Target-Guided Structured Attention Network for Target-Dependent Sentiment Analysis

Ji Zhang   Chengyao Chen   Pengfei Liu   Chao He   Cane Wing-Ki Leung

Wisers AI Lab, Wisers Information Limited, HKSAR, China
{jasonzhang, stacychen, chaohe, caneleung}@wisers.com, ppfliu@gmail.com

Abstract

Target-dependent sentiment analysis (TDSA) aims to classify the sentiment of a text towards a given target. The major challenge of this task lies in modeling the semantic relatedness between a target and its context sentence. This paper proposes a novel *Target-Guided Structured Attention Network* (TG-SAN), which captures target-related contexts for TDSA in a fine-to-coarse manner. Given a target and its context sentence, the proposed TG-SAN first identifies multiple semantic segments from the sentence using a target-guided structured attention mechanism. It then fuses the extracted segments based on their relatedness with the target for sentiment classification. We present comprehensive comparative experiments on three benchmarks with three major findings. First, TG-SAN outperforms the state-of-the-art by up to 1.61% and 3.58% in terms of accuracy and Marco-F1, respectively. Second, it shows a strong advantage in determining the sentiment of a target when the context sentence contains multiple semantic segments. Lastly, visualization results show that the attention scores produced by TG-SAN are highly interpretable

1 Introduction

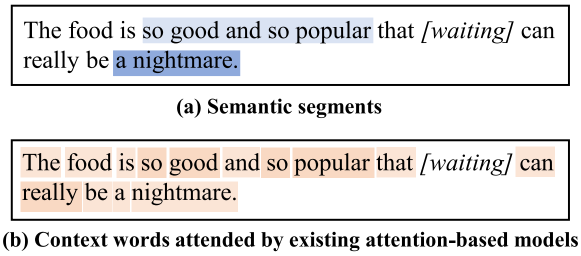
Target-dependent sentiment analysis (TDSA) is an actively studied research topic with the aim to determine the sentiment polarity of a text towards a specific target. For example, given a sentence ''the *food* is so good and so popular that *waiting* can really be a nightmare'', the target-dependent sentiments of *food* and *waiting* are positive and negative, respectively.

The major challenge of TDSA lies in modeling the semantic relatedness between the target and its context sentence (Tang et al., 2016a; Chen et al., 2017). Most recent progress in this area benefits from the attention mechanism, which captures the relevance between the target and every other word in the sentence. Based on such word-level correlations, several models have already been proposed for constructing target-related sentence representations for sentiment prediction (Wang et al., 2016; Tang et al., 2016b; Liu and Zhang, 2017; Yang et al., 2017; Ma et al., 2017).

One important underlying assumption in existing attention-based models is that words can be used as independent semantic units for modeling the context sentence when performing TDSA. This assumption neglects the fact that a sentence is oftentimes composed of multiple semantic segments, where each segment may contain multiple words expressing a certain meaning or sentiment collectively. Furthermore, different semantic segments may even contribute differently to the sentiment of a certain target. Figure 1 shows an example of a restaurant review, which contains two salient semantic segments (highlighted in blue). Intuitively, a TDSA model should be able to identify both segments and determine that the second one is more relevant to the writer's sentiment towards the target *[waiting]*. Existing methods, however, would only attend important words (highlighted in red) such as ''good'', ''popular'', ''really'', and ''nightmare'' individually through the aforementioned assumption.

We hypothesize that the ability to uncover multiple semantic segments and their relatedness with the target from a context sentence will be beneficial for TDSA. In this light, we propose a fine-to-coarse TDSA framework, namely, *Target-Guided Structured Attention Network* (TG-SAN)

Transactions of the Association for Computational Linguistics, vol. 8, pp. 172–182, 2020. https://doi.org/10.1162/tacl_a_00308
Action Editor: Walter Daelemans. Submission batch: 8/2019; Revision batch: 12/2019; Published 4/2020.



The food is so good and so popular that *[waiting]* can really be a nightmare.

(a) Semantic segments

The food is so good and so popular that *[waiting]* can really be a nightmare.

(b) Context words attended by existing attention-based models

Figure 1: A motivating example, where darker shades denote higher contributions to the sentiment of the target *[waiting]*. (a) A TDSA model should be able to identify two salient segments from the sentence, and that the second one is more important for determining the target's sentiment. (b) Existing attention-based models would attend important words individually and fail to determine their relatedness with the target.

in this paper. The core components of TG-SAN include a Structured Context Extraction Unit (SCU) and a Context Fusion Unit (CFU). As opposed to using word-level attention, the SCU utilizes a target-guided structured attention mechanism to encode multiple semantic segments of a sentence as a structured embedding matrix, where each vector in the matrix can be viewed as one target-related context. The CFU then fuses the extracted contexts based on their relatedness with the target to construct the ultimate context representation of the target for sentiment classification.

Our contributions are summarized as follows:

(1) We propose to uncover multiple semantic segments and their relatedness with the target in a sentence for TDSA.

(2) We devise a novel TG-SAN, which uses a fine-to-coarse framework to produce the context representation of the target. TG-SAN utilizes a target-guided structured attention mechanism to encode a sentence as a r-dimensional matrix, where each vector can be viewed as one target-related context. The matrix is further fused into a single context vector by leveraging their relatedness with the target for sentiment classification.

