Neurocomputing 428 (2021) 268–279

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Metaphor identification: A contextual inconsistency based neural sequence labeling approach

Xin Chen [a], Zhen Hai [b], Suge Wang [a,c,*], Deyu Li [a,c], Chao Wang [d], Huanbo Luan [e]

[a] School of Computer and Information Technology, Shanxi University, China
[b] R&D Center Singapore Machine Intelligence Technology Alibaba DAMO Academy, Singapore
[c] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, China
[d] 6ESTATES PTE LTD, Singapore
[e] Tsinghua University, China

ARTICLE INFO

*Article history:*
Received 8 March 2020
Revised 20 September 2020
Accepted 5 December 2020
Available online 16 December 2020
Communicated by Steven Hoi

*Keywords:*
Sequence labeling model
Neural networks
Metaphor identification
Contextual inconsistency

ABSTRACT

Metaphor identification helps improve the performance of various natural language understanding tasks such as word sense disambiguation and sentiment analysis. Though many efforts have been made to deal with metaphor identification, most existing studies largely overlook a fact of contextual inconsistency of the metaphors in natural language. We observe that the greater the semantic inconsistency between current word and contextual words is, the more likely the word belongs to the metaphorical category. In this paper, we formulate the metaphor identification as a sequential tagging problem, and then develop a novel contextual inconsistency based neural sequence labeling approach, which can leverage the semantic contextual inconsistency among words of a sentence to address the problem. We propose to rely on distance metric to measure the contextual inconsistency, and evaluate four widely used distance functions in experiments. Experimental results on publicly available datasets validate the benefit of the proposed model over state-of-the-art baselines for metaphor identification.

## 1. Introduction

Metaphors are ubiquitous in language, on average in every third sentence of general-domain text [1]. A metaphor makes a sentence more vivid and poetic, as well as more obscure in linguistic research [2]. From the perspective of cognition, the essence of a metaphor is the understanding and experiencing of one kind of thing in terms of another [3]. A metaphor is not only a property of language, but also a tool for cognitive activities of humans that help to construct conceptual systems [1]. To alleviate the lack of appropriate words to express new concepts in a certain period of time, humans create new metaphorical meanings of words using active imagination [2]. Some metaphorical and unmetaphorical sentences are shown as follows:

*Example 1: The painting* **won** *critical acclaim.*

*Example 2: This army won a battle.*

*Example 3: My sister's* **memory** *is a* **camera** *that remembers everything we see.*

*Example 1* shows a metaphorical sentence, where *painting* belongs to the *culture and art* category, which is literally different from *won* from the *battle* category. Clearly, it would make the description of *painting* more powerful by using the word *won*. In contrast, in the unmetaphorical sentence of *Example 2*, both words *army* and *won* literally come from the same category *battle*. We observe that the literal difference between the *source* word or category (*battle*) and *target* word or category (*culture and art*) forms the semantic contextual inconsistency, and thus may result in the metaphor in the sentence. The essential feature of a metaphor, i.e., contextual inconsistency, is also observed in the metaphorical sentence of *Example 3*, as shown by the words *memory* and *camera*. Generally, the rich imagination and creativity of human minds may lead to the understanding of senses of natural language beyond the narrowly conceived conception [4]. Thus, it is often difficult to figure out the polysemous meanings and metaphorical uses of words in different contexts of language.

Metaphor identification is an imperative technology for semantic understanding tasks in natural language processing, such as machine translation, information retrieval, and sentiment analysis.

* Corresponding author at: School of Computer and Information Technology, Shanxi University, China.
*E-mail address:* wsg@sxu.edu.cn (S. Wang).

https://doi.org/10.1016/j.neucom.2020.12.010
0925-2312/© 2020 Elsevier B.V. All rights reserved.

Metaphorical meanings account for 20% of word interpretation tasks, which pose a great challenge to word sense disambiguation. Due to the obscurity of metaphor, 44% of metaphorical expressions are translated incorrectly in Google Translate [5]. The poetic nature and imagery of metaphorical expressions cause that they tend to express more implicit emotions than literal expressions [6], which is an obstacle for sentiment analysis.

By exploiting a range of metaphor properties, various efforts have been made to cope with metaphor identification, for example, selectional preference [7,8], concreteness and imageability [9–11], and conceptual information [5,12,13]. Metaphor identification has been formulated as a classification problem in previous studies. Traditional methods focus on carefully designing various task-specific features by investigating the linguistic properties of metaphors [4,14–18]. Recently, various kinds of neural network models have been proposed [19–27], and have achieved good performances. To our knowledge, the semantic contextual inconsistency, a strong indicator of metaphors, has not been studied in previous work.

In this work, following previous studies, we formulate the metaphor identification as a sequential tagging problem. To deal with the problem, we propose to learn from the semantic contextual inconsistency among words, in addition to sequential context semantics of given sentences. Generally, a metaphorical word is used to modify related words in an imaginative way, and is often literally inconsistent from the context in a sentence. We observe that the greater the semantic inconsistency between current word and contextual words is (e.g., from two different domains), the more likely the word would be categorized as metaphor. In order to measure contextual inconsistency, we represent certain semantic meaning of each word within a sentence by using an *abstract distributed representation*, and we propose to employ the distributional distance between the abstract distributed representations of each pair of words in the sentence. Then, we develop a novel neural sequence labeling model, named SEQ-CI, which can exploit the contextual inconsistency property of natural language to improve metaphor identification. One key benefit of SEQ-CI is that it exploits the contextual inconsistency as regularization, and may allow better learning of model parameters from real-life natural language data.

We have made the following main contributions in this paper:

- We find that the contextual inconsistency is an essential feature of a metaphor. To the best our knowledge, this is the first work that leverages semantic contextual inconsistency to help identify metaphor.
- A new enhanced sequence labeling approach to metaphor identification is presented, which can learn from the contextual relationships between pair-wise words in a sentence.
- We evaluate the proposed model on publicly available data, and the results show the superiority of our model over the state-of-the-art baselines.

The rest of this paper is organized as follows. Section 2 presents the related work to metaphor identification. Section 3 describes the proposed contextual inconsistency based sequence labeling approach. Section 4 illustrates the experimental results and analysis. In Section 5, we present discussions on several main aspects of the proposed model, and conclude the paper in Section 6.

## 2. Related work

A variety of work has been done to cope with metaphor identification, and can be roughly grouped into following categories, i.e., selectional preference, concreteness and imageability, conceptual information, and classification formulation.

### 2.1. Selectional preference

Generally, metaphorical words tend to be less frequent than normal use of literal expressions in natural language. Based on the observation, Haagsma et al. [7] utilized selectional preference violation as a tool for metaphor detection. Lederer [8] measured the relative frequency of a particular lexical item in the specialized corpus. The effectiveness of using the selectional preference for metaphor identification was shown in previous studies, but it may fail to detect the most conventional metaphors that are frequent.

### 2.2. Concreteness and imageability

A metaphor is commonly used to describe intuitive concepts in terms of more concrete or physical experiences. The concreteness and imageability have been recognized as indicators of the metaphorical use of words in sentences. Turney et al. [9] introduced a measure for concreteness of concepts based on the MRC Pyscholinguistic Database [28], in order to identify metaphorical use. Following the study [9], Bulat et al. [10] constructed a cognitive representation of concepts, while Köper et al. [11] investigated norms for multi-word units (e.g., verb-noun pairs). These studies led to better metaphor detection. Based on the experiments, concreteness and imageability are shown to be helpful for metaphor identification. However, approaches rely largely on the use of human-crafted knowledge resources, for example, MRC Pyscholinguistic Database [28] and the property norm dataset [29], which may limit the applications across different domains in reality.

### 2.3. Conceptual information

A metaphor often involves two different but related concepts or conceptual domains, i.e., target and source concepts. To automatically extract linguistic metaphors, Heintz et al. [12] used latent Dirichlet allocation (LDA) as a proxy for detection of the source or target concept. Shutova et al. [13] obtained conceptual domains by using clustering. However, the methods were limited to seed selection. Then, Shutova et al. [5] used a hierarchical clustering method to find the relationship between the source and target concepts, which can get rid of the dependence on the seed mapping expressions.

### 2.4. Classification formulation

Metaphor identification has been formulated as a classification problem in previous studies. Traditional methods focus on carefully designing various task-specific features by investigating a range of properties of metaphors, including grammar-based features [14,15], resource-based features [4,15,16] and corpus-based features [15–18]. Based on the designed features, they then rely on various well-known algorithms, such as logistic regression, support vector machines, and conditional random fields, etc., to classify each given expression as the literal meaning or metaphorical meaning. By treating the metaphor detection as a changing process, Schlechtweg et al. [30] detected metaphoric change via word entropy. Rai et al. [31] formulated a soft metaphor identification problem, and employed a fuzzy c-means method to identify a given expression as literal meaning, metaphorical meaning, or possibly metaphorical meaning.

