suffers from the corresponding challenge.

The first perspective treats idioms as long words whose meanings are stipulated arbitrarily and can not be predict from its constituent (Katz, 1963; Fraser, 1970). However, a lot of idioms have shown certain degree of flexibility in term of morphology and lexeme, so this kind of method handles variation badly and fails to generalize.

The second perspective considers idioms as linguistic expressions (Nunberg et al., 1994), whose meanings are determined by the meanings of their constituent parts and some compositional rules can be used to combine them. This fully compositional view may handle lexical variations, but it suffers from the idiomaticity problem, for the meanings of idioms are opaque.

3 Proposed Models

We propose an end-to-end neural model for idiom-aware distributed semantic representation. Specifically, in terms of invisibility, we introduce a neural idiom detector to adaptively distinguish literal and idiomatic meaning of each phrase when learning sentence representations. For the literal phrase, we compute its semantics from its constituents with Tree-structured LSTM (TreeLSTM) (Tai et al., 2015; Zhu et al., 2015). For the idiomatic phrase, we design two different ways to learn representations of idioms grounded in two different linguistic views of idioms, which considers the idiomaticity and flexibility properties of idioms. Figure 1 illustrates the framework of our proposed models, which consist of three modules: literal interpreter, idiom detector and idiomatic interpreter.

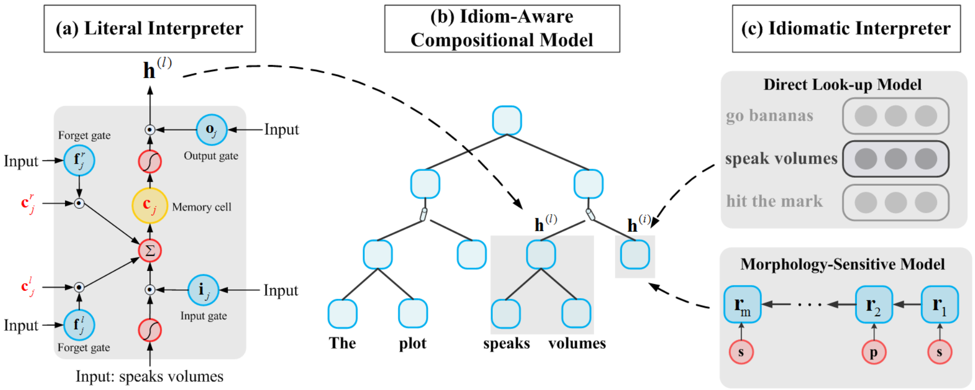

Figure 1: Illustration of different modules for our proposed idiom-aware composition network corresponding to the sentence "`The plot speaks volumes`"

3.1 Literal Interpreter

The literal interpreter is basically a compositional semantic model, in which the semantics of a phrase is composed by literal meanings of its constituent words. Several existing models could serve as literal interpreter. In this paper, we adopt TreeLSTM (Tai et al., 2015) due to its superior performance.

Formally, given a binary constituency tree T induced by a sentence, each non-leaf node corresponds to a phrase. We refer to $\mathbf{h}_j$ and $\mathbf{c}_j$ as the hidden state and memory cell of each node j. The transition equations of node j are as follows:

$$\begin{bmatrix} \tilde{\mathbf{c}}_j \\ \mathbf{o}_j \\ \mathbf{i}_j \\ \mathbf{f}^l_j \\ \mathbf{f}^r_j \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} T_{\mathbf{A},\mathbf{b}} \begin{bmatrix} \mathbf{x}_j \\ \mathbf{h}^l_j \\ \mathbf{h}^r_j \end{bmatrix}, \tag{1}$$

$$\mathbf{c}_j = \tilde{\mathbf{c}}_j \odot \mathbf{i}_j + \mathbf{c}^l_j \odot \mathbf{f}^l_j + \mathbf{c}^r_j \odot \mathbf{f}^r_j, \tag{2}$$

$$\mathbf{h}_j = \mathbf{o}_j \odot \tanh\left(\mathbf{c}_j\right), \tag{3}$$

where $\mathbf{x_j}$ denotes the input vector and is non-zero if and only if it is a leaf node. The superscript l and r represent the left child and right child respectively. σ represents the logistic sigmoid function and $\odot$ denotes element-wise multiplication. $T_{\mathbf{A},\mathbf{b}}$ is an affine transformation which depends on parameters of the network $\mathbf{A}$ and $\mathbf{b}$. Figure 1-(a) gives an illustration of TreeLSTM unit.