(3) We empirically demonstrate that TG-SAN outperforms a variety of baselines and the state-of-the-art on three benchmarks, and that it is effective in handling sentences composed of multiple semantic segments. We also present visualization results to reveal the superior explanatory power of the proposed model.

2 Related Work

Given a target and its context sentence, the major challenge of TDSA lies in identifying target-related contexts in the sentence for determining the target's sentiment. Early work adopted rule-based methods or statistical methods to solve this problem (Ding et al., 2008; Zhao et al., 2010; Jiang et al., 2011). These methods relied either on handcrafted features, rules, or sentiment lexicons, all of which required massive manual efforts.

In recent years, neural networks have achieved great success in various fields for their strong representation capability. They have also been proven effective in modeling the relatedness between the target and its contexts. Recursive neural networks were first used by Dong et al. (2014) and Nguyen and Shirai (2015) for TDSA. Specifically, the target was first converted into the root node of a parsing tree, and then it contexts were composed based on syntactic relations in the tree. As such approaches rely strongly on dependency parsing, they fall short when analyzing nonstandard texts such as comments and tweets, which are commonly used for sentiment analysis.

Another line of work applied recurrent neural network (RNN) and its extensions to TDSA for their natural way of encoding sentences in a sequential fashion. For instance, Tang et al. (2016a) utilized two RNNs to individually capture the left and the right contexts of the target, and then combined the two contexts for sentiment prediction. Zhang et al. (2016) elaborated on this idea by using a gate to leverage the contributions of the two contexts for sentiment prediction. However, such RNN-based methods place more emphasis on the words near the target while ignoring the distant ones, regardless of whether they are target-related.

Recently, attention mechanisms have become widely used for modeling the relatedness between every context word and the target for TDSA (Wang et al., 2016; Yang et al., 2017; Liu and Zhang, 2017; Ma et al., 2017). For example, Yang et al. (2017) assigned attention scores to each context word according to their relevance to the target, and combined all context words with their attention scores to constitute the context representation of the target for sentiment classification.

The aforementioned attention-based methods used a single attention layer to capture target-related contexts. One drawback of this has been recently examined by Chen et al. (2017) and Li

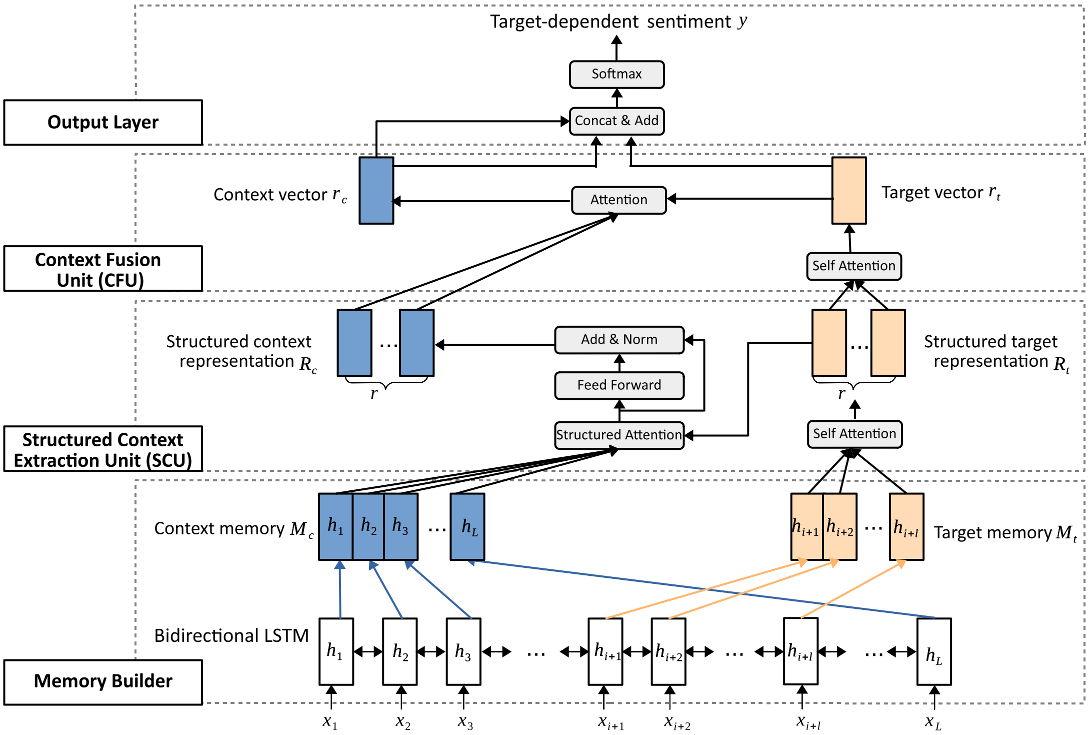

Target-dependent sentiment y
Softmax
Concat & Add
Output Layer
Context vector r_c
Attention
Target vector r_t
Context Fusion Unit (CFU)
Self Attention
Structured context representation R_c
Add & Norm
Feed Forward
Structured Attention
Structured target representation R_t
r
r
Self Attention
Structured Context Extraction Unit (SCU)
Context memory M_c
h_1 h_2 h_3 … h_L
h_{i+1} h_{i+2} … h_{i+l}
Target memory M_t
Bidirectional LSTM
h_1 h_2 h_3 … h_{i+1} h_{i+2} … h_{i+l} … h_L
Memory Builder
x_1 x_2 x_3 x_{i+1} x_{i+2} x_{i+l} x_L