Recently, various kinds of neural network models have been proposed to deal with metaphor identification [19–27].

Specifically, Rei et al. [23] designed a supervised similarity network to identify metaphors by using the similarity between pairs of words. Wu et al. [26] proposed a combined CNN-LSTM model to capture local frequent patterns, while Gao et al. [22] developed contextualized RNN_ELMo models to exploit sequential context of sentences for metaphor identification. Inspired by the theories of the linguistic metaphor identification procedure (MIP) and selectional preference violation (SPV), Mao et al. [27] presented end-to-end sequential recurrent neural networks that can leverage the theories to identify metaphors.

Though existing deep neural network models may work well for metaphor identification, they largely overlook the fact of contextual inconsistency of the metaphor words from all other literal words in natural language. In this paper, we extend a neural sequence labeling model, and propose to incorporate the contextual inconsistency among words as regularization in a new learning framework for improvement of metaphor identification.

## 3. Methodology

### 3.1. Overview

We aim to deal with the metaphor identification from natural language, and following previous studies [22,26,27], we formulate it as a sequential classification problem. Formally, let $C$ be a labeled collection of sentences, $C = \{(S_1, Y_1), \cdots, (S_l, Y_l), \cdots, (S_L, Y_L)\}$, where $S_l$ is an input sentence with $N_l$ words, $S_l = \{w_1, w_2, \ldots, w_{N_l}\}$, while $Y_l = \left\{y_1, y_2, \ldots, y_{N_l}\right\}$ denotes the labels of respective words in $S_l$. Note that $y_i \in \{0, 1\}$, and $y_i = 1$ means the word $w_i$ is labeled as metaphor $(i = 1, 2, \ldots, N_l)$. The metaphor identification task is to build a sequence labeling model based on the labeled data $C$, and then employs the model to predict a sequence of binary labels for individual words of a given unknown sentence.

We observe that, if a sentence contains metaphors, the metaphorical words are often literally and semantically inconsistent from rest contextual words in the sentence. Specifically, the greater the semantic inconsistency between current word (target) and contextual words is, the more likely the current word is a metaphorical expression. Following previous work [32], we represent certain semantic meaning of each word within a sentence by using an *abstract distributed representation*. Then, in order to measure the contextual inconsistency between each pair of words of a given sentence, we propose to employ the distributional distance between the abstract distributed representations of each pair of words in the sentence. Therefore, to deal with the metaphor identification problem, we propose to leverage the semantic contextual inconsistency (CI) in natural language, and develop a new sequence labeling model, called SEQ-CI, as shown in Fig. 1. Our new SEQ-CI consists of five main components, i.e., the embedding layer, the contextualized word representation layer, the abstract distributed representation layer, the semantic contextual inconsistency module, and the metaphor prediction layer, which would be presented in detail in subsequent sections.

### 3.2. Word embedding and contextualized representation

Generally, the real meanings of metaphorical (and unmetaphorical) words are determined not only by their original meanings but also by the context of words. We propose to adopt two types of word embeddings, i.e., *static* and *contextual* semantic representations, as the initialized combined representations of each input sequence of words. Next, the well-known Bi-LSTM RNN model [33] is employed to learn the contextualized word representations.

Formally, given an input sentence with $N$ words $S = \{w_1, w_2, \ldots, w_N\}$, (For simplicity, we omit the subscript of input sentence), we propose to use the combined representation $x_i$ for each word $w_i$ $(i = 1, 2, \ldots, N)$ in the input embedding layer, as shown in Eq. (1).

$$x_i = [g_i || e_i] \tag{1}$$

where $g_i$ denotes the *static* embedding of English word from the Glove model [32] or Chinese word from Tencent AI Lab Embeddings [34], respectively, $e_i$ denotes the contextual embedding from the ELMo model [35], and the symbol "||" means concatenation.

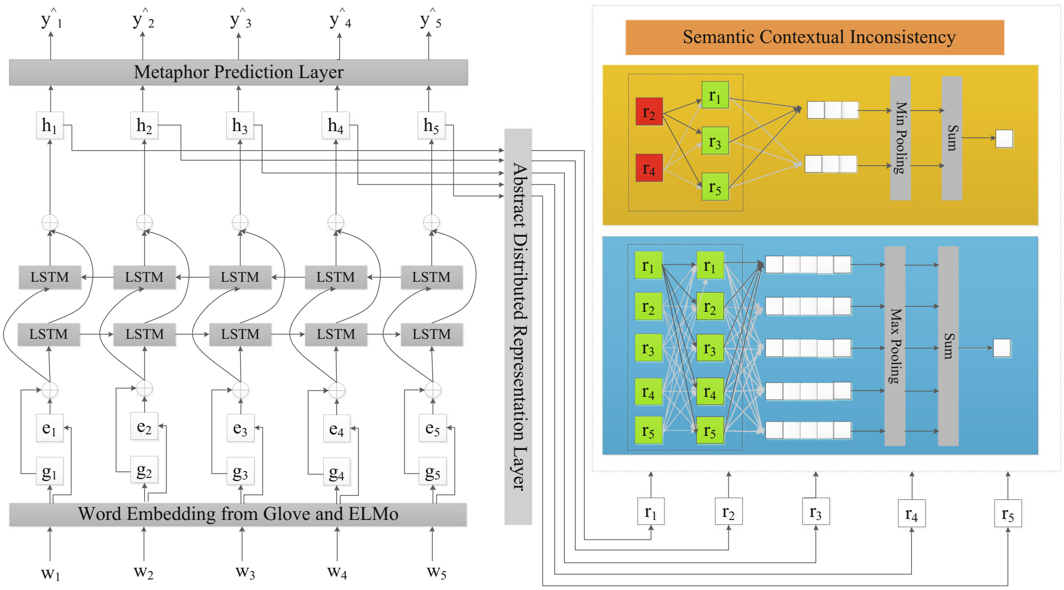


**Fig. 1.** The architecture of the proposed metaphor sequence labeling model with semantic contextual inconsistency (SEQ-CI), which consists of five main components, i.e., the embedding layer, the contextualized word representation layer, the abstract distributed representation layer, the semantic contextual inconsistency module, and the metaphor prediction layer. The contextual inconsistency module differentiates two cases, i.e., sentences with (*upper, orange*) and without (*lower, blue*) metaphorical expressions. For the sentences with metaphors, the red and green colors denote the distributed representations of metaphorical and unmetaphorical words, respectively.

Then, taking as input the combined representations of words $x_i$ $(i = 1, 2, \ldots, N)$, we use the Bi-LSTM model as sequence encoder to generate the contextualized word representations. In particular, the LSTM unit is composed of three gate structures, where the input gate controls which information enters into the unit, the forget gate determines which information is dropped from the unit, and the output gate decides which information is produced from the unit. In addition, the cell state records all historical information which flows up to the current time. Thus, the contextualized representation of $w_i$ based on preceding text is computed by a forward LSTM unit via Eqs. (2)–(7):

$$z_i = \sigma\left(W_z x_i + U_z \overrightarrow{h}_{i-1} + b_z\right) \tag{2}$$

$$o_i = \sigma\left(W_o x_i + U_o \overrightarrow{h}_{i-1} + b_o\right) \tag{3}$$

$$f_i = \sigma\left(W_f x_i + U_f \overrightarrow{h}_{i-1} + b_f\right) \tag{4}$$

$$\tilde{c}_i = tanh\left(W_c x_i + U_c \overrightarrow{h}_{i-1} + b_c\right) \tag{5}$$

$$c_i = f_i \odot c_{i-1} + z_i \odot \tilde{c}_i \tag{6}$$

$$\overrightarrow{h}_i = o_i \odot tanh(c_i) \tag{7}$$

where $z_i, o_i, f_i$ denote the input gate, the output gate, and the forget gate, respectively, $\tilde{c}_i$ refers to new memory content, $c_i$ is the contextual memory content, $\overrightarrow{h}_i$ stands for the final output of forward LSTM unit. $W_z, W_o, W_f, W_c$ are weight parameters for current input $x_i$; $U_z, U_o, U_f, U_c$ represent weights parameters for previous hidden state $\overrightarrow{h}_{i-1}$; $b_z, b_o, b_f, b_c$ denote bias for the input gate, the output gate, the forget gate and the cell hidden state, respectively. $\sigma$ and *tanh* denote the sigmoid and hyperbolic tangent activation functions.

