Improving Sentence Classification by Multilingual Data Augmentation and Consensus Learning

Yanfei Wang[†], **Yangdong Chen**[†], **Yuejie Zhang**[*]
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai 200433, China
{17210240046, 19110240010, yjzhang}@fudan.edu.cn

Abstract

Neural network based models have achieved impressive results on the sentence classification task. However, most of previous work focuses on designing more sophisticated network or effective learning paradigms on monolingual data, which often suffers from insufficient discriminative knowledge for classification. In this paper, we investigate to improve sentence classification by multilingual data augmentation and consensus learning. Comparing to previous methods, our model can make use of multilingual data generated by machine translation and mine their language-share and language-specific knowledge for better representation and classification. We evaluate our model using English (i.e., source language) and Chinese (i.e., target language) data on several sentence classification tasks. Very positive classification performance can be achieved by our proposed model.

1 Introduction

Sentence classification is a task of assigning sentences to predefined categories, which has been widely explored in past decades. It requires modeling, representing and mining a degree of semantic comprehension, which are mainly based on the structure or sentiment of sentences. This task is important for many practical applications, such as product recommendation (Dong et al., 2013), public opinion detection (Pang et al., 2008), and human-machine interaction (Clavel and Callejas, 2015), etc.

Recently, deep learning has achieved state-of-the-art results across a range of Computer Vision (CV) (Krizhevsky et al., 2012), Speech Recognition (Graves et al., 2013), and Natural Language Processing tasks (NLP) (Kalchbrenner et al., 2014a). Especially, Convolutional Neural Network (CNN) has gained great success in sentence modelling. However, training deep models requires a great diversity of data so that more discriminative patterns can be mined for better prediction. Most existing work on sentence classification focuses on learning better representation for a sentence given limited training data (i.e., *source language*), which resorts to design a sophisticated network architecture or learning paradigm, such as attention model (Yang et al., 2016), multi-task learning (Liu et al., 2016), adversarial training (Liu et al., 2017), etc. Inspired by recent advances in Machine Translation (MT) (Wu et al., 2016), we can perform an input data augmentation by making use of multilingual data (i.e., *target language*) generated by machine translation for sentence classification tasks. Such generated new language data can be used as the auxiliary information, and provide the additional knowledge for learning a robust sentence representation. In order to effectively exploit multilingual data, we further propose a novel deep consensus learning framework to mine the language-share and language-specific knowledge for sentence classification. Since the machine translation model can be pre-trained off-the-shelf with great generalization ability, it is worth noting that we do not directly introduce other language data comparing to existing methods in the training and testing phase.

Our main contributions are of two-folds: 1) We first propose utilizing multilingual data augmentation to assist sentence classification, which can provide more beneficial auxiliary knowledge for sentence

[†]: Equal contribution
[*]: Corresponding author

modeling; 2) A novel deep consensus learning framework is constructed to fuse multilingual data and learn their language-share and language-specific knowledge for sentence classification. In this work, we use English as our source language and Chinese/Dutch as the target language from an English-Chinese/Dutch translator. The related experimental results s how that our model can achieve very promising performance on several sentence classification tasks.

2 Related Work

2.1 Sentence Classification

Sentence classification is a well-studied research area in NLP. Various approaches have been proposed in last a few decades (Tong and Koller, 2001; Fernández-Delgado et al., 2014). Among them, Deep Neural Network (DNN) based models have shown very good results for several tasks in NLP, and such methods become increasing popular for sentence classification. Various neural networks are proposed to learn better sentence representation for classification. An influential one is the work of Kim (2014), where a simple Convolutional Neural Network (CNN) with a single layer of convolution was used for feature extraction. Following this work, Zhang and LeCun (2015) used CNNs for text classification with character-level features provided by a fully connected DNN. Liu et al. (2016) used a multi-tasking learning framework to learn multiple related tasks together for sentence classification task. Based on Recurrent Neural Network (RNN), they utilized three different mechanisms of sharing information to model text. In practice, they used Long Short-Term Memory Network (LSTM) to address the issue of learning long-term dependencies. Lai et al. (2015) proposed a Recurrent Convolutional Neural Network (RCNN) model for text classification, which applied a recurrent structure to capture contextual information and employed a max-pooling layer to capture the key components in texts. Jiang et al. (2018) proposed a text classification model based on deep belief network and softmax regression. In their model, a deep belief network was introduced to solve the sparse high-dimensional matrix computation problem of text data. They then used softmax regression to classify the text. Yang et al. (2016) used Hierarchical Attention Network (HAN) for document classification in their model, where a hierarchical structure was introduced to mirror the hierarchical structure of documents, and two levels of attention mechanisms were applied both at the word and sentence level.