Figure 2: Graphical illustration of TG-SAN. The *Memory Builder* (Section 3.2) takes a sequence of dense word vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_L\}$ as input, and obtains the contextualized word representations $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_i, \ldots, \mathbf{h}_L\}$ via a Bi-LSTM. H is then split into a context memory $\mathbf{M}_c$ and a target memory $\mathbf{M}_t$ based on the positions of the target. The *SCU* (Section 3.3) applies a self-attentive operation on the target memory to obtain a structured target representation $\mathbf{R}_t$, which is used to guide the extraction of r target-related segments $\mathbf{R}_c$ from the context memory through a structured attention mechanism. The *CFU* (Section 3.4) generates the target vector $\mathbf{r}_t$ through a self-attentive operation on $\mathbf{R}_t$, and then learns the contribution of each context to obtain the ultimate context vector $\mathbf{r}_c$. Finally, the *Output Layer* (Section 3.5) composes the context vector and the target vector for predicting the target's sentiment.

et al. (2018). They argued that using one layer of attention to attend all context words may introduce noises and degrade classification accuracy. To alleviate this problem, Chen et al. (2017) proposed refining the attended words in an iterative manner, whereas Li et al. (2018) used a convolutional neural network to extract *n*-gram features whose contributions were decided by their relative positions to the target in the context sentence.

To the best of our knowledge, no existing study has explicitly considered uncovering a sentence's semantic segments and learning their contributions to a target's sentiment. We address this problem with a novel target-guided structured attention network in this work.

3 Approach

We first mathematically formulate the TDSA problem addressed in this paper, and then describe the proposed TG-SAN. Figure 2 depicts the architecture of TG-SAN.

3.1 Problem Formulation

A sentence is a sequence of words $S = \{\mathbf{w}_1, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_L\}$, where $\mathbf{w}_i$ is the one-hot representation of a word and L is the length of the sequence. Given a target, the positions of its mentions in S are denoted by $T = \{i_j^1, \ldots, i_j^t, \ldots, i_j^l\}_{j=1}^m$, where l is the number of word tokens in the target and m is the number of times the target appears in S. $L_t = l * m$ is therefore the total number of word tokens of the target in the sentence. Note that by allowing $m \geq 1$, our problem formulation explicitly models the situation where the target has multiple mentions in a sentence, whereas existing attention-based TDSA models only addressed a single mention situation ($m = 1$).

Given a context sentence S and a target's mentions indexed by T, our task is to predict the sentiment polarity $y \in \mathbf{O}$ of the target, where $\mathbf{O} = \{-1, 0, 1\}$ denote negative, neutral, and positive sentiments, respectively.

3.2 Memory Builder

The Memory Builder constructs the target memory and the context memory from the input sentence as follows. A lookup table $\mathbf{E} \in \mathbb{R}^{d_e \times |V|}$ is first built to represent the semantics of each word by word vectors, where d_e is the dimension of the word vectors and $|V|$ is the vocabulary size. The one-hot representation of the word sequence S is then converted into a sequence of dense word vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_L\}$, where $\mathbf{x}_i = \mathbf{E}\mathbf{w}_i$.

A Bi-LSTM layer is placed on top of the word vectors to obtain their contextualized word representations. The output of this Bi-LSTM layer is a sequence $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_i, \ldots, \mathbf{h}_L\}$, where each hidden state $\mathbf{h}_i \in \mathbb{R}^{2d_h}$ is built by concatenating the outputs of two LSTMs $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$.

$$\begin{aligned}
\overrightarrow{\mathbf{h}_i} &= \overrightarrow{LSTM}\left(\mathbf{x}_i, \overrightarrow{\mathbf{h}_{i-1}}\right) \in \mathbb{R}^{d_h} && (1)\\
\overleftarrow{\mathbf{h}_i} &= \overleftarrow{LSTM}\left(\mathbf{x}_i, \overleftarrow{\mathbf{h}_{i-1}}\right) \in \mathbb{R}^{d_h} && (2)\\
\mathbf{h}_i &= [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}] \in \mathbb{R}^{2d_h} && (3)
\end{aligned}$$

where d_h denotes the dimension of each hidden state.

The sequence $H \in \mathbb{R}^{L \times 2d_h}$ is further split into a target memory $\mathbf{M}_t$ and a context memory $\mathbf{M}_c$ according to the positions of target mentions T. $\mathbf{M}_t \in \mathbb{R}^{L_t \times 2d_h}$ consists of the representations of the target words, while $\mathbf{M}_c \in \mathbb{R}^{L_c \times 2d_h}$ consists of those of the context words, where $L_c = L - L_t$.

3.3 Structured Context Extraction Unit (SCU)

Given the target memory and the context memory, the next step is to extract the target-related segments which may appear in different parts of the context sentence. Recently, Lin et al. (2017) proposed a structured self-attention mechanism, which represents a sentence as multiple semantic segments, and applied such mechanism successfully to document-level sentiment analysis. In TDSA, however, not all semantic segments are related to the target. We therefore build on the idea of Lin et al. (2017) to devise a SCU, which is able to capture target-related segments as the contexts for determining the target's sentiment.

