A Human-Like Semantic Cognition Network for Aspect-Level Sentiment Classification

Zeyang Lei,[1] Yujiu Yang,[*,1] Min Yang,[2] Wei Zhao,[3] Jun Guo,[4] Yi Liu[5]
[1]Graduate School at Shenzhen, Tsinghua University, [2]Shenzhen Institutes of Advanced Technology, CAS,
[3]Technische Universitt Darmstadt, [4]TBSI[†], Tsinghua Universitys, [5]Peking University Shenzhen Institute,
leizy16@mails.tsinghua.edu.cn, yangyj@gmail.com, min.yang1129@gmail.com
eeguojun@outlook.com, eeyliu@gmail.com

Abstract

In this paper, we propose a novel Human-like Semantic Cognition Network (HSCN) for aspect-level sentiment classification, motivated by the principles of human beings' reading cognitive process (pre-reading, active reading, post-reading). We first design a word-level interactive perception module to capture the correlation between context words and the given target words, which can be regarded as pre-reading. Second, to mimic the process of active reading, we propose a target-aware semantic distillation module to produce the target-specific context representation for aspect-level sentiment prediction. Third, we further devise a semantic deviation metric module to measure the semantic deviation between the target-specific context representation and the given target, which evaluates the degree we understand the target-specific context semantics. The measured semantic deviation is then used to fine-tune the above active reading process in a feedback regulation way. To verify the effectiveness of our approach, we conduct extensive experiments on three widely used datasets. The experiments demonstrate that HSCN achieves impressive results compared to other strong competitors.

Introduction

Sentiment analysis has attracted increasing attention recently due to its broad applications. The majority of literature addressed the sentiment for a whole piece of text, such as a document, a sentence etc. (Liu 2012). However, in real world, people may mention several target entities in one document/sentence. For example, the sentence "*although the service is not that great, I still like the food*" is positive regarding the food of the restaurant, but negative with regard to its service. Considering merely the overall sentiment of the sentence fails to capture the aspect-level sentiments. When inferring the sentiment in response to a given aspect, it is essential to effectively capture the relatedness of the aspect with its context words. Several recent studies have been proposed to build aspect-level sentiment classifiers with attention mechanisms and explicit memory. For example, Tang, Qin, and Liu (2016) developed deep memory networks to capture importance of context words. Wang et al. (2016) proposed to learn an embedding vector for each aspect, and these aspect embeddings were used to calculate the attention weights to capture important information with regard to a given aspect. As far as our knowledge, human-beings' reading cognitive process has rarely been explored in aspect-level sentiment classification.

When human-beings read and comprehend text, their exploration of the reading process organizes itself most naturally into an examination of three phrases: pre-reading, active reading (i.e., during-reading), and post-reading (Pressley M 1995; Avery and Graves 1997; Toprak and Almacıoğlu 2009). In the pre-reading stage, humans set the purpose of reading and preview the text with prior knowledge to form the initial reading cognition. Active reading is a complex cognitive process by which the reader constructs meaning from text. During the active reading stage, good readers use skimming ability to find relevant information to specific target by locating distinguishable target-related context, instead of thorough reading. It is beneficial to eliminate the distraction of irrelevant information so as to focus on the target-specific information.

As more information becomes available, the readers will evaluate their understanding of the text and revise their hypothesis when necessary. This post-reading stage (re-reading) creates opportunities for deeper understanding and error correction. If one desires to create a machine intelligence imitating such a reading comprehensive skill of humans, studying these three-stage human-beings' reading cognitive process is quite necessary.

In this paper, we propose a novel Human-like Semantic Cognition Network (HSCN) to simulate human-beings' reading cognitive process. HSCN consists of three components corresponding to the three stages in the human-beings' reading cognitive process (pre-reading, active reading, and post-reading). Concretely, we first design a word-level interactive perception module to capture the correlation between words in the context and the given target words, which can be regarded as pre-reading. Second, to mimic the process of active reading, we propose a target-aware semantic distillation module to produce the target-specific context representation for aspect-level sentiment prediction. Specifically, we devise a target-aware skip-reading mechanism to select target-related words from the context, acting as skimming. Then these target-related words are fed into the semantic

[*] Corresponding author

[†] Tsinghua-Berkeley Shenzhen Institute

composition module to encode the target-specific context representation. Third, we further design a semantic deviation metric module to measure the semantic deviation between our target-specific context representation and the given target, which evaluates the degree we understand the target-specific context semantics. The measured semantic deviation is then used to fine-tune the above active reading process in a feedback regulation way.

The main contributions can be summarized as follows:

- We propose HSCN, a human-like semantic cognition network to simulate the three stages in human-beings' reading cognitive process: pre-reading, active reading, and post reading.
- To fully model human-beings' reading cognitive process, our proposed HSCN is a hierarchical semantic network, including the word-level interactive perception module, target-aware semantic distillation module, semantic feedback module.
- To verify the effectiveness of our approach, we conduct extensive experiments on three widely used datasets. The experimental results demonstrate that our model achieves impressive results compared to other strong competitors.