Another direction of solutions for sentence classification is to use more effective learning paradigms. Yogatama et al. (2017) combined Generative Adversarial Networks (GAN) with RNN for text classification. Billal et al. (2017) solved the problem of multi-label text classification in semi-supervised learning manner. Liu et al. (2017) proposed a multi-task adversarial representation learning method for text classification. Zhang et al. (2018a) attempted to learn structured representation of text via deep reinforcement learning. They tried to learn sentence representation by discovering optimized structures automatically and demonstrated two attempts of Information Distilled LSTM (ID-LSTM) and Hierarchically Structured LSTM (HS-LSTM) to build structured representation.

However, these tasks do not take into account the auxiliary language information corresponding to the source language. This auxiliary language can provide the additional knowledge to learn more accurate sentence representation.

2.2 Deep Consensus Learning

Existing sentence classification works (Kim, 2014; Zhang and LeCun, 2015; Lai et al., 2015; Jiang et al., 2018; Yogatama et al., 2017; Billal et al., 2017; Zhang et al., 2018a) mainly focus on feature representation or learning a structured representation (Zhang et al., 2018a). Deep learning based sentence classification models have obtained impressive performance. Those approaches are largely due to the powerful automatic learning and representation capacities of deep models, which benefit from big labelled training data and the establishment of large-scale sentence/document datasets (Yogatama et al., 2017; Billal et al., 2017; Zhang et al., 2018a). However, all of the existing methods usually consider only one type of language information by a standard single language process. Such methods not only ignore the potentially useful information of other different languages, but also lose the opportunity of mining the correlated complementary advantages across different languages. A similar model is [20], which used synthetic source sentences to improve the performance of Neural Machine Translation (NMT). While sharing the high-level multilingual feature learning spirit, the proposed consensus learning model significantly has the following three outstanding characteristics. (1) Beyond the language concatenation based on fusion, our model uniquely considers a synergistic cross-language interaction learning and regularization by consensus propagation. This aims to overcome the challenge of learning discrepancy in multilingual feature optimization. (2) Instead of the traditional single loss design, a multi-loss concurrent supervision mechanism is deployed by our model. This enforces and improves the model's individuality learning power of language-specific feature. (3) Through NMT, we can eliminate some of the ambiguous words and highlight some key words.

3 Methodology

We aim to learn a deep feature representation model for sentence classification based on language-specific input, without any specific feature transformation. Figure 1 depicts our proposed framework, which consists of two stages. The first stage performs multilingual data augmentation from an off-the-shelf machine translator; and the second one feeds the source language data and generated target language data to our deep consensus learning model for sentence classification.

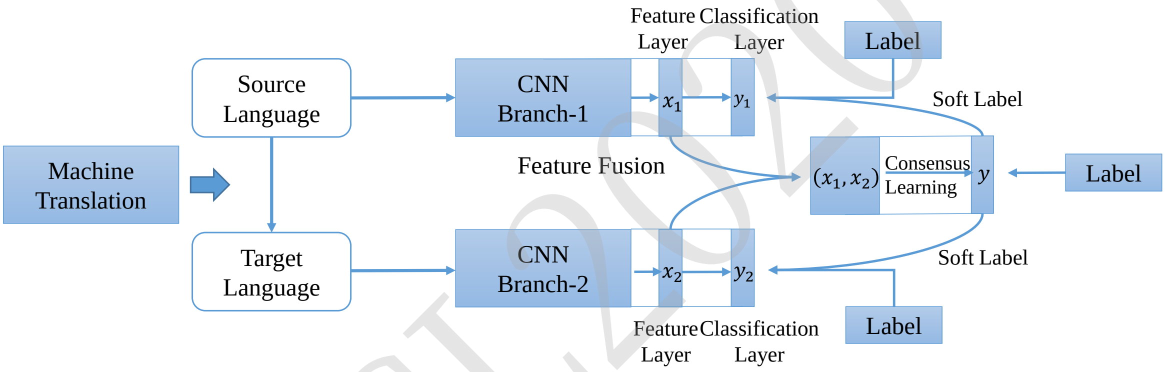

Figure 1. The framework of our proposed model for sentence classification.