Structured target representation. The target memory $\mathbf{M}_t$ is converted into a structured representation using the self-attentive operation (Lin et al., 2017) as follows:

$$\begin{aligned}
\mathbf{A}_t &= \operatorname{softmax}\left(\mathbf{W}_t^2 \tanh(\mathbf{W}_t^1 {\mathbf{M}_t}^T)\right) && (4)\\
\mathbf{R}_t &= \mathbf{A}_t \mathbf{M}_t && (5)
\end{aligned}$$

where $\mathbf{A}_t \in \mathbb{R}^{r \times L_t}$ is a weight matrix and $\mathbf{R}_t \in \mathbb{R}^{r \times 2d_h}$ is the embedding matrix representing the target. $\mathbf{W}_t^1$ and $\mathbf{W}_t^2$ are two parameters for the self-attentive layer. r is a hyper-parameter referring to the number of rows in the target matrix. In other words, r represents the number of structured representations transformed from the target memory $\mathbf{M}_t$.

Following Lin et al. (2017), a penalization term P is used in the loss function to encourage the diversity of rows captured in $\mathbf{R}_t$.

$$P = \| \left(\mathbf{A}_t {\mathbf{A}_t}^T - \mathbf{I}\right) \|_F^2 \quad (6)$$

Target-guided contexts extraction. Given the target matrix $\mathbf{R}_t$, target-related semantic segments are uncovered from the context memory $\mathbf{M}_c$ as follows. A matrix $\mathbf{A}_c \in \mathbb{R}^{r \times L_c}$ is first built to capture the relatedness between the target matrix and the context memory using a bilinear attention operation. It is then used to build a context matrix $\widetilde{\mathbf{R}}_c \in \mathbb{R}^{r \times 2d_h}$, where each row in the matrix can be viewed as a target-related semantic segment:

$$\begin{aligned}
\mathbf{A}_c &= \operatorname{softmax}\left(\mathbf{R}_t \mathbf{W}_c {\mathbf{M}_c}^T\right) && (7)\\
\widetilde{\mathbf{R}}_c &= \mathbf{A}_c \mathbf{M}_c && (8)
\end{aligned}$$

where $\mathbf{W}_c$ is the parameter of the bilinear attention operation.

A feed-forward network is further placed on top of the context matrix $\widetilde{\mathbf{R}}_c$ to produce its transformed representation $\widehat{\mathbf{R}}_c$. A residual connection (He et al., 2016) is then used to compose both matrices to obtain the final structured context representation $\mathbf{R}_c \in \mathbb{R}^{r \times 2d_h}$.

$$\begin{aligned}
\widehat{\mathbf{R}}_c &= \operatorname{ReLU}(\widetilde{\mathbf{R}}_c \mathbf{W}_s^1 + \mathbf{b}_s^1)\mathbf{W}_s^2 + \mathbf{b}_s^2 && (9)\\
\mathbf{R}_c &= \operatorname{LayerNorm}\left(\widehat{\mathbf{R}}_c + \widetilde{\mathbf{R}}_c\right) && (10)
\end{aligned}$$

where $\mathbf{W}_s^1, \mathbf{b}_s^1, \mathbf{W}_s^2, \mathbf{b}_s^2$ are learnable parameters of the feed-forward network. The layer normalization (Ba et al., 2016) used in Equation (10) helps to prevent gradient vanishing and exploding.

3.4 Context Fusion Unit (CFU)

The CFU learns the contributions of the different extracted contexts to the target's sentiment, and

	Tweet		Laptop		Restaurant	
	training	testing	training	testing	training	testing
# Positive	1561	173	979	340	2158	728
# Negative	1560	173	858	128	800	194
# Neutral	3127	346	454	171	631	196

Table 1: Statistics of the experimental datasets.

produces the ultimate context vector of the target. Specifically, a self-attentive operation is utilized to fuse the target matrix $\mathbf{R}_t$ into a target vector $\mathbf{r}_t$.

$$\begin{aligned} \mathbf{a}_t &= \operatorname{softmax}\left(\mathbf{w}_m^2 \tanh(\mathbf{W}_m^1 \mathbf{R}_t{}^T)\right) && (11) \\ \mathbf{r}_t &= \mathbf{a}_t \mathbf{R}_t && (12) \end{aligned}$$

where $\mathbf{w}_m^2$ and $\mathbf{W}_m^1$ are learnable parameters.

Given the target vector $\mathbf{r}_t$, the contribution of each context is then learned to produce the ultimate context vector $\mathbf{r}_c \in \mathbb{R}^{2d_h}$:

$$\mathbf{r}_c = \sum_{i=1}^{r} \alpha_i \mathbf{R}_c[i] \quad (13)$$

$$\alpha_i = \frac{\exp(\beta_i)}{\sum_{j=1}^{r} \exp(\beta_j)} \quad (14)$$

$$\beta_i = \mathbf{R}_c[i] \mathbf{U} \mathbf{r}_t{}^T \quad (15)$$

where $\mathbf{U}$ is a weight matrix, $\mathbf{R}_c[i] \in \mathbb{R}^{2d_h}$ represents the i-th target-related context and α_i denotes its normalized contribution score.