Related work

Sentiment analysis is commonly explored at three levels: document level, sentence level and entity level. In this paper, we mainly focus on entity level sentiment classification. In order to capture the sentiments towards target entities, there are a variety of approaches being proposed for target-dependent or aspect-level sentiment classification. In (Hu and Liu 2004), the features of a product were considered to predict the sentiment polarity towards the target entity. Kiritchenko et al. (2014) trained one multi-class SVM classifier with two copy features, where each feature had two copies *f_general* (for all the aspect categories) and *f_c* (for the specific category). Tang et al. (2015) extended LSTM by considering the target word. Inspired by the recent success of attention-based neural networks (Bahdanau, Cho, and Bengio 2014; Yang et al. 2016), Wang et al. (2016) proposed an attention-based LSTM method to learn an aspect embedding for each aspect and made aspects participate in computing attention weights, and Liu and Zhang (2017) also designed an attention module for targeted sentiment classification. Tang, Qin, and Liu (2016) developed a deep memory network for aspect level sentiment classification which learned the importance of each context word and then utilized this information to calculate continuous text representation. Chen et al. (2017) employed a position-weighted memory network to capture sentiment features separated by a long distance. Ma et al. (2017) proposed the interactive attention networks (IAN) to interactively learn attention weights in the contexts and targets. Wang and Lu (2018) proposed a segmentation attention based LSTM model to effectively capture the structural dependencies between the target and the sentiment expressions with a linear-chain conditional random field (CRF) layer. Li et al. (2018) proposed a transformation network to model target-oriented sentiment classification by overcoming the drawbacks of attention mechanism and convolution neural networks. Xue and Li (2018) proposed a model based on convolutional neural networks and gating mechanisms.

Currently, some researchers began to attempt introducing human cognitive behaviors into their researches (Lake, Salakhutdinov, and Tenenbaum 2015; Zhao et al. 2017; Guo and Zhu 2018). To date, no work exploits the human-beings' reading cognitive process in aspect-level sentiment classification. Our work takes the lead in this topic.

Methodology

Overview

As discussed in Section 1, human-beings' reading cognitive process mainly consists of three phases: pre-reading, active reading, and post-reading. Accordingly, we propose a Human-like Semantic Cognition Network (HSCN) to simulate human-beings' reading cognitive process, which also consists of three components: word-level interactive perception module, target-aware semantic distillation module, and semantic feedback module.

Our goal is to strengthen the target-specific context representation learning through simulating human-beings' reading cognitive process. First, the word-level interactive perception module pre-reads text content and the given target to form an initial cognition, which captures the correlation between words in the context and the given target words. After initial perception, a target-aware skip-reading mechanism is proposed to select target-related words of the context. With the selected target-related words as inputs, we design a semantic composition module to obtain the target-specific context representation. Finally, we design a semantic feedback module to perform post-reading, which measures the semantic deviation between the target-specific context representation and the representation of the given target. This semantic deviation, to a certain extent, reflects the degree that we understand the semantics of the context. We use the semantic deviation to fine-tune the above active reading process in a feedback regulation way. In this section, we elaborate the three components in detail. To prevent conceptual confusion, we use a superscript "c" to indicate the variables that are related to context and a superscripts "t" to indicate the variables that are related to the given target for each context.

Word-level Interactive Perception Module

The word-level interactive perception module acts as the pre-reading in human-beings' reading cognitive process. It captures the correlation between the context words and the given target words by calculating the mutual information between them. The detailed implementation is similar with (Seo et al. 2017; Lei, Yang, and Yang 2018).

Assuming that a context x^c with n words can be denoted as $x^c = [w_1^c, w_2^c, ..., w_n^c]$ and a target x^t containing m words can be represented as $x^t = [w_1^t, w_2^t, ..., w_m^t]$. Here, each w represents a specific word. First, Each word w was embedded into a low-dimensional and dense vector space $\mathbf{e}(w) \in \mathbf{R}^d$, where $\mathbf{e}(\cdot)$ denotes the embedding operation and d refers to the dimension of the word vector. Accordingly, The