3.2 Idiom Detector

Despite the success of TreeLSTM, there is still existing potential weakness of the hypothesis that the meaning of a phrase or a sentence can be composed from the meanings of its constituents. Previous neural sentence models are poor at learning the meanings of idiomatic phrases, not to mention modeling the idiomatic variations.

Therefore, we introduce a parameterized idiom detector, which is used for detecting the boundary between literal and idiomatic meanings. Specifically, if a compositional interpretation is nonsensical in the context of a sentence, then the detector is supposed to check whether an idiomatic sense should be taken and whether it makes sense. This **literal-first** model of idiom comprehension is motivated by the psycholinguistic hypothesis proposed by Bobrow and Bell (1973).

Due to ignoring the context information, TreeLSTM suffers from the problem of disambiguation. For example, the phrase "`in the bag`" is compositional in the sentence "`The dictionary is in the bag`" while it has idiomatic meaning in the sentence "`The election is in the bag unless the voters find out about my past`". To address this problem, we explicitly model the context representation and integrate it into the process of sentence composition.

Context Representation More concretely, for each non-leaf node i and its corresponding phrase p_i, we define C as a word set which contains words surrounding the phrase p_i. Then the context representation s_i can be obtained as follow:

$$\mathbf{s}_i = f(c_1, c_2, ..., c_k; \theta) \tag{4}$$

where f is a function with learnable parameter θ. Here, the function is implemented in two approaches, NBOW and LSTM.

Detector The detector outputs a scalar α to determine whether the meaning of a phrase is literal or idiomatic. Formally, for the phrase i (non-leaf node i) with its context information $\mathbf{s}_i$ and literal meaning $\mathbf{h}_i^{(l)}$, we compute the semantic compositional score α_i using a single layer multilayer perceptron.

$$\alpha_i = \sigma(\mathbf{v}_s^T \tanh(\mathbf{W}_s[\mathbf{h}_i^{(l)}, \mathbf{s}_i])), \tag{5}$$

where $\mathbf{W}_s \in \mathbb{R}^{d \times 2d}$ and $\mathbf{v}_s \in \mathbb{R}^d$ are learnable parameters.

3.3 Idiomatic Interpreter

Idiomatic phrases pose a clear challenge to current compositional models of language comprehension. However, until recently, there is little investigation of learning idiomatic phrases in a real-world task. Here, based on different views of idioms (Katz, 1963; Fraser, 1970; Nunberg et al., 1994), we propose two idiomatic interpreters to model them.

Direct Look-Up Model Inspired by the direct access theory for idiom comprehension (Glucksberg, 1993), in this model, once a phrase p is detected as an idiom, it will be regarded as a long word like a key, and then their meanings can be directly retrieved from an external memory M, which is a table and stores idiomatic information for each idiom as depicted in the top subfigure in Figure 1-(c). Formally, the idiomatic meaning for a phrase can be obtained as:

$$\mathbf{h}^{(i)} = \mathbf{M}[k] \tag{6}$$

where k denotes the index of the corresponding phrase p.

Morphology-Sensitive Model Since most idioms enjoy certain flexibility in morphology, lexicon, syntax, the above model suffers from the problem of idiom variations. To remedy this, inspired by the compositional view of idioms (Nunberg et al., 1994) and recent success of character-based models (Kim et al., 2016; Lample et al., 2016; Chung et al., 2016), we propose to use CharLSTM to directly encode the meaning of a phrase in an idiomatic space and generate an idiomatic representation, which is not contaminated by its literal meaning and sensitive to different variations.