3.5 Output Layer and Model Training

Consider the examples (a) ‘‘It takes a long time to *boot up*’’, and (b) ‘‘The *battery life* is long’’. Although both targets (in italic) have similar contexts, their sentimental orientations are totally different. It is therefore necessary to consider the target itself along with its contexts to predict its sentiment.

In the output layer, the context vector $\mathbf{r}_c$ and the target vector $\mathbf{r}_t$ are concatenated, and transformed via a non-linear function. The transformed vector is further used in conjunction with $\mathbf{r}_c$ to build the final feature vector $\mathbf{r}_{ct}$:

$$\mathbf{r}_{ct} = \mathbf{r}_c + f(\mathbf{W}_f[\mathbf{r}_c; \mathbf{r}_t]) \quad (16)$$

where $f(\cdot)$ denotes a non-linear activation function, and the ReLU function is adopted in this paper. A $\mathrm{softmax}$ layer is then applied to convert the feature vector into a probability distribution:

$$q(y|\mathbf{r}_{c_t}) = \operatorname{softmax}(\mathbf{W}_q \mathbf{r}_{c_t} + \mathbf{b}_q) \quad (17)$$

where $\mathbf{W}_q \in \mathbb{R}^{\|\mathbf{O}\| \times 2d_h}$ and $\mathbf{b}_q \in \mathbb{R}^{\|\mathbf{O}\|}$ are parameters of the $\mathrm{softmax}$ layer.

For a number of D training instances, cross-entropy loss with a L_2 regularization term is adopted as the loss function:

$$\mathcal{L} = -\sum_{i=1}^{D} y_i \log(q_i) + \lambda_1 \sum_i P_i + \frac{\lambda_2}{2} \|\theta\|_2^2 \quad (18)$$

where y_i is the true sentiment label, $\boldsymbol{q_i}$ is the predicted probability of the true label, θ is the set of parameters of TG-SAN, λ_1 and λ_2 are regularization coefficients, and P_i is the penalization term for the i-th training instance (see Equation (6)).

4 Experiments

4.1 Experimental Setup

Datasets

We evaluate the proposed TG-SAN on three public benchmark datasets, namely, Tweet, Laptop, and Restaurant. The Tweet dataset contains tweets collected from Twitter (Dong et al., 2014). The Laptop, and Restaurant datasets are from the SemEval 2014 challenge (Pontiki et al., 2014), containing customer reviews on laptops and restaurants, respectively. We discarded data instances labeled as ‘‘Conflict’’ in the Laptop and Restaurant datasets following previous studies. Table 1 summarizes statistics of the datasets.

We use classification accuracy and macro-F_1 as evaluation metrics in all experiments.

Compared Models

To demonstrate the ability of the proposed model, we compare it with three baseline approaches, four attention-based models, and the state-of-the-art.

SVM (Kiritchenko et al., 2014): This was a top-performing system in SemEval 2014. It utilized various types of handcrafted features to build a SVM classifier.

AdaRNN (Dong et al., 2014): This utilized a recursive neural network based on dependency tree structure to iteratively compose target-related contexts from a sentence for sentiment classification.

TD-LSTM (Tang et al., 2016a): This employed two LSTMs to separately model the left and the right contexts of a given target, and concatenated their last hidden states to predict the target's sentiment.

ATAE-LSTM (Wang et al., 2016): This used a LSTM layer to model a sentence, and used an attention layer to produce a weighted representation of the sentence with respect to a given target.

IAN (Ma et al., 2017): This used two LSTMs to separately model the sequence of target words and that of context words in a sentence. It then applied an interactive attention mechanism to capture the relatedness between the target and its context for sentiment classification.

MemNet (Tang et al., 2016b): This applied multiple hops of attention on the word embeddings of the context sentence, and treated the output of the last hop as the final representation of the target.

RAM (Chen et al., 2017): This proposed a recurrent neural attention mechanism to iteratively refine the context representation, and took the combination of all constructed contexts as the final representation for sentiment classification.

TNet (Li et al., 2018): It is the state-of-the-art in target-dependent sentiment analysis. It first transformed words considering their positions with respect to the target, and used a convolutional neural network to extract *n*-gram features from the context sentence for sentiment classification. Note that the published results of TNet were based on the authors' implementation with a bug in data preprocessing.[1] We fixed the identified bug, retrained the TNet model with the parameters suggested in the work of Li et al. (2018), and reported the revised results in this paper for empirical comparison.

Experimental Settings

As no standard validation set is available for the benchmark datasets, we randomly held out 20% of the training set as the validation set for tuning the hyper-parameters of TG-SAN. Settings producing the highest validation accuracy are listed in Table 2, and are adopted in the subsequent experiments unless otherwise specified.

Parameter	**Value**
Word embedding dimension d_e	300
LSTM hidden dimension d_h	150
Dropout rate	0.5
No. of structured representations r	2
Penalization term coefficient λ_1	0.1
Regularization term coefficient λ_2	10^{-6}
Batch size	64

Table 2: Hyper-parameter settings of TG-SAN.