Eq. (8) summarizes the computational steps for each forward LSTM:

$$\overrightarrow{h}_i = \overrightarrow{LSTM}\left(x_i, \overrightarrow{h}_{i-1}, \overrightarrow{\theta}_i\right) \tag{8}$$

where $\overrightarrow{\theta}_i$ are all the parameters of the forward LSTM unit.

Similarly, we also calculate the contextual representations based on subsequent content of each input sequence via a backward LSTM, as shown in Eq. (9).

$$\overleftarrow{h}_i = \overleftarrow{LSTM}\left(x_i, \overleftarrow{h}_{i+1}, \overleftarrow{\theta}_i\right) \tag{9}$$

where $\overleftarrow{\theta}_i$ are all the parameters of the backward LSTM unit.

The forward LSTM layer produces a sequence of forward hidden states $\left\{\overrightarrow{h}_1, \overrightarrow{h}_2, \cdots, \overrightarrow{h}_N\right\}$, while the backward LSTM yields another sequence of hidden states $\left\{\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_N\right\}$. The two sequences of hidden states are concatenated together to form the contextualized representations $H$, as shown in Eq. (10).

$$H = \{h_1, \cdots, h_N\}, \tag{10}$$

where $h_i = \left[\overrightarrow{h}_i || \overleftarrow{h}_i\right]$ $(i = 1, 2, \ldots, N)$, and $h_i \in \mathbb{R}^{d_1}$.

The contextualized representations $H$ is then fed to the subsequent metaphorical layer as input for the sequential metaphor classification task.

### 3.3. Abstract distributed representation

To characterize the distributional semantics of words in a given sentence, we design an abstract distributed representation layer. Specifically, we propose to adopt a *softmax* function, which takes the contextualized representations $H$ as input, and generates corresponding distributed representations of individual words.

Let $h_i$ be the contextualized representation of $w_i$, and the sequence of abstract distributed representations of the sentence: $R = \{r_1, r_2, \ldots, r_N\}$ can be obtained via Eqs. (11)–(12).

$$m_i = W_r h_i + b_r \tag{11}$$

$$r_i = softmax(m_i) \tag{12}$$

where

$$r_{i,j} = \frac{e^{m_{i,j}}}{\sum_{k=1}^{d_2} e^{m_{i,k}}} \tag{13}$$

$W_r$ and $b_r$ are trainable parameters, $W_r \in \mathbb{R}^{d_2 \times d_1}$, and $b_r \in \mathbb{R}^{d_2}$.

### 3.4. Semantic Contextual Inconsistency

The semantic contextual inconsistency among words within each input sentence is significant for metaphor identification. To measure the contextual inconsistency, we take into account two contrastive cases, i.e., sentences with metaphors and sentences without metaphors.

#### 3.4.1. Sentences with metaphors

For the sentences with metaphors, there naturally contains contextual inconsistency between metaphorical words and contextual unmetaphorical words. We observe that the greater the contextual inconsistency between current metaphorical word and context is, the greater possibility that the metaphorical word would be correctly predicted. We propose to rely on overall distributional distances between metaphor words and contextual words of an input sentence to calculate the contextual inconsistency of the sentence. In particular, we compute the sum of minimum distributional distances of metaphorical words against all the rest contextual unmetaphorical words. Given a sentence containing metaphors, it is expected that the overall contextual inconsistency of the sentence is relatively high. With the objective of maximizing the overall inconsistency of the sentence, selecting to use the minimum distributed distances may ensure that the local inconsistency between every pair of words can be properly modeled. The inconsistency $Inc(S)$ of sentence $S$ is computed via Eq. (14).

$$Inc(S) = \sum_{k=1}^{K} \min_{1 \leqslant u \leqslant U} \left\{Dis\left(r_k^{(met)}, r_u^{(un)}\right)\right\} \tag{14}$$

where both $r_k^{(met)}$ and $r_u^{(un)}$ are the abstract distributed representations of the metaphorical word $w_k$ and unmetaphorical word $w_u$ of the sentence $S$ $(k = 1, 2, \ldots, K; u = 1, 2, \ldots, U; K/U$: number of metaphorical/unmetaphrical words in $S$). Note that during training the ground truth labels of words of each input sentence are known. $Dis\left(r_k^{(met)}, r_u^{(un)}\right)$ denotes a distance function of $r_k^{(met)}$ and $r_u^{(un)}$. A variety of distance metrics are available in practice, and we evaluate four well-known distance functions in the experiments, i.e., modified cosine distance, Euclidean distance, Gaussian distance, and Kullback-Leibler divergence, as shown below.

- Modified cosine distance:

$$Dis(r_k, r_u) = -\frac{\sum_{m=1}^{d_2} r_{k_m} r_{u_m}}{\sqrt{\sum_{m=1}^{d_2} (r_{k_m})^2}\sqrt{\sum_{m=1}^{d_2} (r_{u_m})^2}} \tag{15}$$

- Euclidean distance:

$$Dis(r_k, r_u) = \sqrt{\sum_{m=1}^{d_2} (r_{k_m} - r_{u_m})^2} \tag{16}$$

- Gaussian distance:

$$Dis(r_k, r_u) = exp\left( -\frac{\sqrt{\sum_{m=1}^{d_2} (r_{k_m} - r_{u_m})^2}}{2\sigma^2} \right) \tag{17}$$

- Kullback-Leibler divergence:

$$Dis(r_k, r_u) = \sum_{m=1}^{d_2} r_{k_m} * log\left(\frac{r_{k_m}}{r_{u_m}}\right) \tag{18}$$

where $r_k, r_u \in \mathbb{R}^{d_2}$.

### 3.4.2. Sentences without Metaphors

For the sentences that do not contain metaphors, it is expected that their contextual inconsistency should be as small as possible. With the aim of minimizing the overall contextual inconsistency $Inc(S)$ of the sentence $S$ without metaphors, we propose to measure the inconsistency $Inc(S)$ by using the sum of maximum distributional distance of each pair of words in the sentence, as shown in Eq. (19) below.

$$Inc(S) = \sum_{k=1}^{N} \max_{1 \leqslant u \leqslant N, u \neq k} \{Dis(r_k, r_u)\} \tag{19}$$

where both $r_k$ and $r_u$ are the abstract distributed representations of each pair of words $w_k$ and $w_u$ of the given sentence $(k, u = 1, 2, \ldots, N)$.

## 3.5. Metaphor word identification

In the metaphor identification layer, a softmax classifier takes the contextualized representation of each word of an input sentence as inputs, and then predicts whether each word is a metaphor or not. Formally, let $S = \{w_1, w_2, \ldots, w_N\}$ be an input sequence of words and $H = \{h_1, h_2, \cdots, h_N\}$ be the sequence of contextualized representations. Then, the predicted sequence of probability scores corresponding to respective words can be computed via Eqs. (20)–(21).

$$q_i = W_t h_i + b_t \tag{20}$$

$$\hat{y}_i = softmax(q_i) \tag{21}$$

where

$$\hat{y}_{i,j} = \frac{e^{q_{i,j}}}{\sum_{k=1}^{|Labels|} e^{q_{i,k}}} \tag{22}$$

$h_i$ denotes the contextualized representation of word $w_i$ of the sentence, $|Labels|$ is the number of predictive labels, both $W_t$ and $b_t$ are trainable task-specific parameters, $W_t \in \mathbb{R}^{d_3 \times d_1}$, and $b_t \in \mathbb{R}^{d_3}$.

## 3.6. Model training

Previous study has shown the benefit of focal loss over other loss functions (e.g., cross-entropy) for learning from imbalanced data (see Table 1) [36]. Based on the focal loss, we introduce an inconsistency regularization, and then develop a new loss function, which allows the propose model SEQ-CI to leverage the semantic contextual inconsistency for learning from the training data, as shown in Eq. (23).

$$Loss = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{i=1}^{N_l} -\alpha_{y_i}(1-\hat{y}_i)^{\gamma} y_i log(\hat{y}_i) - \lambda R(S_l)\right) \tag{23}$$

where

$$R(S_l) = \begin{cases} Inc(S_l) & S_l \text{ with metaphors} \\ -Inc(S_l) & S_l \text{ without metaphors} \end{cases} \tag{24}$$

Both $y_i$ and $\hat{y}_i$ are ground-truth and predicted labels of word $w_i$ of the sentence $S_l$, $N_l$ is number of words in the sentence, and $L$ is the number of sentences in the dataset. The hyper-parameter $\alpha_{y_i}$ denotes the weight of label $y_i$, setting $\gamma > 0$reduces the relative loss for well-classified examples, and $\lambda$ denotes the importance of the contextual inconsistency term in the loss.