Formally, for each non-leaf node i and its corresponding phrase p_i in a constituency tree, we apply charLSTM to phrase p_i as depicted in the bottom subfigure in Figure 1-(c) and utilize the emitted hidden states $\mathbf{r}_j$ to represent the idiomatic meaning of the phrase.

$$\mathbf{r}_j = \textbf{Char-LSTM}(\mathbf{r}_{j-1}, \mathbf{c}_{j-1}, \mathbf{x}_j) \tag{7}$$

where $j = 1, 2, \cdots, m$ and m represents the length of the input phrase.

Then, we can obtain the idiomatic representation:

$$\mathbf{h}^{(i)} = \mathbf{r}_m \tag{8}$$

After obtaining the literal or idiomatic meanings, we can compute the final representation for phrase p_i:

$$\mathbf{h}_i = \alpha_i \mathbf{h}_i^{(i)} + (1 - \alpha_i)\mathbf{h}_i^{(l)} \tag{9}$$

3.4 Analysis of Two Proposed Idiomatic Interpreters

Given a phrase, both interpreters can generate a corresponding semantic representation, which is not contaminated by its literal meaning. The difference is that Look-Up Model takes a totally non-compositional view that the meanings of idioms can be directly accessed from an external dictionary. This straightforward retrieval mechanism is more efficient and can introduce external prior knowledge by utilizing pre-trained external dictionary.

By contrast, Morphology-Sensitive Model holds the idea that idiomatic meanings can still be composed in an idiomatic space, which allows this model to understand idioms better in terms of flexibility. Besides, this kind of model does not require an extra dictionary.

4 iSent: A Benchmark for Idiom-Enriched Sentiment Classification Dataset

To evaluate our models, we need a task that heavily depends on the understanding of idioms. In this paper, we choose sentiment classification task due to following reasons: 1) Idioms typically imply an affective stance toward something and they are common in reviews and comments (Williams et al., 2015). 2) The error analysis of sentiment classification results reveals that a large number of errors are caused by idioms (Balahur et al., 2013).

In this section, we will first give a brief description of the most commonly used datasets for sentiment classification so as to motivate the need for a new benchmark dataset.

Dataset	Train	Dev.	Test	Class	L_{avg}	$\|\mathcal{V}\|$
MR	9596	-	1066	2	22	21K
SST-1	8544	1101	2210	5	19	18K
SST-2	6920	872	1821	2	18	15K
SUBJ	9000	-	1000	2	21	21K

Table 2: Statistics of the four mainstream datasets for sentiment classification. L_{avg} denotes the average length of documents; $|\mathcal{V}|$ denotes the size of vocabulary.

4.1 Mainstream Datasets for Sentiment Classification

We list four kinds of datasets which are most commonly used for sentiment classification in NLP community. Additionally, we also evaluate our models on these datasets to make a comparison with many recent proposed models. Each dataset is briefly described as follows.

- **SST-1** The movie reviews with five classes (negative, somewhat negative, neutral, somewhat positive, positive) in the Stanford Sentiment Treebank[1] (Socher et al., 2013).
- **SST-2** The movie reviews with binary classes. It is also derived from the Stanford Sentiment Treebank.
- **MR** The movie reviews with two classes [2](Pang and Lee, 2005).
- **SUBJ** Subjectivity data set where the goal is to classify each instance (snippet) as being subjective or objective. (Pang and Lee, 2004)

The detailed statistics about these four datasets are listed in Table 2.

4.2 Reasons for a New Dataset

Differing from previous work, which evaluating idiom detection as a standalone task, we want to integrate idiom understanding into sentiment classification task. However, most of existing sentiment datasets do not cover enough idioms or related linguistic phenomenon. To better evaluate our models on idiom understanding task, we proposed an idiom-enriched sentiment classification dataset, in which each sentence contains at least one idiom.

Additionally, considering most idioms have certain flexibility in morphology, lexicon and syntax, we enrich our dataset by introducing different types of idiom variations so that we can further evaluate the ability that the model handle different idiomatic variations. As shown in Table 3, we sum up two types of phenomena towards idiom variations and, for each variation, we obtain several corresponding sentences from a large corpora.