We initialized the embedding layer of TG-SAN with the pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014), and fixed the word vectors during the training process. The recurrent weight matrices were initialized with random orthogonal matrices. All other weight matrices were initialized by randomly sampling from the uniform distribution $\mathcal{U}(-0.01, 0.01)$. All bias vectors were initialized to zero. RMSProp was used for network training by setting the learning rate as 0.001 and the decay rate as 0.9. Dropout (Srivastava et al., 2014) and early stopping were adopted to alleviate overfitting. Dropout was applied on the inputs of the Bi-LSTM layer and the output layer with the same dropout rate shown in Table 2.

4.2 Main Results

We report the experimental results of TG-SAN ($r = 2$) and the compared models in Table 3. In summary, TG-SAN outperforms all compared models on the Tweet and the Restaurant datasets. On the Laptop dataset, it also achieves the best accuracy among all models, and macro-F1 comparable to the best-performing model, RAM (Chen et al., 2017). Such results demonstrate the efficacy of the proposed TG-SAN. We also observe that the attention-based models perform better than the baseline models in general. This is not surprising, as different context words can be of different importance to the sentiment of a target, a phenomenon that can be naturally captured by the attention mechanism.

TNet and RAM are the most competitive among all compared models, attributed to their efforts on alleviating the noise produced by using a single layer of attention, as already shown in previous studies. However, we observe that their prediction abilities vary across datasets: RAM performs better than TNet on Laptop and Restaurant, and vice versa on Tweet. In contrast, TG-SAN produces satisfactory performance consistently on

[1] https://github.com/lixin4ever/TNet/issues/4.

Models		Tweet		Laptop		Restaurant	
		Accuracy	Macro-F_1	Accuracy		Accuracy	Macro-F_1
Baselines	SVM (2014)	0.6340$^\sharp$	0.6330$^\sharp$	0.7049*	–	0.8016*	–
	AdaRNN (2014)	0.6630*	0.6590*	–	–	–	–
	TD-LSTM (2016a)	0.6662$^\sharp$	0.6401$^\sharp$	0.7183$^\sharp$	0.6843$^\sharp$	0.7800$^\sharp$	0.6673$^\sharp$
Attention-based	ATAE-LSTM (2016)	–	–	0.6870*	–	0.7720*	–
	IAN (2017)	–	–	0.7210*	–	0.7860*	–
	MemNet (2016b)	0.6850$^\sharp$	0.6691$^\sharp$	0.7033$^\sharp$	0.6409$^\sharp$	0.7816$^\sharp$	0.6583$^\sharp$
	RAM (2017)	0.6936*	0.6730*	0.7449*	**0.7135***	0.8023*	0.7080*
State-of-the-art	TNet (2018)	0.7327	0.7132	0.7465	0.6985	0.8005	0.6901
Proposed Model	TG-SAN	**0.7471**	**0.7365**	**0.7527**	**0.7118**	**0.8166**	**0.7259**
Ablations	w/o CFU	0.7312	0.7141	0.7465	0.7042	0.8095	0.7189
	w/o SCU & CFU	0.7153	0.6975	0.7058	0.6559	0.8023	0.6960
	w/o TG	0.7269	0.7093	0.7324	0.6923	0.8131	0.6986

Table 3: Comparison of Accuracy and Macro-F_1 among different models. Results marked with $\sharp$ are adopted from Chen et al. (2017), and those with $*$ are adopted from the original papers. Performance improvements of the proposed TG-SAN model over the state-of-the-art, TNet (Li et al., 2018), are statistically significant at $p < 0.01$.

all datasets, demonstrating the capability of the proposed fine-to-coarse attention framework in capturing the semantic relatedness between the target and the context sentence for TDSA.

To conclude, we validated the efficacy of TG-SAN through comparative experiments. The advantage of TG-SAN over existing methods confirms our hypothesis that semantic segments are the basic units for understanding target-dependent sentiments. It also shows that such segments can be effectively captured by the proposed target-guided structured attention mechanism.

4.3 Ablation Studies

Three ablation models are designed to reveal the effectiveness of each compoent in TG-SAN.

w/o CFU: This ablation model uses the SCU to capture target-related segments in a sentence, and averages all context vectors to constitute the vector $\mathbf{r}_c$ in Equation (13) without distinguishing their different contributions.

w/o SCU & CFU: In this ablation model, the combination of SCU and CFU is replaced by a simple attention layer. Specifically, the target is represented as the averaged vector of the target memory. It is then utilized to attend the most relevant words in the context sentence to build the context vector. In the output layer, the context vector and the target vector are both composed for sentiment prediction.

w/o TG: In this ablation model, the guidance of the target in the SCU is removed to explore the effect of the target on context extraction. Hence, the SCU is reduced to the one proposed by Lin et al. (2017), which extracts semantic segments from the sentence using the self-attentive mechanism.

Table 3 reports the results of the three ablation models. We observe that performance degrades when the attention layer capturing the contributions of contexts is removed in *w/o CFU*. This indicates that some contexts are indeed more important than the others in deciding the sentiment of a target, and the difference is well captured by CFU. Results also show that the use of SCU is crucial. Comparing *w/o CFU* and *w/o SCU & CFU*, the macro-F1 of the latter drops drastically by 1.66%, 4.83%, and 2.29% on Tweet, Laptop, and Restaurant respectively. Furthermore, results worsened when the target's guidance is replaced with the self-attentive mechanism as in *w/o TG*. This indicates that not all semantic segments appearing in the sentence are related to the target, and it is necessary to extract the related ones for TDSA.