3.1 Multilingual Data Augmentation

Data augmentation is a very important technique in machine learning that allows building better models. It has been successfully used for many tasks in areas of CV and NLP, such as image recognition (Krizhevsky et al., 2012) and MT (Zhang et al., 2018a). In MT, Back-translation is a common data augmentation method (Sennrich et al., 2016; Zhang et al., 2018b), which allows us to combine monolingual training data. Especially when the existing data is insufficient to learn a discriminative representation for a specific task, the data augmentation methods can be used.

In sentence classification, given an input sentence in one language, we perform data augmentation by translating the sentence to another language using existing machine translation methods. We name the input language as *source language* and the translated language as *target language*. This motivation comes from the recent great advance in NMT (Wu et al., 2016). Given an input sentence in source language, we simply call the *Google* Translation API [0] to get the translated data in target language. Comparing to other state-of-art NMT models, the *Google* translator has the advantage of both effectiveness and efficiency in real application scenarios. Since target language is used for multilingual data augmentation and the type of it is not important to the proposed model, we random choose Chinese and Dutch respectively as the target language for multilingual data augmentation, and the source language depends on the language of input sentence.

[0] *https://cloud.google.com/translate/*

3.2 Deep Consensus Learning Model

Learning a consensus classification model with the combination of several beneficial information into one final prediction can lead to a more accurate result (Chen et al., 2017). Thus we use two languages of data, $\{S_1, S_2, S_3, \cdots, S_{N-1}, S_N\}$ and $\{T_1, T_2, T_3, \cdots, T_{N-1}, T_N\}$, to perform consensus learning for sentence classification. As shown in Figure 1, our model has three parts: (1) Two branches of language-specific subnetworks for learning the most discriminative features for each language data; (2) One fusion branch responsible for learning the language-share representation with the optimal integration of two kinds of language-specific knowledge; and (3) Consensus propagation for the feature regularization and learning optimization. The design of architecture components will be described in detail as below.

Language-specific Network We utilize the *TextCNN* architecture (Kim, 2014) for each branch of language-specific network, which has been proved to be very effective for sentence classification. *TextCNN* can be divided into two stages, that is, one with convolution layers for feature learning, and another with full connected layers for classification. Given training labels of input sentence, the Softmax classification loss function is used to optimize the category discrimination. Formally, given a corpus of sentences of source language $\{S_1, S_2, S_3, \cdots, S_{N-1}, S_N\}$, the training loss on a batch of n sentences can be computed as:

$$L_{S\_brch} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\exp\left(w_{y_i}^{T} S_i\right)}{\sum_{k=1}^{c}\exp\left(w_k^{T} S_i\right)}\right) \tag{1}$$

where c is the number of categories of sentences; y_i denotes the category label of the sentence S_i; and w is the prediction function parameter of the training category class k. The training loss for target language branch $L_(T\_brch)$ can be computed in the same manner. Meanwhile, since the source language and target language belong to different language spaces, such two branches of language-specific networks are trained with the uniform architecture but different parameters.

Language-share Network We perform the language-share feature learning from two language-specific branches. For this purpose, we firstly perform the language-share learning by fusing across from these two branches. For design simplicity and cost efficiency, we achieve the feature fusion on the feature vectors from the concatenation layer before dropout in *TextCNN* by an operation of *Concat→FC→Dropout→FC→Softmax*. This produces a category prediction score for input pair (a sentence in source language and its translated one in target language). We similarly utilize the Softmax classification loss L_ST for the language-share classification learning as that in the language-specific branches.

Consensus Propagation Inspired by the teacher-student learning approach, we propose to regularize the language-specific learning by consensus feedback from the language-share network. More specifically, we utilize the consensus probability $P_{ST} = \left[p_{ST}^{1}, p_{ST}^{2}, \cdots, p_{ST}^{c-1}, p_{ST}^{c}\right]$ from the language-share network as the *teacher* signal (called "*soft label*" versus the ground-truth one-hot "*hard label*") to guide the learning process of all language-specific branches (*student*) concurrently by an additional regularization, which can be formulated in a cross-entropy manner as:

$$\mathcal{H}_S = -\frac{1}{c}\sum_{i=1}^{c}\left(p_{ST}^{i}\ln\left(p_s^{i}\right) + \left(1 - p_{ST}^{i}\right)\ln\left(1 - p_s^{i}\right)\right) \tag{2}$$

where $P_S = [p_S^1, p_S^2, p_S^3, \cdots, p_S^{c-1}, p_S^c]$ defines the probability prediction over all c sentence classes by the source language branch. Thus the final loss function for the language-specific network can be redefined via enforcing an additional regularization in Eq. (1).

$$L_S = L_{S\_brch} + \lambda\mathcal{H}_S \tag{3}$$

where λ controls the importance tradeoff between two terms. The regularization terms $\mathcal{H}_T$ and L_T for target language branch can be computed in the same way.