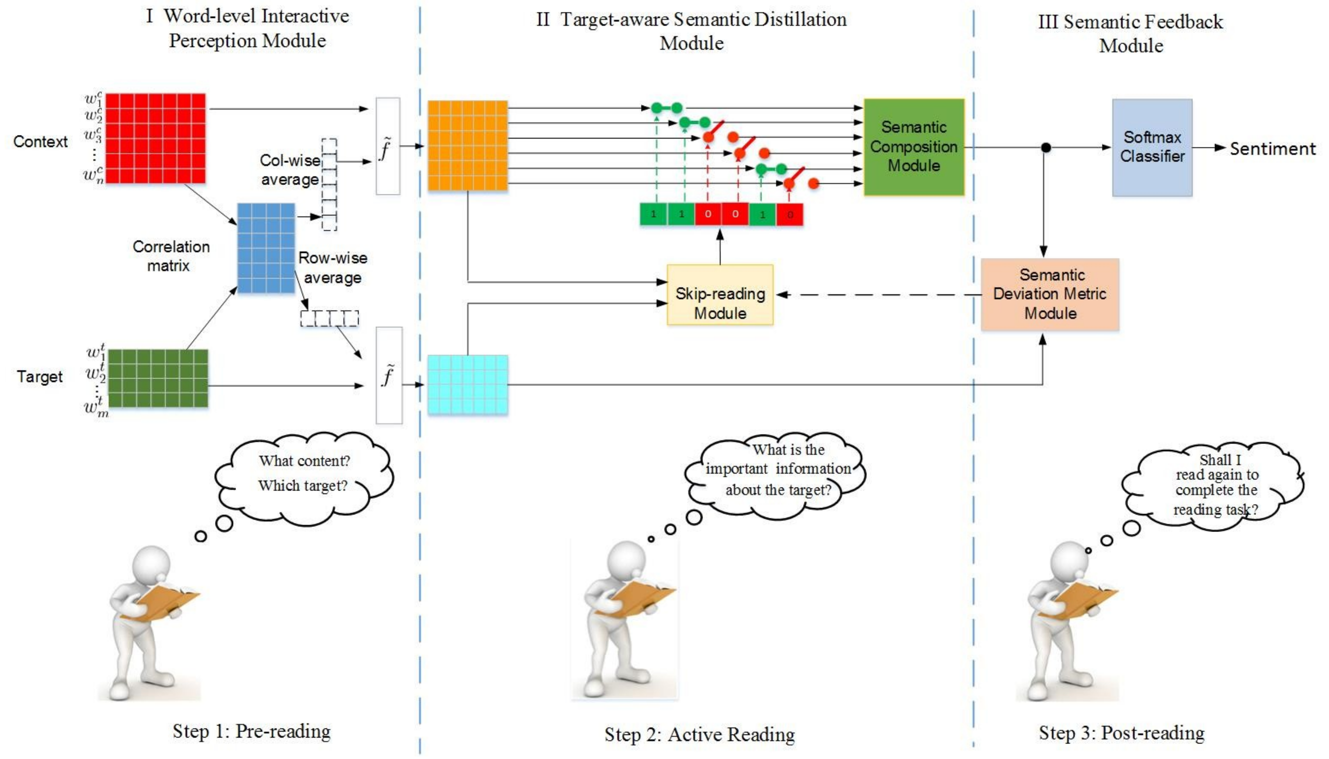
I Word-level Interactive Perception Module
II Target-aware Semantic Distillation Module
III Semantic Feedback Module
Context
w_1^c
w_2^c
w_3^c
w_n^c
Col-wise average
Correlation matrix
Row-wise average
Target
w_1^t
w_2^t
w_m^t
$\tilde{f}$
$\tilde{f}$
1 1 0 0 1 0
Semantic Composition Module
Skip-reading Module
Softmax Classifier
Sentiment
Semantic Deviation Metric Module
What content? Which target?
What is the important information about the target?
Shall I read again to complete the reading task?
Step 1: Pre-reading
Step 2: Active Reading
Step 3: Post-reading

Figure 1: The overall Framework of Human-like Semantic Cognition Network (HSCN)

embeddings of the context x^c and the corresponding target x^t can be obtained as $\mathbf{e}(x^c) = [\mathbf{e}(w_1^c), \mathbf{e}(w_2^c), ..., \mathbf{e}(w_n^c)]$ and $(x^t) = [\mathbf{e}(w_1^t), \mathbf{e}(w_2^t), ..., \mathbf{e}(w_m^t)]$, respectively.

Then, we compute the correlation matrix M to represent the relatedness between the context and the target, as follow:

$$M = \mathbf{e}(x^c)^T \cdot \mathbf{e}(x^t) \in \mathbb{R}^{n \times m} \tag{1}$$

where each element $M_{i,j}$ in the correlation matrix refers to the relevance between the w_i^c in the context and the w_j^t in the target.

Then, we perform mean-pooling operation on the values of each row of S and obtain a column vector of size n. This obtained column vector is fed into a *softmax* layer, producing an attention vector for the context:

$$\boldsymbol{\rho} = softmax(\frac{\sum_{i=1}^{m} M[:, i]}{m}) \tag{2}$$

Similarly, a row vector of size m can be obtained by computing the average values of each column of S. We use the row vector as the input of a *softmax* layer to get an attention vector for the target words:

$$\boldsymbol{\sigma} = softmax(\frac{\sum_{j=1}^{n} M[j, :]}{n}) \tag{3}$$

Finally, with the context embedding matrix $\mathbf{e}(x^c)$ and the attention vector $\boldsymbol{\rho}$ as inputs, we can compute the target-enhanced context representation matrix $W^c \in \mathbb{R}^{k \times n}$ as follows:

$$W^c = \mathbf{ELU}(U^c(\mathbf{e}(x^c) + (\mathbf{e_d} \otimes \boldsymbol{\rho}) \odot \mathbf{e}(x^c))) \tag{4}$$

where U^c is a projection parameter, $\mathbf{e_d} = [1, 1, .., 1]^T$ represents a d-dimensional all-ones vector, $\mathbf{e_d} \otimes \boldsymbol{\rho} = [\boldsymbol{\rho}; \boldsymbol{\rho}; ...; \boldsymbol{\rho}]$ denotes the kronecker product operation between $\mathbf{e_d}$ and $\boldsymbol{\rho}$, $\odot$ refers to the element-wise multiplication, and exponential linear unit (ELU) (Clevert, Unterthiner, and Hochreiter 2016) is a nonlinear activation function. In the same way, we can also obtain the contextual enhanced target representation matrix $W^t \in \mathbb{R}^{k \times m}$ as follow:

$$W^t = \mathbf{ELU}(U^t(\mathbf{e}(x^t) + (\mathbf{e_d} \otimes \boldsymbol{\sigma}) \odot \mathbf{e}(x^t))) \tag{5}$$

where U^t is a projection parameter, $\mathbf{e_d} = [1, 1, .., 1]^T$ denotes a d-dimensional all-ones vector, $\mathbf{e_d} \otimes \boldsymbol{\sigma} = [\boldsymbol{\sigma}; \boldsymbol{\sigma}; ...; \boldsymbol{\sigma}]$ denotes the kronecker product operation between $\mathbf{e_d}$ and $\boldsymbol{\sigma}$.

The Eq. (1-5) can help establish the relatedness between the context and the target and thus we call the process as word-level interactive perception module, which is beneficial for the following active reading process.

Target-aware Semantic Distillation Module

As depicted in Figure 1, the target-aware semantic distillation module contains two parts: the skip-reading module for selecting target-related words and the semantic composition module for encoding the target-specific context representation. In the following, we will elaborate the two parts in detail.

Skip-reading Module After the pre-reading procedure, we design a target-aware skip-reading module to comprehend the target-specific context and select out the crucial target-related words from the context. Specifically, with the context representation matrix W^c and the target representation matrix W^t as inputs, we build a semantic decision-making scheme to extract target-related information and determine whether a word in the context should be deleted or retained. Formally, we formulate this process as follows:

$$\boldsymbol{\gamma} = softmax(U^{\boldsymbol{\gamma}}\{W^t\}^T W^c) \tag{6}$$

$$\mathbf{g} = argmax(\boldsymbol{\gamma}^{\boldsymbol{T}}) \tag{7}$$

where $U^{\boldsymbol{\gamma}}$ is a learnable parameter, $\boldsymbol{\gamma} \in \mathbb{R}^{2\times n}$ denotes a skip-reading state matrix with two dimensions indicating the probability of being skipped and not be skipped respectively, $\mathbf{g} \in \mathbb{R}^n$ is a gate-controlled vector with each element value being 0 or 1. The value being 0 denotes that the corresponding word should be deleted while the value being 1 denotes that the corresponding word should be retained.

Note that the $argmax$ operation is a hard-decision process and non-differentiable, direct application of the gradient propagation may be difficult. For the gradient estimation of discrete variables, some feasible approaches achieve unbiased gradient estimation using the REINFORCE algorithm (Williams 1992; Ke et al. 2018). However, the REINFORCE algorithms suffer from high variance of the gradient estimation, model instability, and hard training (V et al. 2015; Maddison, Mnih, and Teh 2017). We instead use a **Gumbel-Softmax** distribution (Gumbel and Lieblein 1954; Seo et al. 2018) to approximate Equation (7), which is fully differentiable. The approximate implementation $\tilde{\mathbf{g}}$ of $\mathbf{g}$ is shown as follow:

$$\tilde{\mathbf{g}} = \frac{exp((log(\boldsymbol{\gamma}^T) + \boldsymbol{\delta})/\tau)}{\sum exp((log(\boldsymbol{\gamma}^T) + \boldsymbol{\delta})/\tau)} \tag{8}$$

where $\boldsymbol{\delta}$ is an independent sample from Gumbel(0,1)=-log(-log(Uniform(0,1))) and τ is the temperature coefficient, an inherent hyper-parameter of Gumbel-Softmax function to adjust the approximation degree towards one-hot sampling. Note that when $\tau \rightarrow 0$, the gumbel-softmax operation approximately approaches discrete one-hot sampling, namely the $argmax$ operation. Generally, $\tau <= 0.1$ can obtain a sharper distribution and $\tau = 0.5 \sim 1$ can obtain a more smoothing distribution.

Given $\tilde{\mathbf{g}}$[1] as the final gate-controlled vector, we formulate the distilled representation matrix W^r of the context that is composed of the selected target-related words, as follow:

$$W^r = W^{cT} \odot (\mathbf{e_k} \otimes \tilde{\mathbf{g}}) \tag{9}$$

where $\mathbf{e_k} = [1, 1, .., 1]^T$ denotes a k-dimensional all-ones vector, $\odot$ denotes the element-wise multiplication, and $\mathbf{e_k} \otimes \tilde{\mathbf{g}}$ denotes the kronecker product operation between $\mathbf{e_k}$ and $\tilde{\mathbf{g}}$.