4.3 Data collection

We crawl the website `rottentomatoes.com` to excerpt movie reviews with corresponding scores and collect the idioms from dictionary (Flavell and Flavell, 2006). The idiom dictionary contains lexical variations while has no morphology variations. To address this problem, we manually annotate the morphological variation of each idiom in term of verb(tense), noun(plural or singular).

Then we filter these movie reviews with idioms ensuring that each sentence covers at least one idiom. After that, we obtain nearly 15K movie reviews covering 1K idioms. To further improve the quality of these idiom-enriched sentences, we take some strategies to filter the dataset and finally construct 13K idiom-enriched sentences.

- If the occurrence of an idiom in all the reviews is less than 3, we threw this idiom and corresponding reviews. [3]
- We find some "idioms" in sentences are movie names rather than expressing idiomatic meanings and we filtered this kind of noise.

4.4 Statistics

The iSent dataset finally contains 9752 training samples , 1020 development samples and 2003 test samples. Besides, the development and test sets cover different types of idiom variations allowing us to test the model's generalization. Table 4 shows the detailed statistics and Figure 2 shows the distribution of the number of reviews over different lengths.

5 Experiment

We first evaluate our proposed models on four popular sentiment datasets, so that we can make a comparison with varieties of competitors. And then, we use the newly-introduced dataset to make more detailed experiment analyses.

[1] `http://nlp.stanford.edu/sentiment.`

[2] `https://www.cs.cornell.edu/people/pabo/movie-review-data/.`

[3] The reason is that, for some idioms, we should split their corresponding reviews into train/dev/test sets.

Variations		Examples
Morp.	Verb Noun	go bananas → went bananas in a nutshell → in nutshells
Lexical	Add. Sub.	in the lurch → in the big lurch on the same page → on different pages

Table 3: Idiom variations at morphological and lexical level. Add. and Sub. refer to lexical addition and substitution respectively.

	Train	Dev			Test		
		O	M	L	O	M	L
Idiom	1247	124	21	40	124	21	40
Sent.	9752	720	200	100	1403	400	200

Table 4: Key statistics for the idioms and sentences in iSent dataset. O(Original) denotes the idioms in dev/test sets are in original forms and have appeared in training set. M(Morphology) and L(Lexical) represent the morphology and lexical idiom variations respectively and they are unseen in training set.

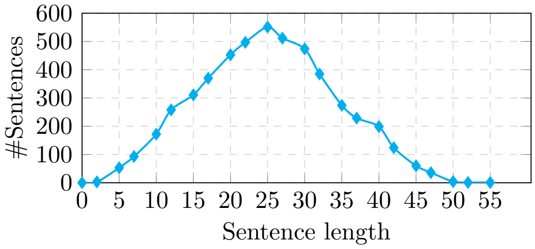

Figure 2: The distribution of the number of reviews over different lengths.

5.1 Experimental Settings

Loss Function Given a sentence and its label, the output of neural network is the probabilities of the different classes. The parameters of the network are trained to minimise the cross-entropy of the predicted and true label distributions. To minimize the objective, we use stochastic gradient descent with the diagonal variant of AdaGrad (Duchi et al., 2011).

Initialization and Hyperparameters In all of our experiments, the word embeddings for all of the models are initialized with the GloVe vectors (Pennington et al., 2014). The other parameters are initialized by randomly sampling from uniform distribution in $[-0.1, 0.1]$.

For each task, we take the hyperparameters which achieve the best performance on the development set via a small grid search over combinations of the initial learning rate $[0.1, 0.01, 0.001]$, l_2 regularization $[0.0, 5E{-}5, 1E{-}5]$ The final hyper-parameters are as follows. The initial learning rate is 0.1. The regularization weight of the parameters is $1E{-}5$.

For all the sentences from the five datasets, we parse them with constituency parser (Klein and Manning, 2003) to obtain the trees for our and some competitor models.

5.2 Competitor Models

We give some descriptions about the setting of our models and several baseline models.