The training of our proposed model proceeds in two stages. First, we rely on training the language-specific network separately, which is terminated by the early stopping strategy. Afterwards, the language-share network and consensus propagation loss are introduced. We use the whole loss defined in Eq.

(3) and L_{ST} to train the language-specific network and language-share network at the same time. In the testing time, given an input sentence and its translated sentence, the final prediction is obtained by averaging the three prediction scores from the language-specific networks and the language-share network.

4 Experiment and Analysis

In this section, we investigate the empirical performance of our proposed architecture on five benchmark datasets for sentence classification.

4.1 Datasets and Experimental Setup

The sentence classification datasets include:

(1) ***MR***: This dataset includes movie reviews with one sentence per review, in which the classification involves detecting positive/negative reviews (Pang and Lee, 2005).

(2) ***CR***: This dataset contains annotated customer reviews of 5 products, and the target is to predict positive/negative reviews (Hu and Liu, 2004).

(3) ***Subj***: This dataset is a subjectivity dataset, which includes subjective or objective sentiments (Pang and Lee, 2004).

(4) ***TREC***: This dataset focuses on the question classification task that involves 6 question types (Li and Roth, 2002).

(5) ***SST*-1**: This dataset is Stanford Sentiment Treebank, an extension of *MR*, which contains training/development/testing splits and fine-grained labels (very positive, positive, neutral, negative, very negative) (Socher et al., 2013).

Similar with (Kim, 2014), the initialized word vectors for source language are obtained from the publicly available *word2vec* vectors that were trained on 100 billion words from *Google News*. For target language of Chinese, we retrain the *word2vec* models on *Chinese Wikipedia Corpus*; and for target language of Dutch, we retrain the *word2vec* models on *Dutch Wikipedia Corpus*. In our experiments, we choose the *CNN-multichannel* model variant of *TextCNN* because of its better performance.

4.2 Ablation Study

We first compare our proposed model with several baseline models for sentence classification. Here, we use ***S+T*** to indicate that the model's input contains the source language and the target language. ***T***(*) indicates the type of target language, i.e., ***T(CH)*** indicates that the target language is Chinese, and ***T(DU)*** indicates that the target language is Dutch. Figure 2 and 3 show the comparison results of classification accuracy rate on five benchmark datasets. *CNN(S)* denotes the *CNN-multichannel* model variant of *TextCNN*, which only uses the source language data of English for training and testing. *CNN(T)* is a retrained *TextCNN* model on the translated target language data of Chinese(CH)/Dutch(DU), and the other settings keep the same as *CNN(S)*. *Ours(S+T(*))* denotes our model by combining multilingual data augmentation with deep consensus learning. We can find that *Ours(S+T(*))* performs much better than those baselines, which proves the effectiveness of our framework. It is obvious that multilingual data augmentation can provide the beneficial additional discrimination for learning a robust sentence representation for classification. It is worth noting that *CNN(T)* is even better than *CNN(S)* on *MR*. This indicates that existing machine translation methods can not only keep the discriminative semantics of source language, but also create useful discrimination in target language space.

Similar to *TextCNN*, we also use several variants of the model to demonstrate the effectiveness of our model. As we know, when lacking a large supervised training set, we usually use word vectors obtained from unsupervised neural language models to initialize word vectors for performance improvement. Thus we use various word vector initialization methods to validate the model.

The different word vector initialization methods include:

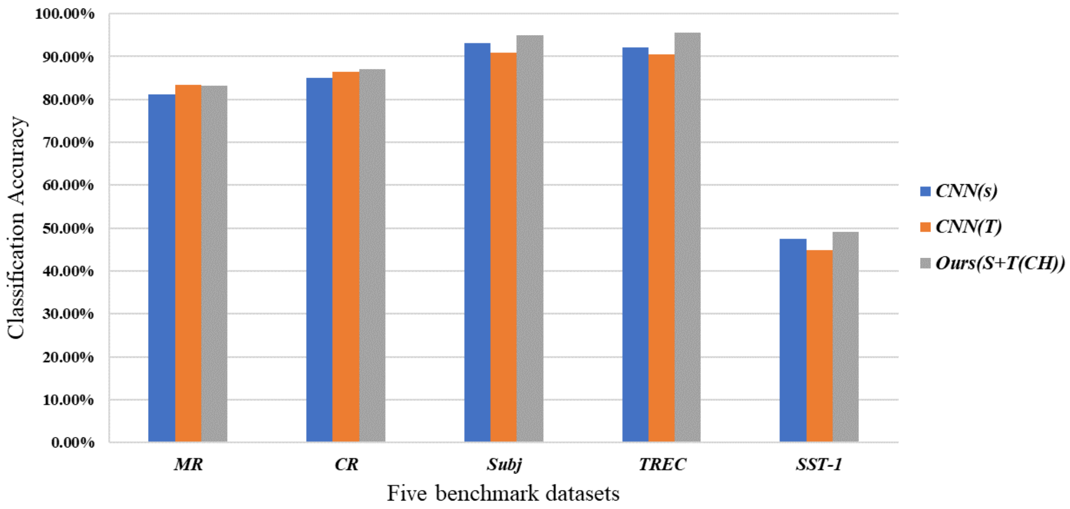

Figure 2. The comparison results with existing baseline models based on English→ Chinese MT.

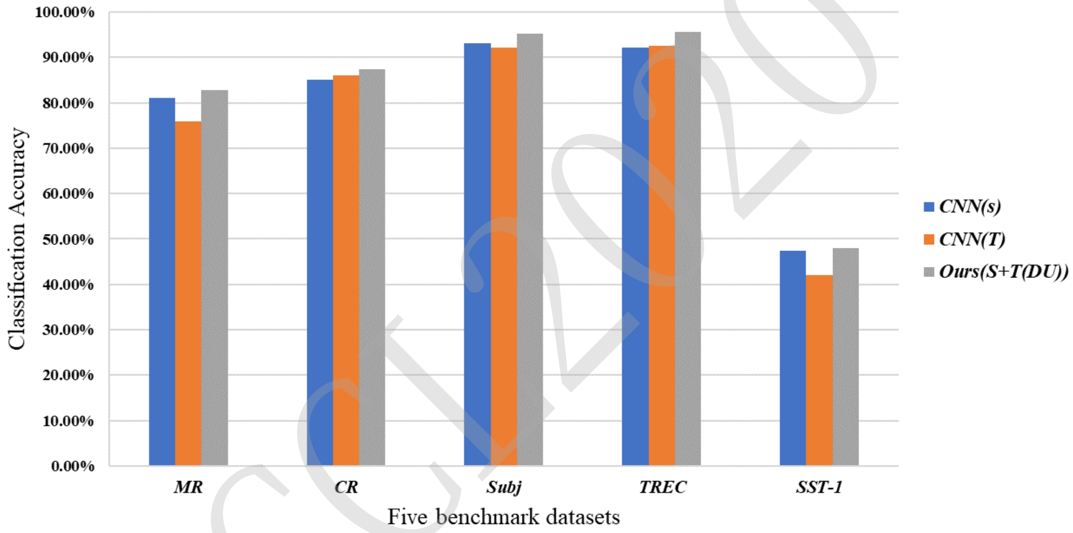

Figure 3. The comparison results with existing baseline models based on English→ Dutch MT.

(1) **Rand**: All words are randomly initialized and can be trained during training.

(2) **Static**: All words of input language are initialized by pre-trained vectors from the corresponding language *word2vec*. Simultaneously, all these words are kept static during training.

(3) **Non-static**: This is an initialization method same to Static, but the pre-trained vectors can be fine-tuned during training.

(4) **Multichannel**: This model contains two types of word vector, which are treated as different channels. One type of word vector can be finetuned during training, while the other keeps static. Two types of word vector are initialized with the same word embedding form *word2vec*.

In Table 1, we show the experimental results of different model variants based on English→ Chinese MT. Compared to the source language *S*, the accuracy rates of the target language *T(CH)* classification are partly improved or decreased, which shows the strong dataset dependency. Considering that the proposed *S+T(CH)* model with Multichannel obtains the current optimal results, we choose the model with Multichannel as our final results. Similar to Table 1, we show the experimental results of different model variants based on English→ Dutch MT in Table 2. Combining the experimental results in Tables 1 and 2, we have enough reasons to prove the validity of our consensus learning method.