4.4 Effects of r

One important hyper-parameter in TG-SAN is r, which refers to the number of structured representations extracted from the context sentence. We vary the value of r from 1 to 5 to investigate its effects on the TDSA task in this experiment. It

r =	Tweet		Laptop		Restaurant	
	Accuracy	**Macro-F_1**	**Accuracy**	**Macro-F_1**	**Accuracy**	**Macro-F_1**
1	0.7399	0.7261	0.7512	0.6998	0.8131	0.7167
2	**0.7471**	**0.7365**	**0.7527**	**0.7118**	0.8166	0.7259
3	0.7355	0.7210	0.7496	0.7063	0.8184	0.7348
4	0.7399	0.7236	0.7433	0.7028	**0.8220**	**0.7447**
5	0.7327	0.7182	0.7433	0.6972	0.8184	0.7407

Table 4: Effects of r, the number of structured representations extracted from the context sentence. Results show that capturing multiple contexts (r>1) is beneficial for TDSA.

Model	Tweet		Laptop		Restaurant	
	Accuracy	**Macro-F_1**	**Accuracy**	**Macro-F_1**	**Accuracy**	**Macro-F_1**
w/o SCU & CFU	0.6316	0.5250	0.6937	0.6415	0.8097	0.6995
TG-SAN ($r = 1$)	0.6842	0.5667	0.7487	0.6946	0.8230	0.7213
TG-SAN	**0.7368**	**0.6850**	**0.7513**	**0.7114**	**0.8291**	**0.7366**

Table 5: Results on multi-segment sentences, where each sentence contains multiple targets or multiple mentions of the same target. TG-SAN outperforms its degenerated version and the baseline model, showing the advantage of the proposed structured attention mechanism in uncovering multiple target-related contexts.

is worth noting that the attention mechanism of the model degenerates into simple attention when setting r as 1. Table 4 reports the results.

TG-SAN performs best when $r = 2$ on the Tweet and Laptop datasets, and $r = 4$ on the Restaurant dataset. In general, we conclude that the best setting of r is always greater than 1. This demonstrates that multiple contexts are indeed beneficial for predicting target-dependent sentiments, which are well captured by the structured attention mechanism. We also observe that when $r > 1$, model performance may decrease as r increases. The reason might be that a growing r increases the complexity of the model, making it more difficult to train and less generalizable.

4.5 Studies on Multi-segment Sentences

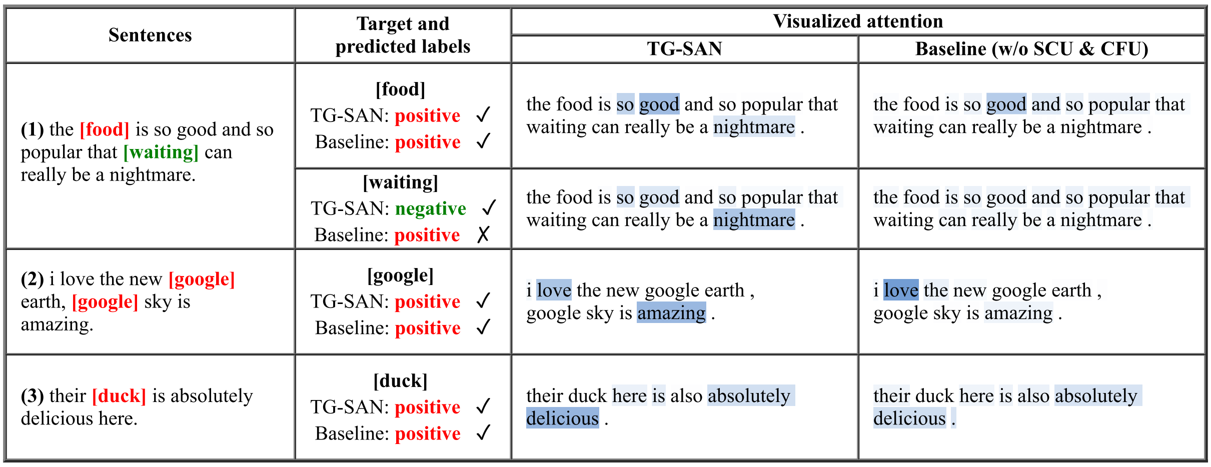
To better understand the advantage of structured attention in TDSA, we further examine a specific group of instances containing multiple semantic segments. Specifically, each instance considered in this experiment either contains multiple different targets, or multiple mentions of the same target. We identified in total 38, 382, and 825 such instances from the Tweet, Laptop, and Restaurant datasets, respectively. It is worth noting that multi-segment instances are particularly common in Laptop and Restaurant, accounting for 59.78% and 73.79% of all instances, respectively.

In this experiment, we compare TG-SAN with two models relying on a simple attention mechanism. One is its degenerated version with $r = 1$, and the other is a baseline model (w/o SCU & CFU). Table 5 reports the comparative results.