- CharLSTM: Character level LSTM.
- TLSTM: Vanilla tree-based LSTM, proposed by Tai et al. (2015).
- Cont-TLSTM: Context-dependent tree-based LSTM, introduced by Bowman et al. (2016).
- iTLSTM-Lo: Proposed model with Look-Up idiomatic interpreter.
- iTLSTM-Mo: Proposed model with Morphology-Sensitive interpreter.

5.3 Evaluation over Mainstream Datasets

The experimental results are shown in Table 5. We can see Cont-TLSTM outperforms TLSTM on all four tasks, showing the importance of context-sensitive composition. Besides, both iTLSTM-Lo and iTLSTM-Mo achieve better results than TLSTM and Cont-LSTM, which indicates the effectiveness of our introduced modules (detector and idiomatic interpreter). Additionally, compared with iTLSTM-Lo, iTLSTM-Mo behaves better, suggesting its char-based idiomatic interpreter is more powerful.

Although four mainstream datasets are not rich in idioms, we could also observe substantial improvement gained from our models. We attribute this success to the power of introduced detector in identifying other non-compositional collocations besides idioms. We will discuss about this later.

5.4 Evaluation over iSent Dataset

Since iSent is a newly-introduced dataset, there is no existing baselines. Nevertheless, we provide several strong baselines implemented by ourselves as shown in Table 6, and we can observe that:

- Differing from the improvement achieved on mainstream datasets, proposed models have shown their advantages on idiom-enriched sentences. They obtain more significant improvements.

Model	MR	SST-1	SST-2	SUBJ
NBOW	77.2	42.4	80.5	91.3
RAE (Socher et al., 2011)	77.7	43.2	82.4	-
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	-
RNTN (Socher et al., 2013)	-	45.7	85.4	-
DCNN (Kalchbrenner et al., 2014)	-	48.5	86.8	-
CNN-multichannel (Kim, 2014)	81.5	47.4	88.1	93.2
CharLSTM	75.2	44.0	85.2	90.2
TLSTM	78.7	48.5	86.1	91.0
Cont-TLSTM	79.5	48.9	86.4	91.7
iTLSTM-Lo	81.6	49.9	87.7	93.2
iTLSTM-Mo	**82.5**	**51.2**	**88.2**	**94.5**

Table 5: Accuracies of our models on four datasets against state-of-the-art neural models.

Model	Train	Dev.	Test
NBOW	80.9	77.1	74.5
LSTM	87.5	76.9	75.0
BiLSTM	93.4	76.8	76.3
CharLSTM	92.4	75.1	74.4
TLSTM	88.2	75.3	74.9
Cont-TLSTM	90.8	76.2	75.5
iTLSTM-Lo	88.9	79.6	78.1
iTLSTM-Mo	91.3	81.1	80.0

Table 6: Accuracies of our models on iSent dataset against typical baselines. BiLSTM represents bidirectional LSTM.

- Additionally, iTLSTM-Lo performs worse than iTLSTM-Mo while still surpasses baseline models, which also indicates the variation-sensitive model (iTLSTM-Mo) of idioms could further improve the performance.

5.5 Analysis

In this section, we will provide more detailed quantitative and qualitative analysis in terms of three properties of idioms described in Table 1: flexibility, invisibility and idiomaticity.

Flexibility Besides the overall accuracies on the test set, we also list the performance achieved by different models over the different parts of test set: original, morphological and lexical, which represents different types of variations and have been described in Table 4.

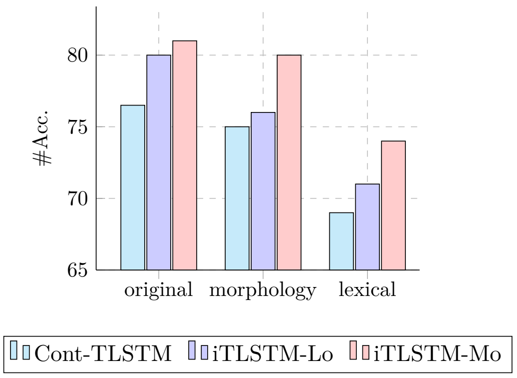

Figure 3: Performances achieved by different models are subdivided into three parts. Original, Morphology, Lexical represents accuracies achieved by corresponding part of test data.