Semantic Composition Module After obtaining the target-related words in the context, we devise an appropriate semantic composition module to encode the target-specific context representation. To be specific, with the distilled context representation matrix W^r as inputs, we first employ **GRU** networks (Chung et al. 2015) to obtain the hidden states H^r of the context,

$$H^r = \mathbf{GRU}(W^r) \tag{10}$$

Then, motivated by the superiority of attention models in sentence semantic composition (Lin et al. 2017; Shen et al. 2018; Shaw, Uszkoreit, and Vaswani 2018), we also design a target-aware attention mechanism to learn the compositional weights about the hidden states of the context words. Concretely, we first encode the contextual enhanced target representation matrix $W^{\boldsymbol{t}}$ to obtain its hidden representation H^t. Then we execute the mean-pooling operation over H^t. Finally, we utilize the result of mean-pooling operation as attention source to learn the attention weights over the context. The detailed process can be summarized as follows:

$$H^t = \mathbf{GRU}(W^t) \tag{11}$$

$$\mathbf{z_t} = \sum_{i=1}^{m} H^t[i,:]/m \tag{12}$$

$$\boldsymbol{\varphi} = softmax(f(\hat{U}[H^r; (\mathbf{e_n} \otimes \mathbf{z_t}])) \tag{13}$$

$$\mathbf{o} = \{H^r\}^T \boldsymbol{\varphi} \tag{14}$$

where $\hat{U}$ is a learnable parameter, φ refers to the attention weights for the target-related context words, f denotes the nonlinear activation function. Here, we may choose $tanh$, elu, or $relu$ as f. $H^t[i,:]$ denotes the i-th row vector of H^t and $\mathbf{o}$ is the final target-specific context representation.

Semantic Feedback Module

After obtaining the target-specific context representation, we design a semantic deviation metric module to measure the semantic deviation between our target-specific context representation and the given target, which models the degree we understand the target-specific context semantics. In particular, we adopt a fully connection network to project the target-specific context representation to the target semantic space. In the target semantic space, we directly use the residual vector between the projected context vector and the target vector as semantic comprehension deviation. The obtained semantic residual vector is then used to fine-tune the aforementioned skip-reading step in a feedback regulation way. The detailed implementation process is formulated as follow.

$$\Delta\mathbf{r} = \mathbf{z^t} - tanh(U^o\mathbf{o} + \mathbf{b^o}) \tag{15}$$

$$\hat{\boldsymbol{\gamma}} = softmax(U^r \Delta\mathbf{r}^T H^r) \tag{16}$$

$$\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma} + \eta\hat{\boldsymbol{\gamma}} \tag{17}$$

where $\Delta\mathbf{r}$ is the semantic comprehending deviation, $\hat{\boldsymbol{\gamma}}$ is the incremental skip-reading state matrix via semantic deviation, $\tilde{\boldsymbol{\gamma}}$ is the final augmented skip-reading state matrix by fine-tuning the pervious skip-reading result γ, and U^o, U^r, $\mathbf{b^o}$ are projection parameters,η is the hyperparameter, which can be set as 1 in this paper.

Finally, with the augmented skip-reading state matrix $\tilde{\boldsymbol{\gamma}}$, the Equation (8-14) (noted as $\mathbf{SR}$) can be performed once

[1] To simply, We randomly choose any one of the two dimension of $\tilde{\mathbf{g}}$ as the final $\tilde{\mathbf{g}}$ to participate in the following calculation.

again to obtain the regulated final enhanced context semantics $\mathbf{o_1}$ via **rereading** as follow.

$$\mathbf{o_1} = \mathbf{SR}(\tilde{\boldsymbol{\gamma}}) \tag{18}$$

where $\mathbf{SR}$ denotes the entire transformation process of one-turn rereading. In real world, human beings may perform multi-turn rereading to determine the final semantics of the sentences.

Sentiment Classifier

The final context representation $\tilde{\mathbf{o}} = [\mathbf{o}, \mathbf{o_1}]$ can be fed to a softmax function to predict sentiment polarity distribution:

$$\hat{y} = softmax(U_o^T \tilde{\mathbf{o}} + b_o) \tag{19}$$

where $\hat{y}$ is the predicted sentiment polarity distribution, U_o and b_o are learned parameters. Suppose that a training corpus contains N training samples (x_i, y_i), we can train the entire networks by minimizing the following loss function.

$$L(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j log(\hat{y}_i^j) + \lambda(\sum_{\theta \in \Theta} \theta^2) \tag{20}$$

where y_i^j is the ground truth sentiment polarity, C is the number of sentiment polarity categories, $\hat{y}_i^j$ denotes the predicted sentiment probabilities, θ represents each parameter to be regularized, Θ is a collection of all parameters, λ is the weight coefficient for L_2 regularization.

Experimental Setup

Datasets

SemEval-14 This data is constructed for the aspect based sentiment analysis task (SemEval-2014 Task 4)[2]. Two domain-specific datasets for *restaurants* (S-res.) and *laptops* (S-laptop) have been provided for training and testing. Table 1 shows the statistics of these two datasets. There are five aspect categories in the restaurants dataset, including *price, food, service, ambience, anecdotes/miscellaneous*. Since the laptop dataset does not have the aspect-specific polarities, we modified the dataset to include annotations for aspect categories and their sentiment polarities (i.e., positive, negative, neutral). The tags of the aspect category field for the laptop dataset are *performance, price, quality, appearance*.