# 4. Experiments

## 4.1. Data

To evaluate the propose model SEQ-CI for metaphor identification, we adopted three commonly used English datasets, i.e., VUA [37], TroFi [38], and MOH-X [20], and one publicly available Chinese metaphor dataset called CHI.[1] Table 1 lists the statistics of the datasets.

**VUA**: This is the largest dataset publicly available for metaphor detection, and each word of the dataset is annotated. VUA covers various types of text, including fiction, academic writing, newspaper and conversation. Following [22,27], we split the dataset VUA (including words of all POS tags) into three subsets, i.e., VUA_trn (training), VUA_tst (testing), and VUA_dev (development), and also applied the model trained on VUA to the randomly sampled VUA verb testing set VUA_Verbs_tst, which is actually one shared task of metaphor identification in NAACL 2018 [39].

**TroFi**: This metaphorical corpus was constructed by using bootstrapping method, and only a single target verb was annotated in every sentence. Following [22,27], we conducted the 10-CV (cross-validation) experiments on this dataset.

**MOH-X**: MOH extracts conceptual words with multiple ($3 \sim 10$) meanings from WordNet, and annotates the examples to form a metaphorical corpus. MOH-X is a metaphorical subset of MOH, where each sentence has a single annotated target verb. Similar to **TroFi**, we conducted the 10-CV experiments on the dataset.

**CHI**: The metaphor dataset was created for Chinese metaphor identification in the 17th China national conference on computational linguistics (CCL 2018).[2] The dataset contains 5,494 sentences, where the verbs and nouns of each sentence were annotated. It was split into three subsets of CHI_trn (training), CHI_dev (development) and CHI_tst (testing), as shown in Table 1.

## 4.2. Comparison systems

For the proposed SEQ-CI approach to metaphor identification, various distance metrics can be used to compute the contextual

[1] http://ir.dlut.edu.cn/File/Download?cid=3.

[2] http://www.cips-cl.org/static/CCL2018/call-evaluation.html.

**Table 1**
The statistics of datasets [27]. # Tgt tok is the number of target tokens whose metaphoricity is to be identified. % M is the percentage of metaphoric tokens among target tokens. # Seq is the number of sequences. Avg # seq len is the average of the number of sequence lengths. Avg # M/S is the average number of metaphors per metaphorical sentence. #Met Seq and #Non Met Seq denote the number of the sequences with metaphors and without metaphors, respectively.

| DataSet | #Tgt tok | %M | #Seq | Avg # Seq len | Avg # M/S | #Met Seq | #Non Met Seq |
|---|---|---|---|---|---|---|---|
| VUA | 205,425 | 11.6 | 10,567 | 19.4 | 3.4 | 6,897 | 3,670 |
| VUA_trn | 116,622 | 11.2 | 6,323 | 18.4 | 3.3 | 3,974 | 2,349 |
| VUA_dev | 38,628 | 11.6 | 1,550 | 24.9 | 4.0 | 1,116 | 434 |
| VUA_tst | 50,175 | 12.4 | 2,694 | 18.6 | 3.4 | 1,807 | 887 |
| VUA_Verbs_tst | 5,873 | 30.0 | 5,873 | 18.6 | 1.5 | 1,759 | 4,114 |
| CHI_trn | 47,069 | 22.8 | 3,954 | 11.9 | 2.9 | 3,674 | 280 |
| CHI_dev | 6,371 | 15.9 | 440 | 14.5 | 2.7 | 370 | 70 |
| CHI_tst | 13,707 | 25.4 | 1,100 | 12.4 | 3.2 | 1,088 | 12 |
| TroFi | 3,737 | 43.5 | 3,737 | 28.3 | 1.0 | 1,627 | 2,110 |
| MOH-X | 647 | 48.7 | 647 | 8.0 | 1.0 | 315 | 332 |

inconsistency. In our experiments, we evaluated four different metrics, i.e., the modified cosine distance, Euclidean distance, Gaussian distance, and Kullback-Leibler divergence, and thus implemented four variants of the proposed model, i.e., **SEQ-CI + Cos**, **SEQ-CI + Edis**, **SEQ-CI + Gaussian**, and **SEQ-CI + KL**, respectively. We compared the proposed models with the following well-established baselines.

- Semantic-Feature: This method presents a variety of semantic features, such as difference of word concreteness of verb-noun or adjective-noun, word stem features, and then relies on a logistic regression model built on these features for metaphor identification [4].
- Abstractness: This method exploits the abstractness degree of words for metaphor detection. It presents the abstractness features, such as single word and multiword units, and then constructs a logistic regression classifier for identifying metaphors [11].
- Word-Similarity: The metaphorical use of words often involves two different but related *target* and *source* concepts. A supervised similarity network leverages the similarity between pairs of words for metaphor detection [23].
- CNN+RNN$_{\text{ensmb}}$: The combination of CNN and RNN models can capture both local and long dependency information of sentence simultaneously. A combined CNN-LSTM model [26] was designed to identify metaphors. The model has achieved the best performance in shared tasks on the VUA dataset at the NAACL 2018 workshop for metaphorical language processing [39].
- RNN_ELMo$_{\text{SEQ}}$: The contextual information is often helpful for understanding the meaning of metaphorical words. Based on LSTM, the contextualized neural sequence labeling model [22] was proposed to identify metaphors from natural language.
- RNN_HG: The metaphor identification procedure (MIP) shows that a metaphor can be recognized by the contrast between the contextual and literal meanings of each target word. A contrastive recurrent neural network, which could leverage the comparison between hidden states and original Glove embeddings [32] of individual words, was presented to identify metaphors [27].
- RNN_MHCA: As shown in the selectional preference violation, metaphoricity can be identified by detecting the incongruity between a target word and its context in a given sentence. A recurrent neural network which can exploit multi-head contextual attention was developed for metaphor detection problem [27].

### 4.3. Experimental setting

Under the unified learning framework of the proposed SEQ-CI, we carefully tuned the values of the dimensionality of BiLSTM states and the abstract semantic representation (ASR) via grid search, as shown in the Table 2. We applied the Adam optimizer [40] to learning of the SEQ-CI model, and the batch size was set as 32 for VUA and CHI, and 10 for TroFi and MOH-X. Following [26,27], we set the value of hyper-parameter $\alpha_{y_i}$ as 2.0 for the label $y_i = 1$ and 1.0 for $y_i = 0$, respectively. Following [36], the hyper-parameter $\gamma$ was set as 2.0. We tuned the hyper-parameter $\lambda$ via grid search, and selected the value as 1.0. The contextual ELMo embeddings are not available on the Chinese dataset CHI, and thus were not used in the input embedding layer of our models.

In the follow sections, we report the metaphor detection results of the proposed models and baselines in terms of precision (P), recall (R), and F1 score (F1) metrics ("-": not available).

### 4.4. The results of metaphor identification for all POS track

We compare the proposed models with the state-of-the-art baselines not only on English data (VUA) but also on Chinese data (CHI). Table 3 shows the results of metaphor detection for words of all part-of-speech (POS) tags.

In particular, the baseline CNN+RNN$_{\text{ensmb}}$ loses the competition with all the SEQ-CI models and other sequential neural models for metaphor detection. Generally, the CNN module in CNN+RNN$_{\text{ensmb}}$ is useful for acquiring local information within a fixed window, and may not be able to capture properly the sequential semantic dependency among words in sentences. Though the baselines, i.e., RNN_ELMo$_{\text{SEQ}}$, RNN_HG, and RNN_MHCA, perform well for metaphor identification, overall the proposed models, i.e., SEQ-CI + Cos and SEQ-CI + Gaussian, achieve the best performance in terms of F1 across VUA and CHI datasets. This may suggest that learning from contextual inconsistency among words in addition to from contextualized semantics is really beneficial for identifying metaphorical words from input sequences. In addition, the performance gap (F1) between the proposed models and baselines on CHI (69.5 vs 67.9) is larger than that on VUA (74.8 vs 74.3). It may suggest that the proposed contextual inconsistency based sequential labeling model can be also generalized to different languages (Chinese) for metaphor detection.