We can see from Figure 3, both idiom-aware models achieve better performance than Cont-TLSTM by a large margin on the original part of test set, which indicates the importance of understanding idiomatic phrases during sentence modelling. Additionally, iTLSTM+Mo outperforms the other two models on the test set, suggesting the effectiveness of morphology-based model for modeling idiom variations.

Invisibility and Idiomaticity Previous experimental results have shown the effectiveness of our models. Here, we want to know how the introduced idiom detector contributes to the improvement of performance. Toward this end, we analyze all the 157 samples which our model predicts correctly while baseline model (Cont-TLSTM) fails on iSent, and find more than 120 sentences are given wrong sentiment by Cont-TLSTM due to ignoring the figurative meanings of idioms. For example, as shown in Figure 4, we randomly sample a sentence and analyze the changes of the predicted sentiment score at different nodes of the tree.

The sentence “`The movie enable my friends to blow a gasket`” has negative sentiment. Cont-TLSTM gives a wrong prediction due to ignoring the information expressed by the idiomatic phrase “`blow a gasket`”. By contrast, our model correctly detects this idiom, whose meaning plays a major role in final sentiment prediction.

Non-compositional Phrases Detection Besides idioms, we find the introduced detector can also pick up other types of non-compositional phrases[4]. We roughly sum up these non-

[4] An idiom itself is a non-compositional phrase.

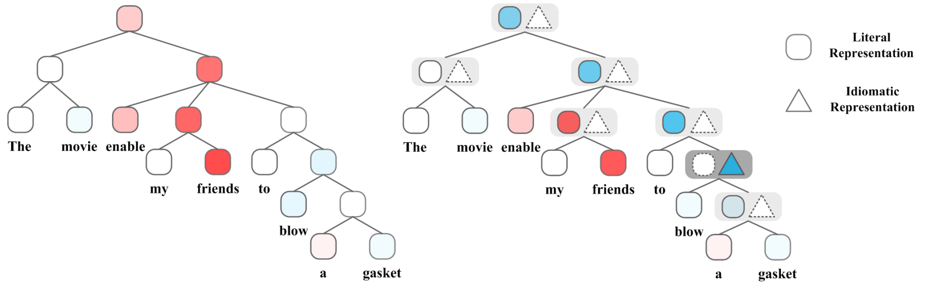

Figure 4: The change of the predicted sentiment score at different nodes of the tree. The red and blue color represent positive and negative sentiment respectively, where darker indicates higher confidence. Dashed triangle or square box denote not being selected by detector.

PN	VP	NP	AP
Notting Hill	*let alone*	*short cuts*	*on earth*
Holly Bolly	*take cover*	*deja vu*	*at once*
star wars saga	*rips off*	*black comedy*	*all in all*
Barry Skolnick	*thumb down*	*femme fatale*	*not only*
Apollo 13	*fall apart*	*no problem*	*at times*

Table 7: PN, VP, NP and AP represent proper noun, noun phrase, verb phrase and adverbial phrase respectively.

compositional phrases picked up by introduced detector from all the five development sets and list them in Table 7.

From the table, we can see that most of these phrases either imply an affective stance toward something: "`thumbs down`", or are critical to the understanding of sentences such as the "Verb Phrases" and "Adverb Phrases". For example, the sentence "`More often than not, this mixed bag hit its mark`" has a positive sentiment. Cont-TLSTM pays much more attention to the word "`not`" without realizing that it belongs to the collocation "`more often than not`", which expresses neutral emotion. In comparison, our model regards this collocation as a whole with neutral sentiment, which is crucial for the final prediction.

6 Related Work

Previous work related to idioms focused on their identification, which falls in two kinds of paradigms: idiom type classification (Gedigian et al., 2006; Shutova et al., 2010) and idiom token classification (Katz and Giesbrecht, 2006; Li and Sporleder, 2009; Fazly et al., 2009; Peng et al., 2014; Salton et al., 2016). Different with these work, we integrate idioms understanding into a real-world task and consider different peculiarities of idioms in an end-to-end trainable framework.