Tweets

The original dataset is a collection of tweets from Twitter by (Dong et al. 2014), using keywords (e.g., “bill gates”, “google”) to query the Twitter API. Each tweet has a manually labeled sentiment polarity (i.e., positive, neutral or negative) for the target entity (keywords). The training data consists of 6,248 tweets, and the testing data has 692 tweets. The percentages of positive, negative and neutral tweets in both the training set and the test set are 25%, 25%, 50%, respectively.

Dataset	Pos.	Neg.	Neu.	Total
S-res. (train)	2164	807	637	3608
S-res. (test)	728	196	196	1120
S-laptop (train)	994	870	464	2238
S-laptop (test)	341	128	169	638

Table 1: Statistics of SemEval-14 dataset.

Baselines

In the experiments, we evaluate and compare our model with several baseline methods, including **SVM-feature** (Kiritchenko et al. 2014), **LSTM**, **TD-LSTM** (Tang et al. 2015), **ATAE-LSTM** (Wang et al. 2016), **MemNet** (Tang, Qin, and Liu 2016),**RAM** (Chen et al. 2017),**IAN** (Ma, Yuan, and Wu 2017), **SA-LSTM-P** (Wang and Lu 2018),**PRET+MULT** (He et al. 2018). For more details, please refer to the corresponding references.

Implementation Details

Model hyper-parameters are set by a grid search. To avoid overfitting, we use a dropout strategy (Wager, Wang, and Liang 2013) to randomly omit part of the feature detectors on each training case. Meanwhile, to reduce the vulnerability of neural networks, we adopt the *label smoothing technique* (Szegedy et al. 2016), which replaces output vectors $y_{label} = [1, 0, 0, 0, ..., 0]$ with $y_{label} = [1 - \epsilon, \frac{\epsilon}{C-1}, \frac{\epsilon}{C-1}, \frac{\epsilon}{C-1}, ..., \frac{\epsilon}{C-1}]$, where ϵ is smoothing coefficient.

In our experiments, we use 300-dimensional GloVe[3] vectors to initialize the word embeddings for words in the context and target words, and all out-of-vocabulary words are initialized by sampling from the uniform distribution U(-0.25,0.25). We initialize all the weight matrices as random orthogonal matrices, and all the bias vectors are initialized to zero. The dimension of GRU hidden states is set as 10. We conduct mini-batch (with size 40) training using RMSprop optimization algorithm to train the model. The dropout rate is set to 0.5, and the coefficient λ of L_2 normalization is set to 10^{-5}. Label smoothing coefficient ϵ is set as 0.01.

Experimental Results

Main Results

In our experiments, the evaluation metrics are classification accuracy and Macro-averaged F1 (D. Manning, Schtitze, and Lee 2002). We report the best results and the average results by running our model five times. The mean and the standard deviation of our model is also shown in Table 2.

We summarize the experimental results in Table 2. As is shown in Table 2, SVM-feature obtains an impressive result in Laptop and Restaurant datasets. But it needs labor-intensive feature engineering works and an amout of extra linguistic resources, which limits its advantage. The LSTM performs poorly since they do not consider the target-specific information when deciding the sentiment polarity of different aspects. TD-LSTM obtains a better result

[2] http://alt.qcri.org/semeval2014/

[3] http://nlp.stanford.edu/projects/glove

Methods	Laptop		Restaurant		Tweets	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
SVM-feature	70.5	NA	80.2	NA	63.4*	63.3*
LSTM	66.5	60.1	74.3	63.0	66.5	64.7
TD-LSTM	68.1	63.9	75.6	64.5	66.6*	64.0*
ATAE-LSTM	68.7	64.2	77.6	65.3	67.7	65.8
MemNet	68.9#	62.8#	76.9#	66.4#	68.5	66.9
RAM	72.1#	68.4#	78.5#	68.5#	69.4	67.3
IAN	72.1	NA	78.6	NA	NA	NA
SA-LSTM-P	75.1	NA	**81.6**	NA	60.9	NA
PRET+MULT	71.2	67.5	79.1	69.7	NA	NA
HSCN	**76.1** (74.9 ± 1.0)	**72.5** (70.9 ± 1.1)	77.8 (77.1 ± 0.6)	**70.2** (70.0 ± 0.3)	**69.6** (68.8 ± 0.8)	66.1 (65.4 ± 0.7)

Table 2: Evaluation results. The best result on each dataset is in bold. The results with $*$ are retrieved from (Chen et al. 2017), and the results with # are retrieved from (He et al. 2018). For HSCN, the upper results represent the best performance for the model and the lower are the mean and the standard deviation when running the model five times.

than LSTM because it considers the left and right context with target when modeling the sentence representation. The attention-based models such as ATAE-LSTM and IAN consistently perform better than LSTM and TD-LSTM. This may be because that these attention-based models capture the important information in response to the given aspect. Memory-based methods such as MemNet and RAM also obtain comparable results, which verifies the effectiveness of integrating memory into sentiment modeling. **Our model** performs even better than the strong attention-based or memory-based competitors by simulating human-beings' reading cognitive process. For example, for the Laptop dataset, the average classification accuracy increases by 2.8% and the average Macro-F1 also increases by 2.5% than RAM, the strongest baselines until 2017. As for the Restaurant dataset, the average Macro-F1 also obtains the start-of-the-art result compared with other methods. For Tweets, our model can also get a comparable result, which verifies our model's effectiveness to some extent.