We observe that TG-SAN outperforms the other two models on all datasets. This demonstrates that the structured attention mechanism provides a richer context representation ability to identify the target-related contexts more effectively, which is in line with our motivation.

4.6 Case Studies

We demonstrate through case studies that TG-SAN produces not only superior classification performances, but also highly interpretable results. Figure 3 presents test instances covering three different situations: (1) multiple targets, multiple segments; (2) single target, multiple segments; and (3) single target, single segment. For each instance, we plot a heat map to visualize the attention results produced by TG-SAN and a baseline model (w/o SCU & CFU) for comparison. Note that the attention score of each word in TG-SAN is produced by the product of the context weights $\boldsymbol{\alpha} \in \mathbb{R}^r$ (see



Sentences	Target and predicted labels	Visualized attention	
		TG-SAN	Baseline (w/o SCU & CFU)
(1) the **[food]** is so good and so popular that **[waiting]** can really be a nightmare.	**[food]** TG-SAN: **positive** ✓ Baseline: **positive** ✓	the food is so good and so popular that waiting can really be a nightmare .	the food is so good and so popular that waiting can really be a nightmare .
	[waiting] TG-SAN: **negative** ✓ Baseline: **positive** ✗	the food is so good and so popular that waiting can really be a nightmare .	the food is so good and so popular that waiting can really be a nightmare .
(2) i love the new **[google]** earth, **[google]** sky is amazing.	**[google]** TG-SAN: **positive** ✓ Baseline: **positive** ✓	i love the new google earth , google sky is amazing .	i love the new google earth , google sky is amazing .
(3) their **[duck]** is absolutely delicious here.	**[duck]** TG-SAN: **positive** ✓ Baseline: **positive** ✓	their duck here is also absolutely delicious .	their duck here is also absolutely delicious .

Figure 3: Visualization results (best viewed in color). Targets are shown in square brackets. Positive and negative sentiments are highlighted in red and green respectively. In the visualized attention results, the darker the shading of a word, the higher the attention weight it receives from the corresponding model. In general, TG-SAN demonstrates a stronger interpretability than the baseline model. It effectively uncovers all sentiment-related contexts in each case, and identifies the most important ones with respect to a specific target. In contrast, contexts captured by the baseline model are incomplete and inaccurate, as can be seen obviously from the attention results it generates for ''waiting'' in sentence (1) and ''google'' in sentence (2).

Equation (14)) and the word contributions of each context $\mathbf{A}_c \in \mathbb{R}^{r \times L_c}$ (see Equation (7)), denoted by $\boldsymbol{\alpha}^T \mathbf{A}_c$.

Visualization results show that TG-SAN has a strong ability in uncovering semantic segments in a sentence. It can also effectively identify the relatedness between a segment and a certain target. For example, sentence (1) contains two segments expressing opposite sentiments towards the targets ''food'' and ''waiting''. TG-SAN identifies both segments, and places more emphasis on the segment ''so good'' (respectively, ''nightmare'') when predicting the sentiment of ''food'' (respectively, ''waiting''). In contrast, whereas the baseline model identifies all sentiment-related words, it fails to determine accurately the relatedness between each word and the target. As a result, it produces a wrong sentiment prediction for ''waiting''. Similar observations can be made from sentence (2). In this sentence, TG-SAN explicitly captures two target-related segments, whereas the baseline model identifies only one. In case (3), we observe that even when a context sentence contains only one target-related segment, TG-SAN still produces a reasonable explanation for its prediction.

5 Conclusions and Future Work

In this paper, we develop a novel *Target-Guided Structured Attention Network* (TG-SAN) for target-dependent sentiment analysis (TDSA). As opposed to the simple word-level attention mechanism used by existing models, TG-SAN uses a fine-to-coarse attention framework to uncover multiple target-related contexts and then fuse them based on their relatedness with the target for sentiment classification. The effectiveness of TG-SAN is validated through comprehensive experiments on three public benchmark datasets. It also demonstrates superior ability in handling multi-segment sentences, which contain multiple targets or multiple mentions of the same target. In addition, the attention results it produces are highly interpretable as visualization results shown.

As future work, we may extend this study in two directions. First, the SCU is currently utilized once to extract target-related contexts from a sentence, but extending such fine-to-coarse framework through iterative use of multiple SCUs is also feasible from the model perspective. Second, we would like to explore the effectiveness of our model in other tasks where semantic relatedness plays an important role as in TDSA, such as the answer sentence selection task for question-answering.

Acknowledgments

We would like to thank all reviewers and the action editor for their constructive suggestions and comments. This work was supported in part by the Enterprise Support Scheme (ESS) of the Hong Kong Innovation and Technology Fund (No. B/E022/18). Any opinions, findings, conclusions or recommendations expressed in this paper do not reflect the views of the Government of the Hong Kong Special Administrative Region, the Innovation and Technology Commission, or the ESS Assessment Panel.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450v1*.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 452–461.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 49–54.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *International Conference on Learning Representations 2017*.

Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 572–577.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2509–2514.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 214–224.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 606–615.

Min Yang, Wenting Tu, Jingxuan Wang, Fei Xu, and Xiaojun Chen. 2017. Attention based LSTM for target dependent sentiment classification. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 5013–5014.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 3087–3093.

Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010*, pages 56–65.