Recently, there are some work exploring the compositionality of various types of phrases (Kartsaklis et al., 2012; Muraoka et al., 2014; Hermann, 2014; Hashimoto and Tsuruoka, 2016). Compared with these work, we focus on how to properly model idioms under the context of sentence representations.

More recently, Zhu et al. (2016) propose a DAG-structured LSTM to incorporate external semantics including non-compositional or holistically learned semantics. Its key characteristic is that a DAG needs be built in advance, which merges some detected n-grams as the non-compositional phrases based on external knowledge. Different from this work, we focus on how to integrate detection and understanding of idioms into a unified end-to-end model, in which an idiomatic detector is introduced to adaptively control the semantic compositionality. Particularly, in the whole process no extra information is given to tell which phrases should be regarded as non-compositional.

7 Conclusion and Future Work

In this paper, we lay idioms understanding in the context of sentence-level semantic representation based on two linguistic perspectives. To apply our model into the real-world task, we introduce a sizeable idiom-enriched sentiment classification dataset, which covers abundant peculiarities of idioms. We make an elaborate experiment design and case analysis to evaluate the effectiveness of our proposed models.

In future work, we would like to investigate more complicated idiom-enriched NLP tasks, such as machine translation.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011 and 61672162), Shanghai Municipal Science and Technology Commission (No. 16JC1420401), and Young Elite Scientists Sponsorship Program By CAST (2015QNRC001).

References

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202* .

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & Cognition* 1(3):343–346.

Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021* .

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147* .

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The JMLR* 12:2493–2537.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The JMLR* 12:2121–2159.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1):61–103.

Linda Flavell and Roger Flavell. 2006. *Dictionary of idioms and their origins*. Kyle Cathie Limited.

Bruce Fraser. 1970. Idioms within a transformational grammar. *Foundations of language* pages 22–42.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*. Association for Computational Linguistics, pages 41–48.

Sam Glucksberg. 1993. Idiom meanings and allusional content. *Idioms: Processing, structure, and interpretation* pages 3–26.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. *arXiv preprint arXiv:1603.06067* .

Karl Moritz Hermann. 2014. Distributed representations for compositional semantics. *arXiv preprint arXiv:1411.3146* .

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

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of ACL*.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *In Proceedings of COLING: Posters*. Citeseer.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, pages 12–19.

Jerrold J Katz. 1963. *Semantic interpretation of idioms and sentences containing them*. Research Laboratory of Electronics, Massachusetts Institute of Technology.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. pages 423–430.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* .

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pages 315–323.

Pengfe Liu, Xipeng Qiu, Jifan Chen, and Xuanjing Huang. 2016a. Deep fusion LSTMs for text semantic matching. In *Proceedings of ACL*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* .

Pengfei Liu, Xipeng Qiu, Yaqian Zhou, Jifan Chen, and Xuanjing Huang. 2016b. Modelling interaction of sentence pair with coupled-lstms. In *Proceedings of the 2016 Conference on EMNLP*.

Masayasu Muraoka, Sonse Shimaoka, Kazeto Yamamoto, Yotaro Watanabe, Naoaki Okazaki, and Kentaro Inui. 2014. Finding the best model among representative compositional models. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation (PACLIC 2014)*. pages 65–74.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language* pages 491–538.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 115–124.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *EMNLP*. pages 2019–2027.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the EMNLP* 12:1532–1543.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese .

Giancarlo D Salton, Robert J Ross, and John D Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 194–204.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 1002–1010.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*. pages 1201–1211.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* .

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications* 42(21):7375–7385.

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-adaptive hierarchical sentence model. *arXiv preprint arXiv:1504.05070* .

Xiao-Dan Zhu, Parinaz Sobhani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In *ICML*. pages 1604–1612.

Xiaodan Zhu, Parinaz Sobhani, and Hongyu Guo. 2016. Dag-structured long short-term memory for semantic compositionality. In *Proceedings of NAACL-HLT*. pages 917–926.