Quantitative Analysis

The Effect of Each Component To further investigate the effect of each component of the HSCN model, we also conduct the ablation test of HSCN in terms of discarding word-level interactive perception module (denoted as HSCN-I), skip-reading module (denoted as HSCN-II), semantic composition module (denoted as HSCN-III), and semantic feedback module (denoted as HSCN-IV). The results are reported in Figure 3. Generally, all four factors contribute, and semantic composition module contributes most. This is within our expectation since the third component refers to the final context semantics comprehension that plays most important role in target-specific text understanding. The pre-reading component also makes great contribution to aspect-level sentiment classification. We believe this is because that the pre-reading module, establishing the correlation relation between the target and the content, is the basis of the latter comprehension stages. This also inspires us in the future text modelling, pre-reading and semantic composition should be paid more attention.

Effect of the Number of Readings

Number of Readings	Acc	F1	Runtime(s)
1	75.6	71.8	7
2	76.1	72.5	9
3	76.1	71.7	13
4	75.5	70.5	15
5	75.1	70.4	18

Table 3: The experimental results for the models with the different numbers of readings on the Laptop dataset. Runtime denotes the running time for each training epoch.

One may wonder how many times of readings is appropriate for a reader. Therefore, we study the performance of the models with different times of reading in this part. Specifically, we implement the HSCN model with different number of copies of semantic feedback module based on the same network infrastructure and run them on the same CPU server. The detailed results are shown in Table 3. We can find that, with reading times increasing from 1 to 3, the classification accuracy increases from 75.6 to 76.1 and the runtime of each training epoch also increases from 7s to 13s. This verifies that rereading can really enhance our understanding for context semantics but suffers expensive time cost. When reading times surpasses 3, we find that the performance declines instead to some extent but the runtime still increases by a certain margin. We infer that one possible reason is that the model is over-fitting with the increasing of hyper-parameters when the number of reading times increases. In our experiment, taking into account the performance and the runtime, we choose only reread once.

Case Study

We use an exemplary case which is randomly selected from the test set of SemEval-14 restaurant data to demonstrate the reading comprehension process of our model by visualizing the attention results of each step. The selected context is "*The falafal was rather over cooked and dried but the*

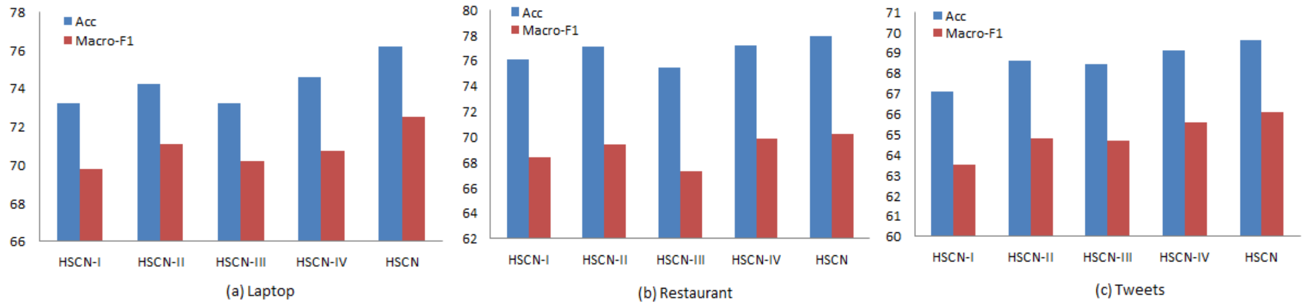

Figure 2: Ablation Test results

chicken was fine.”, and the corresponding target is *“falafal”*. Our model successfully predicts the sentiment polarity towards *“falafal”* as negative. Figure 3 shows how attention weights vary with the influence of a given target in the reading cognitive process (i.e., pre-reading, active reading, post-reading). The color depth indicates the importance degree of the attention. The darker the color, the more important the word. Figure 3(I) denotes the attention weights of the context words toward the given target after performing pre-reading. We reveal that the context words related to the target *“falafal”* such as *“falafal”*, *“over”*, *“cooked”*, *“dried”* are paid much more attention than the irrelevant words such as *“the”*, *“and”*. This verifies that pre-reading can establish the initial correlation between the target and the context. Then, active reading process locates the most important features and ignores the rest. As shown in 3(II), the importance of target-related words are risen, while the target-irrelevant words are ignored. This step can be regarded as skimming in the process of the reading comprehension, which eliminates the distractions of irrelevant words and obtains the distilled if. Finally, post reading is performed to fine tune the active reading process. Figure 3(III) shows that the attention weights after post reading (rereading) sightly. In particular, some new target-related words such as *”rather”,”dried”* are added while some selected irrelevant words are discarded. This verifies that post reading (rereading) creates the opportunities for deeper understanding of the text and error correction.