The VUA dataset consists of four different types of articles, i.e., academic, conversation, fiction, and news, where the metaphor tokens account for 11.2%, 7.4%, 10.7%, 14.9% of total tokens of the respective types of articles. To further validate the advantage of

**Table 2**
The selected values of dimensionality of BiLSTM states and abstract semantic representation on different dataset.

| Type | Dateset | Measurements of the SEQ-CI Model | | | |
|---|---|---|---|---|---|
| | | Cos | Edis | Gaussian | KL |
| BiLSTM | VUA | 64 | 64 | 256 | 64 |
| | TroFi | 32 | 32 | 32 | 32 |
| | MOH-X | 256 | 128 | 128 | 256 |
| | CHI | 32 | 32 | 32 | 32 |
| ASR | VUA | 10 | 30 | 20 | 20 |
| | TroFi | 5 | 5 | 5 | 5 |
| | MOH-X | 5 | 3 | 4 | 3 |
| | CHI | 5 | 5 | 5 | 5 |

**Table 3**
The overall metaphor identification results for all POS tags on the datasets VUA and CHI (%). The results of the state-of-the-art baselines are highlighted via *underline*, while the best results are shown in **boldface**.

| Model | VUA | | | CHI | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| CNN+RNN$_{ensmb}$ | 60.8 | 70.0 | 65.1 | – | – | – |
| RNN_ELMo$_{SEQ}$ | 71.6 | 73.6 | 72.6 | 76.7 | 56.0 | 64.7 |
| RNN_HG | 71.8 | 76.3 | 74.0 | 69.7 | 60.7 | 64.9 |
| RNN_MHCA | 73.0 | 75.7 | 74.3 | 69.1 | 66.7 | 67.9 |
| SEQ-CI + Cos | 72.4 | 76.6 | 74.4 | 70.1 | 69.0 | **69.5** |
| SEQ-CI + Edis | 72.6 | 76.7 | 74.6 | 70.6 | 67.6 | 69.1 |
| SEQ-CI + Gaussian | 75.5 | 74.1 | **74.8** | 69.3 | 68.0 | 68.7 |
| SEQ-CI + KL | 74.1 | 75.1 | 74.6 | 69.3 | 68.8 | 69.0 |

the proposed SEQ-CI models over baselines, we conduct a quantitative breakdown analysis on VUA_tst dataset, and report the metaphor identification results for each of the four types in Table 4. Overall the proposed models outperform the baselines across all different types of texts. Surprisingly, compared to the conversational (F1 = 68.2) or imaginary text (F1 = 70.2), it is perhaps relatively easy for the proposed models as well as baselines to deal with metaphor detection on the formal language data, e.g., academic articles (F1 = 81.0). The largest performance gap between the proposed models and baselines (F1 = 70.2 vs 67.7) is observed for the fiction type. This may be expected, and suggests that the proposed models can be more effective for identifying metaphorical language when applied to imaginary literature such as fiction text.

In addition, we also conduct a breakdown analysis of metaphor identification over four different part-of-speech (POS) tags, i.e., verbs, nouns, adjectives, and adverbs, where the metaphor tokens account for 18.1%, 13.6%, 11.5%, and 6.0% of all the tokens of the four POS tags, respectively. Table 5 shows the metaphor detection results for each POS tag on the VUA_tst dataset. Overall the proposed models again outperform the four state-of-the-art baselines across all the POS tags.

Specifically, compared to the baselines, the proposed models achieve the best metaphor detection performance in terms of F1 except for the adjectives, and show improved results on average for all POS tags. Surprisingly, the proposed models largely improve the metaphor identification for adverbs (F1 = 67.1 vs 63.8). One explanation is that the adverbial words, which are often used to add more information to verbs, adjectives, or other adverbs, may be more likely to cause the semantic contextual inconsistency in individual sentences.

### 4.5. The results of metaphor identification for verb track

To further validate the effectiveness of the proposed models, we conduct experiments for metaphor identification of verb track only. Table 6 shows the comparison results on the three datasets VUA_Verbs_tst, TroFi, and MOH-X. Clearly, the proposed models again outperforms all the baselines across the three datasets. In

**Table 4**
The metaphor identification results for different types of texts on the VUA_tst (%). The results of the state-of-the-art baselines are highlighted via *underline*, while the best results are shown in **boldface**.

| Type | Model | P | R | F1 |
|---|---|---|---|---|
| Academic | CNN+RNN$_{ensmb}$ | 72.5 | 74.6 | 73.5 |
| | RNN_ELMo$_{SEQ}$ | 78.2 | 80.2 | 79.2 |
| | RNN_HG | 76.5 | 83.0 | 79.6 |
| | RNN_MHCA | 79.6 | 80.0 | 79.8 |
| | SEQ-CI + Cos | 78.8 | 81.3 | 80.0 |
| | SEQ-CI + Edis | 80.7 | 81.3 | **81.0** |
| | SEQ-CI + Gaussian | 82.1 | 79.5 | 80.8 |
| | SEQ-CI + KL | 81.3 | 79.1 | 80.2 |
| Conversation | CNN+RNN$_{ensmb}$ | 45.3 | 71.1 | 55.3 |
| | RNN_ELMo$_{SEQ}$ | 64.9 | 63.1 | 64.0 |
| | RNN_HG | 63.6 | 72.5 | 67.8 |
| | RNN_MHCA | 64.0 | 71.1 | 67.4 |
| | SEQ-CI + Cos | 64.6 | 71.4 | 67.9 |
| | SEQ-CI + Edis | 63.1 | 71.5 | 67.0 |
| | SEQ-CI + Gaussian | 65.0 | 67.5 | 66.2 |
| | SEQ-CI + KL | 64.3 | 72.5 | **68.2** |
| Fiction | CNN+RNN$_{ensmb}$ | 48.3 | 69.2 | 56.9 |
| | RNN_ELMo$_{SEQ}$ | 61.4 | 69.1 | 65.1 |
| | RNN_HG | 61.8 | 74.5 | 67.5 |
| | RNN_MHCA | 64.8 | 70.9 | 67.7 |
| | SEQ-CI + Cos | 63.3 | 74.2 | 68.3 |
| | SEQ-CI + Edis | 65.1 | 73.8 | 69.2 |
| | SEQ-CI + Gaussian | 68.5 | 69.8 | 69.2 |
| | SEQ-CI + KL | 66.8 | 74.1 | **70.2** |
| News | CNN+RNN$_{ensmb}$ | 66.4 | 64.7 | 65.5 |
| | RNN_ELMo$_{SEQ}$ | 72.7 | 71.2 | 71.9 |
| | RNN_HG | 71.6 | 76.8 | 74.1 |
| | RNN_MHCA | 74.8 | 75.3 | 75.0 |
| | SEQ-CI + Cos | 74.4 | 74.8 | 74.6 |
| | SEQ-CI + Edis | 72.9 | 75.3 | 74.1 |
| | SEQ-CI + Gaussian | 77.3 | 73.0 | **75.1** |
| | SEQ-CI + KL | 76.0 | 71.9 | 73.9 |

**Table 5**
The metaphor identification results for different POS tags on the VUA_tst (%). The results of the state-of-the-art baselines are highlighted via *underline*, while the best results are shown in **boldface**.

| POS | Model | P | R | F1 |
|---|---|---|---|---|
| VERB | CNN+RNN$_{ensmb}$ | - | - | 67.4 |
| | RNN_ELMo$_{SEQ}$ | 68.1 | 71.9 | 69.9 |
| | RNN_HG | 66.4 | 75.5 | 70.7 |
| | RNN_MHCA | 66.0 | 76.0 | 70.7 |
| | SEQ-CI + Cos | 67.2 | 75.7 | 71.2 |
| | SEQ-CI + Edis | 67.1 | 77.3 | **71.8** |
| | SEQ-CI + Gaussian | 69.5 | 72.5 | 71.0 |
| | SEQ-CI + KL | 68.2 | 74.4 | 71.1 |
| NOUN | CNN+RNN$_{ensmb}$ | - | - | 62.9 |
| | RNN_ELMo$_{SEQ}$ | 59.9 | 60.8 | 60.4 |
| | RNN_HG | 60.3 | 66.8 | 63.4 |
| | RNN_MHCA | 69.1 | 58.2 | 63.2 |
| | SEQ-CI + Cos | 62.3 | 62.4 | 62.4 |
| | SEQ-CI + Edis | 65.5 | 59.1 | 62.1 |
| | SEQ-CI + Gaussian | 70.6 | 58.2 | 63.8 |
| | SEQ-CI + KL | 68.5 | 59.8 | **63.9** |
| ADJ | CNN+RNN$_{ensmb}$ | - | - | 65.1 |
| | RNN_ELMo$_{SEQ}$ | 56.1 | 60.6 | 58.3 |
| | RNN_HG | 59.2 | 65.6 | 62.2 |
| | RNN_MHCA | 61.4 | 61.7 | 61.6 |
| | SEQ-CI + Cos | 60.9 | 62.1 | 61.5 |
| | SEQ-CI + Edis | 63.5 | 63.8 | 63.7 |
| | SEQ-CI + Gaussian | 66.0 | 59.1 | 62.4 |
| | SEQ-CI + KL | 63.2 | 61.0 | 62.1 |
| ADV | CNN+RNN$_{ensmb}$ | - | - | 58.8 |
| | RNN_ELMo$_{SEQ}$ | 67.2 | 53.7 | 59.7 |
| | RNN_HG | 61.0 | 66.8 | 63.8 |
| | RNN_MHCA | 66.1 | 60.7 | 63.2 |
| | SEQ-CI + Cos | 66.9 | 67.2 | **67.1** |
| | SEQ-CI + Edis | 68.3 | 61.9 | 64.9 |
| | SEQ-CI + Gaussian | 74.0 | 60.7 | 66.7 |
| | SEQ-CI + KL | 70.5 | 60.7 | 65.2 |

addition, we conducted *t-test* on TroFi and MOH-X, where the 10-fold cross validation experiments have been implemented. On TroFi, the results of four variants of SEQ-CI are significantly different from that of the RNN_ELMo$_{SEQ}$ model at the given significance level $p = 0.01$. Compared to the baseline RNN_MHCA, the proposed models SEQ-CI + Cos, SEQ-CI + Edis, SEQ-CI + Gaussian and SEQ-CI + KL have resulted in significant improvements at the given significance levels $p = 0.05$, 0.1, 0.1, and 0.05, respectively. On MOH-X, the results of four variant SEQ-CI models are significantly different from that of the RNN_ELMo$_{SEQ}$ model at the given significance level of $p = 0.01$.

**Table 7**
The metaphor identification results for different types of texts on the VUA_Verbs_tst (%). The best results are highlighted via **boldface**.

| Types | Model | P | R | F1 |
|---|---|---|---|---|
| Academic | SEQ-CI + Cos | 74.2 | 82.8 | 78.2 |
| | SEQ-CI + Edis | 74.4 | 83.2 | **78.6** |
| | SEQ-CI + Gaussian | 74.6 | 81.5 | 77.9 |
| | SEQ-CI + KL | 71.9 | 83.2 | 77.1 |
| Conversation | SEQ-CI + Cos | 56.7 | 55.0 | 55.8 |
| | SEQ-CI + Edis | 56.7 | 58.4 | **57.5** |
| | SEQ-CI + Gaussian | 55.4 | 56.4 | 55.9 |
| | SEQ-CI + KL | 56.9 | 55.3 | 56.1 |
| Fiction | SEQ-CI + Cos | 55.2 | 71.4 | 62.3 |
| | SEQ-CI + Edis | 55.8 | 74.0 | **63.6** |
| | SEQ-CI + Gaussian | 58.2 | 65.2 | 61.5 |
| | SEQ-CI + KL | 53.6 | 68.9 | 60.3 |
| News | SEQ-CI + Cos | 73.8 | 80.5 | **77.0** |
| | SEQ-CI + Edis | 70.2 | 78.7 | 74.2 |
| | SEQ-CI + Gaussian | 72.9 | 78.4 | 75.5 |
| | SEQ-CI + KL | 71.8 | 81.6 | 76.4 |

In particular, on the VUA_Verbs_tst dataset, the traditional machine learning methods, i.e., Semantic-Feature and Abstractness, are not as good as the proposed SEQ-CI and other sequential neural network methods. It may suggest that automatically learning contextualized representations via sequential neural network models often leads to better prediction performance, compared to manual feature engineering in the traditional methods.

On MOH-X, the results of the SEQ-CI models are better than that of *Word-Similarity* on all evaluation metrics. Although *Word-Similarity* identified the metaphor by calculating the semantic similarity between pairs of words, it did not use the context and semantic contextual inconsistency of the words. Benefitting from learning from the semantic contextual inconsistency, overall the proposed SEQ-CI models achieve better results than the RNN_HG and RNN_MHCA models on all datasets. It may suggest that exploiting the semantic contextual inconsistency is more robust and generalizable than using the linguistic theories for metaphor identification.

Moreover, Table 7 shows the breakdown analysis of metaphor identification over the four different types of texts on the dataset VUA_Verbs_tst. It agrees well with our previous finding, and specifically, identifying metaphors from formal language (Academic) is relatively easier than that from colloquial text (Conversation) or imaginary literature (Fiction).

**Table 6**
The metaphor identification results for verbs on the three datasets (%). The results of the state-of-the-art baselines are highlighted via *underline*, while the best results are shown in **boldface**.

| Model | VUA_Verbs_tst | | | TroFi | | | MOH-X | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Semantic-Feature | 52.7 | 69.8 | 60.0 | – | – | – | – | – | – |
| Abstractness | – | – | 62.0 | – | – | – | – | – | – |
| Word-Similarity | – | – | – | – | – | – | 73.6 | 76.1 | 74.2 |
| CNN+RNN$_{ensmb}$ | 60.0 | 76.3 | 67.2 | – | – | – | – | – | – |
| RNN_ELMo$_{SEQ}$ | 68.2 | 71.3 | 69.7 | 70.7 | 71.6 | 71.1 | 79.1 | 73.5 | 75.6 |
| RNN_HG | 69.3 | 72.3 | 70.8 | 67.4 | 77.8 | 72.2 | 79.7 | 79.8 | 79.8 |
| RNN_MHCA | 66.3 | 75.2 | 70.5 | 68.6 | 76.8 | 72.4 | 77.5 | 83.1 | 80.0 |
| SEQ-CI + Cos | 68.1 | 75.7 | **71.7** | 70.6 | 76.4 | **73.3** | 82.2 | 79.0 | 80.2 |
| SEQ-CI + Edis | 67.0 | 76.3 | 71.4 | 69.8 | 76.5 | 72.9 | 77.6 | 83.8 | 80.3 |
| SEQ-CI + Gaussian | 68.4 | 73.8 | 71.0 | 69.7 | 76.1 | 72.7 | 79.2 | 82.3 | **80.5** |
| SEQ-CI + KL | 66.5 | 75.9 | 70.9 | 69.3 | 77.5 | 73.1 | 81.2 | 79.4 | 80.0 |

**Table 8**
The experimental results of SEQ and SEQ-CI models on VUA (%). The best result are highlighted via **boldface**.

| Models | P | R | F1 |
|---|---|---|---|
| SEQ | 71.0 | 75.5 | 73.2 |
| SEQ-CI + Cos | 72.4 | 76.6 | 74.4 |
| SEQ-CI + Edis | 72.6 | 76.7 | 74.6 |
| SEQ-CI + Gaussian | 75.5 | 74.1 | **74.8** |
| SEQ-CI + KL | 74.1 | 75.1 | 74.6 |

## 5. Discussions

### 5.1. Benefit of Exploiting Contextual Inconsistency

In this section, we designed the ablation study to evaluate the impact of contextual inconsistency of the proposed SEQ-CI model for metaphor identification. Specifically, we removed the contextual inconsistency term from loss function of our proposed model, which resulted in a reduced SEQ model. Table 8 shows the comparison results of SEQ-CI and SEQ models on VUA. Clearly, exploiting the semantic contextual inconsistency regularization term in learning improves the performance of metaphor identification by a large margin across all four different measurements of the semantic inconsistency. These results manifest remarkable capability of leveraging the semantic contextual inconsistency in the proposed sequential learning framework for identifying metaphorical words.

### 5.2. Influence of different contextual embedding

We then evaluate the influence of employing different state-of-the-art contextual embedding (CE) from ELMo and BERT for metaphor identification. Following the same framework of the proposed model, we replaced the existing ELMo embedding with the BERT embedding [41], which was extracted from the last layer of the BERT model ("bert-base-uncased").[3] The ELMo embedding and BERT embedding were fine-tuned while training the proposed models. Table 9 shows the results of metaphor identification based on different types of contextual embedding on VUA_VErbs_tst, MOH-X and TroFi.

Employing the contextual embedding from ELMo in the proposed models largely improves the metaphor detection results compared to that of the BERT model. BERT model has achieved remarkable results in multiple NLP applications, but it may not be suitable for obtaining contextualized embedding in this metaphorical identification task. In contrast, the ELMo model is introduced to encode context-sensitive dynamic embedding for downstream tasks, and then, as the distributed representations of contextual meanings of words, it may be more appropriate and beneficial for metaphor identification than BERT.