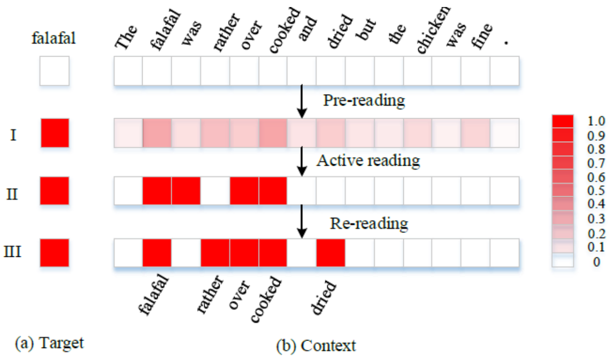

Figure 3: Case Study

Conclusion and Future Work

In this paper, motivated by the principles of human being’s reading cognitive process, we proposed a novel Human-like Semantic Cognition Network (HSCN) for aspect-level sentiment classification. HSCN consists of three components, including word-level interactive perception module, target-aware semantic distillation module, and semantic feedback module. To the best of our knowledge, we take the lead to concretely exploit human-beings’ reading cognitive process for aspect- level sentiment analysis. Experiments on three widely used datasets showed the superiority of HSCN.

As for future work. we will work on the two aspects: one is that we will explore more proper ways to model the human reading cognition process, and the other is that we will attempt applying our model in longer text to further verify the effectiveness of human reading cognition process.

Acknowledgements

We sincerely thank all the anonymous reviewers for their valuable comments to improve this paper. This work was supported in part by the National Key Research and Development Program of China (No. 2017YFC1601004), and Shenzhen special fund for the strategic development of emerging industries (No. JCYJ20170412170118573 and JCYJ20160331104524983). Zeyang Lei and Jun Guo were partly supported by the 2018 Tencent Rhino-Bird Elite Training Program. Min Yang was sponsored by CCF-Tecent Open Research Fund.

References

Avery, P. G., and Graves, M. F. 1997. Scaffolding young learners’ reading of social studies texts. *Social Studies & the Young Learner* 9:10–14.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science*.

Chen, P.; Sun, Z.; Bing, L.; and Yang, W. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *EMNLP*, 452–461.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2015. Gated feedback recurrent neural networks. *Computer Science* 2067–2075.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and accurate deep network learning by exponential linear units (elus).

D. Manning, C.; Schtitze, H.; and Lee, L. 2002. Review: Foundations of statistical natural language processing, christopher d. manning and hinrich schütze.

Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; and Xu, K. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *ACL*, 49–54.

Gumbel, E. J., and Lieblein, J. 1954. *Statistical theory of extreme values and some practical applications: a series of lectures*.

Guo, J., and Zhu, W. 2018. Partial multi-view outlier detection based on collective learning. In *Proc. AAAI Conf. Artificial Intell. (AAAI)*, 298–305.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *Proceedings of ACL(2)*.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *SIGKDD*, 168–177.

Ke, N. R.; Zolna, K.; Sordoni, A.; Lin, Z.; Trischler, A.; Bengio, Y.; Pineau, J.; Charlin, L.; and Pal, C. 2018. Focused hierarchical rnns for conditional sequence processing. In *Proceedings of ICML*.

Kiritchenko, S.; Zhu, X.; Cherry, C.; and Mohammad, S. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *International Workshop on Semantic Evaluation*, 437–442.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338.

Lei, Z.; Yang, Y.; and Yang, M. 2018. Saan: A sentiment-aware attention network for sentiment analysis. In *Proceedings of SIGIR*.

Li, X.; Bing, L.; Lam, W.; and Shi, B. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of ACL*.

Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. In *ICLR*.

Liu, J., and Zhang, Y. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 1–167.

Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2017. Interactive attention networks for aspect-level sentiment classification. In *IJCAI*, 4068–4074.

Ma, B.; Yuan, H.; and Wu, Y. 2017. Exploring performance of clustering methods on document sentiment analysis. *Journal of Information Science* 43(1):54–74.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *ICLR*.

Pressley M, A. P. 1995. *Verbal protocols of reading: The nature of constructively responsive reading*.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of ICLR*.

Seo, M.; Min, S.; Farhadi, A.; and Hajishirzi, H. 2018. Neural speed reading via skim-rnn. In *Proceedings of ICLR*.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. In *NAACL*.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2015. Target-dependent sentiment classification with long short term memory. *Computer Science*.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect level sentiment classification with deep memory network. In *EMNLP*, 214–224.

Toprak, E. L., and Almacıoğlu, G. 2009. Three reading phases and their applications in the teaching of english as a foreign language in reading classes with young learners. *Journal of Language & Linguistic Studies* 5(1):20–36.

V, M.; K, K.; D, S.; AA, R.; J, V.; MG, B.; A, G.; M, R.; AK, F.; and G, O. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Wager, S.; Wang, S.; and Liang, P. S. 2013. Dropout training as adaptive regularization. In *NIPS*, 351–359.

Wang, B., and Lu, W. 2018. Learning latent opinions for aspect-level sentiment classification. In *AAAI*.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, 606–615.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8(3):229–256.

Xue, W., and Li, T. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of ACL*.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Zhao, W.; Kong, Y.; Ding, Z.; and Fu, Y. 2017. Deep active learning through cognitive information parcels. In *Proceedings of the 2017 ACM on Multimedia Conference*, 952–960. ACM.