### 5.3. Quantitative analysis of the contextual inconsistency

The proposed models are based on the observation that the greater the semantic inconsistency between current word and contextual words is, the more likely the word belongs to the metaphorical category. To verify the idea, we visualize the distance or inconsistency of abstract semantic distributed representations of words. Specifically, we obtain the abstract semantic distributed representations $\{m_1, m_2, \ldots, m_N\}$ of an input sequence $\{w_1, w_2, \ldots, w_N\}$ by utilizing the trained SEQ-CI + Edis model via Eq. (11), and show the semantic distances between current word $w_i$ and contextual words using euclidean function (16). Intuitively, the larger the euclidean distance, the higher the contextual inconsistency. Fig. 2 shows the visualization results of the semantic inconsistencies in metaphorical and unmetaphorical example sentences.

Overall, lower contextual inconsistency can be seen in the unmetaphorical sentence compared to metaphorical sentence, as shown by light blue color in Fig. 2 (b). In the metaphorical sentence, the semantic inconsistencies between the word "won" and the contextual words are larger than of rest other words, e.g., "painting", "critical" and "acclaim", as shown in Fig. 2(a). It may suggest that the word "won" would be more likely to be predicted as metaphorical word compared with the other words. It's possible that the functional words ("the") or punctuations may show larger contextual inconsistency than content words in sentences. This is because the function words or punctuations are semantically far from any other words than the content words, as shown in Fig. 2 (a) and (b).

### 5.4. Hyper-parameter selection

In the proposed loss function (Eq. 23), the new introduced hyper-parameter $\lambda$ denotes the weight of the contextual inconsistency term in the loss. It is thus important to find the appropriate value of the hyper-parameter for model learning. In this section, we show the results of empirically selecting the best value of $\lambda$ on the VUA dataset, which may shed some light on the search and find of the best hyper-parameter in practice.

Specifically, we adopted the widely used grid-search method to tune the parameter $\lambda$ (from 0 to 1.0), and then selected the value that may lead to the best metaphor identification results on the development set VUA_dev. Fig. 3 examines the effect of changing the values of $\lambda$ on the F1 scores of the proposed SEQ-CI models on the development set. Clearly, $\lambda = 1.0$ may lead to decent validation results, and was thus selected for learning of the proposed models for detecting metaphorical words.

## 6. Conclusion

In this paper, we have developed a new sequence labeling model, SEQ-CI, which can learn from the semantic contextual inconsistency for improvement of metaphor identification. We conducted extensive experiments not only on publicly available English data but also on Chinese metaphor data, and the experimental results show the benefit of SEQ-CI over all the seven well-established baselines. The proposed metaphor identification model may provide support for downstream tasks in natural language processing. In particular, different from literal expression, metaphorical expressions tend to have a stronger emotional effect, and emotional content is formed by compositing and interacting the meanings of the source and target semantic context in the metaphors. The proposed model can be used to identify metaphorical use of words in natural language, and thus will serve as a functional module for metaphorical sentiment analysis for our future work.

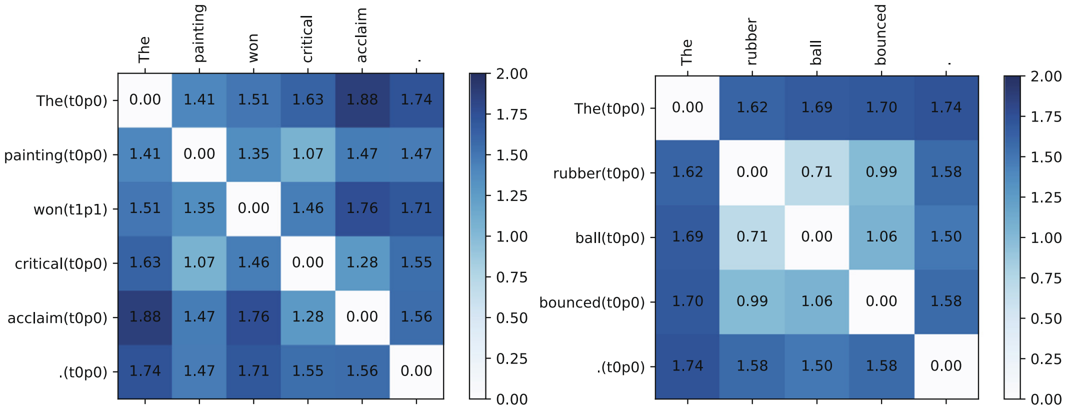
## CRediT authorship contribution statement

**Xin Chen:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing - original draft, Writing - review & editing. **Zhen Hai:** Conceptualization, Formal analysis, Methodology, Writing - review & editing. **Suge Wang:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing - review & editing. **Deyu Li:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing

[3] https://github.com/huggingface/pytorch-pretrained-BERT.

**Table 9**
The experimental results of employing different contextual embedding in the proposed models for metaphor identification on VUA, TroFi and MOH-X(%). The best results are highlighted via **boldface**.

| CE | Model | VUA_VErbs_tst | | | TroFi | | | MOH-X | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 | P | R | F1 |
| ELMo | SEQ-CI + Cos | 68.1 | 75.7 | **71.7** | 70.6 | 76.4 | **73.3** | 82.2 | 79.0 | 80.2 |
| | SEQ-CI + Edis | 67.0 | 76.3 | 71.4 | 69.8 | 76.5 | 72.9 | 77.6 | 83.8 | 80.3 |
| | SEQ-CI + Gaussian | 68.4 | 73.8 | 71.0 | 69.7 | 76.1 | 72.7 | 79.2 | 82.3 | **80.5** |
| | SEQ-CI + KL | 66.5 | 75.9 | 70.9 | 69.3 | 77.5 | 73.1 | 81.2 | 79.4 | 80.0 |
| BERT | SEQ-CI + Cos | 62.3 | 63.6 | 62.9 | 67.9 | 73.8 | 70.6 | 75.4 | 78.7 | 76.8 |
| | SEQ-CI + Edis | 60.2 | 68.1 | **63.9** | 66.7 | 75.7 | **70.7** | 74.9 | 80.7 | **77.4** |
| | SEQ-CI + Gaussian | 63.1 | 63.7 | 63.4 | 67.4 | 73.6 | 70.3 | 74.2 | 79.1 | 76.0 |
| | SEQ-CI + KL | 62.5 | 64.8 | 63.6 | 67.3 | 73.4 | 70.2 | 73.3 | 79.5 | 75.9 |



(a) Metaphorical sentence

(b) Unmetaphorical sentence

**Fig. 2.** The semantic inconsistencies between current word and contextual words in example sentences. ("t": truth, "p": prediction, "0": unmetaphorical, "1": metaphorical).

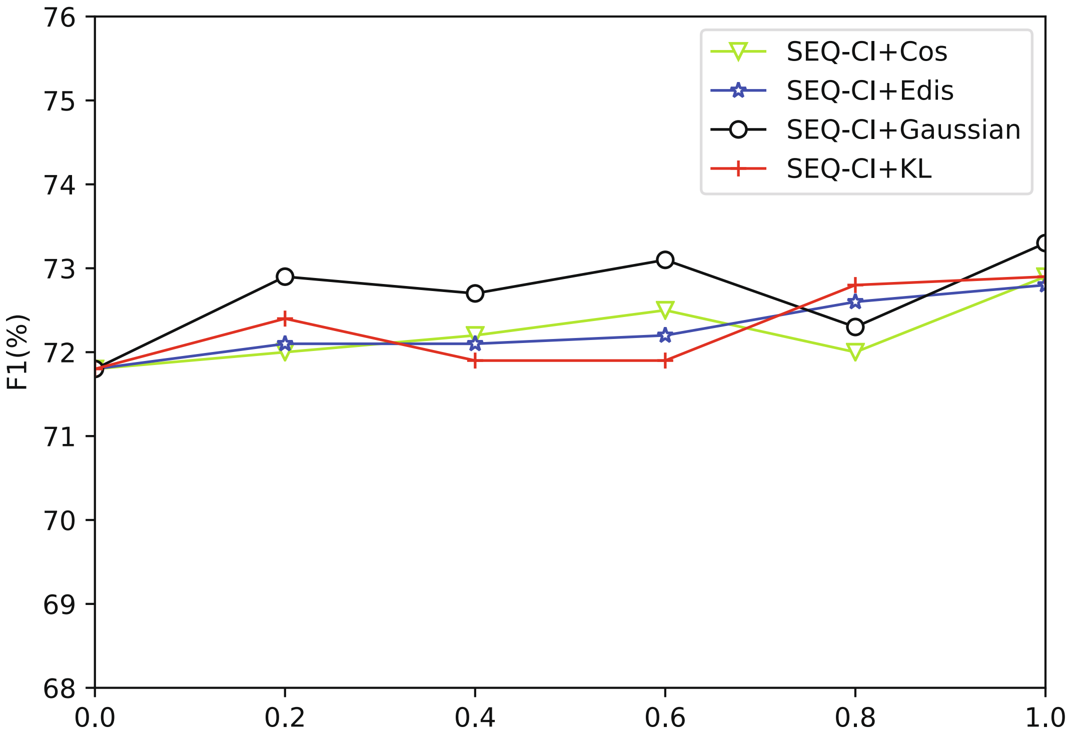


**Fig. 3.** The F1 scores of SEQ-CI models versus different values of *lambda* on VUA_dev dataset.

- review & editing. **Chao Wang:** Methodology. **Huanbo Luan:** Methodology.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors would like to thank all anonymous reviewers for their valuable comments and suggestions which have significantly improved the quality of this work. The work is supported by the National Key Research and Development Project of China (2018YFB1005103), and the National Natural Science Foundation of China (61632011, 61772302).

## References

[1] E. Shutova, Design and evaluation of metaphor processing systems, Comput. Linguist. 41 (4) (2015) 579–623.
[2] Y. Wang, Cognitive Linguistics, Shanghai Foreign Language Education Press, Shanghai, 2006.
[3] G. Lakoff, M. Johnson, Metaphors We Live by, University of Chicago press, 2008.
[4] B.B. Klebanov, C.W. Leong, E.D. Gutierrez, E. Shutova, M. Flor, Semantic classifications for detection of verb metaphors, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, 2016, pp. 101–106.
[5] E. Shutova, L. Sun, E.D. Gutiérrez, P. Lichtenstein, S. Narayanan, Multilingual metaphor processing: experiments with semi-supervised and unsupervised learning, Comput. Linguist. 43 (1) (2017) 71–123.
[6] S. Mohammad, E. Shutova, P. Turney, Metaphor as a medium for emotion: an empirical study, in: Proceedings of the 5th Joint Conference on Lexical and Computational Semantics, 2016, pp. 23–33.
[7] H. Haagsma, J. Bjerva, Detecting novel metaphor using selectional preference information, in: Proceedings of the 4th Workshop on Metaphor in NLP, 2016, pp. 10–17.
[8] J. Lederer, Finding metaphorical triggers through source (not target) domain lexicalization patterns, in: Proceedings of the 4th Workshop on Metaphor in NLP, 2016, pp. 1–9.
[9] P.D. Turney, Y. Neuman, D. Assaf, Y. Cohen, Literal and metaphorical sense identification through concrete and abstract context, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 680–690.
[10] L. Bulat, S. Clark, E. Shutova, Modelling metaphor with attribute-based semantics, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, 2017, pp. 523–528.
[11] M. Köper, S.S. im Walde, Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses, in: Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, 2017, pp. 24–30.
[12] I. Heintz, R. Gabbard, M. Srivastava, D. Barner, D. Black, M. Friedman, R. Weischedel, Automatic extraction of linguistic metaphors with LDA topic modeling, in: Proceedings of the 1st Workshop on Metaphor in NLP, 2013, pp. 58–66.
[13] E. Shutova, S. Teufel, A. Korhonen, Statistical metaphor processing, Comput. Linguist. 39 (2) (2013) 301–353.
[14] I.-H. Chen, Y. Long, Q. Lu, C.-R. Huang, Leveraging eventive information for better metaphor detection and classification, in: Proceedings of the 21st Conference on Computational Natural Language Learning, 2017, pp. 36–46.
[15] S. Rai, S. Chakraverty, D.K. Tayal, Supervised metaphor detection using conditional random fields, in: Proceedings of the 4th Workshop on Metaphor in NLP, 2016, pp. 18–27.
[16] H. Jang, Y. Jo, Q. Shen, M. Miller, S. Moon, C. Rose, Metaphor detection with topic transition, emotion and cognition in context, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2016, pp. 216–225.
[17] A. Mosolova, I. Bondarenko, V. Fomin, Conditional random fields for metaphor detection, in: Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 121–123.
[18] F.U. Jianhui, S. Wang, C. Cao, Chinese metaphor phrase recognition via combining the clustering and classification, J. Chin. Inf. Process. 32 (2) (2018), pp. 22–28+49.
[19] S. Sun, Z. Xie, BiLSTM-based models for metaphor detection, in: Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing, Springer, 2017, pp. 431–442.
[20] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 160–170.
[21] E.-L. Do Dinh, I. Gurevych, Token-level metaphor detection using neural networks, in: Proceedings of the 4th Workshop on Metaphor in NLP, 2016, pp. 28–33.
[22] G. Gao, E. Choi, Y. Choi, L. Zettlemoyer, Neural metaphor detection in context, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 607–613.
[23] M. Rei, L. Bulat, D. Kiela, E. Shutova, Grasping the finer point: a supervised similarity network for metaphor detection, arXiv preprint arXiv:1709.00575.
[24] K. Swarnkar, A.K. Singh, Di-LSTM contrast: a deep neural network for metaphor detection, in: Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 115–120.
[25] M. Pramanick, A. Gupta, P. Mitra, An LSTM-CRF based approach to token-level metaphor detection, in: Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 67–75.
[26] C. Wu, F. Wu, Y. Chen, S. Wu, Z. Yuan, Y. Huang, Neural metaphor detecting with CNN-LSTM model, in: Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 110–114.
[27] R. Mao, C. Lin, F. Guerin, End-to-end sequential metaphor identification inspired by linguistic theories, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3888–3898.
[28] M. Coltheart, The MRC psycholinguistic database, Q. J. Exp. Psychol. Sect. A 33 (4) (1981) 497–505.
[29] K. McRae, G.S. Cree, M.S. Seidenberg, C. McNorgan, Semantic feature production norms for a large set of living and nonliving things, Behav. Res. Methods 37 (4) (2005) 547–559.
[30] D. Schlechtweg, S. Eckmann, E. Santus, S.S. i. Walde, D. Hole, German in flux: detecting metaphoric change via word entropy, arXiv preprint arXiv:1706.04971.
[31] S. Rai, S. Chakraverty, D.K. Tayal, Y. Kukreti, Soft metaphor detection using fuzzy c-means, in: Proceedings of the International Conference on Mining Intelligence & Knowledge Exploration, 2017, pp. 402–411.
[32] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[34] Y. Song, S. Shi, J. Li, H. Zhang, Directional skip-gram: Explicitly distinguishing left and right context for word embeddings, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2, 2018, pp. 175–180.
[35] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2018, pp. 2227–2237.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[37] G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, Metaphor in usage, Cogn. Linguist. 21 (4) (2010) 765–796.
[38] J. Birke, A. Sarkar, A clustering approach for nearly unsupervised recognition of nonliteral language, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 329–336.
[39] C.W.B. Leong, B.B. Klebanov, E. Shutova, A report on the 2018 VUA metaphor detection shared task, in: Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 56–66.
[40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[41] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

**Xin Chen** received the M.S. degree in 2016 and is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Shanxi University, Shanxi, China since 2016. Her current research interests include natural language processing, text sentiment analysis.

**Zhen Hai** received the Ph.D. degree in computer science and engineering from Nanyang Technological University, Singapore. Zhen is with the DAMO Academy, Alibaba Group. His research interests include natural language processing, text mining, and machine learning.

**Suge Wang** received the Ph.D. degree from Shanghai University. She is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University. She has published more than 40 papers in international journals. Her research interests include natural language processing, text sentiment analysis, and machine learning.

**Deyu Li** received the Ph.D. degree from Xi'an Jiaotong University. He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University. He has published more than 60 papers in international journals. His research interests include artificial intelligence, granular computing, data mining, and machine learning.

**Chao Wang** is the CTO of 6Estates. Previously, he was a senior research scientist at Baidu. He holds a PhD in Computer Science from Tsinghua University. His work has appeared in major journals and conferences such as SIGIR, CIKM, TOIS, and IRJ. His main achievements include 2015 SIGIR Best Paper Honorable Mention Award, 2015 Scientific Technology Advance Award of Beijing City (First Prize), and 2016 CIPS Excellent PHD thesis Award. In 6Estates, he led the team to achieve 3rd Position in 2018 Chinese Machine Reading Comprehension competition and 3rd Position in 2019 Chinese Machine Reading Comprehension competition.

**Huanbo Luan** is the deputy director and senior research scientist of NExT Search Center at both Tsinghua University and National University of Singapore. He received his B.S. degree in computer science from Shandong University in 2003 and Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2008. His research interests include NLP, multimedia information retrieval, social media and big data analysis.