Evaluation Pattern	Model Variant	Benchmark Dataset				
		MR	*CR*	*Subj*	*TREC*	*SST*-1
S	**Rand**	76.1%	79.8%	89.6%	91.2%	45.0%
	Static	81.0%	84.7%	93.0%	92.8%	45.5%
	Non-static	81.5%	84.3%	93.4%	93.6%	48.0%
	Multichannel	81.1%	85.0%	93.2%	92.2%	47.4%
T(CH)	**Rand**	79.5%	79.8%	88.5%	85.4%	42.5%
	Static	83.0%	81.4%	89.8%	89.4%	43.6%
	Non-static	82.5%	86.4%	90.1%	90.4%	42.9%
	Multichannel	83.4%	86.4%	90.9%	90.4%	44.8%
S+T(CH)	**Rand**	79.7%	77.2%	92.0%	92.4%	47.1%
	Static	81.8%	86.4%	93.6%	95.0%	47.6%
	Non-static	81.7%	**87.9%**	94.5%	**95.2%**	48.0%
	Multichannel	**83.2%**	87.1%	**95.0%**	**95.6%**	**49.1%**

Table 1. The experimental results of different model variants based on English→ Chinese MT.

Evaluation Pattern	Model Variant	Benchmark Dataset				
		MR	*CR*	*Subj*	*TREC*	*SST*-1
T(DU)	**Rand**	66.5%	78.5%	85.3%	84.8%	35.3%
	Static	75.0%	82.1%	91.6%	89.0%	40.8%
	Non-static	76.6%	86.6%	92.8%	93.0%	42.9%
	Multichannel	76.0%	86.1%	92.1%	92.6%	42.0%
S+T(DU)	**Rand**	76.1%	87.1%	89.5%	90.8%	42.6%
	Static	81.6%	85.6%	93.4%	94.8%	46.2%
	Non-static	81.8%	84.0%	93.9%	**95.6%**	46.8%
	Multichannel	**82.8%**	**87.3%**	**95.3%**	**95.6%**	**47.9%**

Table 2. The experimental results of different model variants based on English→ Dutch MT.

4.3 Comparison with Existing Approaches

To further exhibit the effectiveness of our model, we compare our approach with several state-of-the-art approaches, including recent LSTM-based models and CNN-based models. As shown in Table 3, it can be concluded that our approach can gain very promising results comparing to these methods. The whole performance is measured by the accuracy rate for sentence classification. We roughly divide the existing approaches into four categories. The first category is the RNN-based model, in which Standard-RNN refers to Standard Recursive Neural Network (Socher et al., 2013), MV-RNN is Matrix-Vector Recursive Neural Network (Socher et al., 2012), RNTN denotes Recursive Neural Tensor Network (Socher et al., 2013), and DRNN represents Deep Recursive Neural Network (Irsoy and Cardie, 2014). The second category is the LSTM-based model, in which bi-LSTM stands for Bidirectional LSTM (Tai et al., 2015), SA-LSTM means Sequence Autoencoder LSTM (Dai and Le, 2015), Tree-LSTM is Tree-Structured LSTM (Tai et al., 2015), and Standard-LSTM represents Standard LSTM Network (Tai et al., 2015). The CNN-based model is the third category, in which DCNN denotes Dynamic Convolutional Neural Network (Kalchbrenner et al., 2014b), CNN-Multichannel is Convolutional Neural Network with Multichannel (Kim, 2014), MVCNN refers to Multichannel Variable-Size Convolution Neural Network (Yin and Schütze, 2015), Dep-CNN denotes Dependency-based Convolutional Neural Network (Ma et al., 2015), MGNC-CNN stands for Multi-Group Norm Constraint CNN (Zhang et al., 2016b), and DSCNN represents Dependency Sensitive Convolutional Neural Network (Zhang et al., 2016a). The fourth one is based on other methods, in which Combine-skip refers to skip-thought model with the concatenation of the vectors from uni-skip and bi-skip (Kiros et al., 2015), CFSF indicates initializing Convolutional Filters with Semantic Features (Li et al., 2017), and GWS denotes exploiting domain knowledge via Grouped Weight Sharing (Zhang et al., 2017). Especially on *MR*, our model of ***S+T(CH)*** can achieve the best performance by a margin of nearly 5%. This improvement demonstrates that our multilingual data augmentation and consensus learning can make great contributions to such sentence classification task. Through multilingual data augmentation, important words will be retained. The NMT systems can map those ambiguous words in source language to different word units in target language, which can achieve the result of word disambiguation. Essentially, our method can enable CNNs to obtain better discrimination and generalization abilities. To further demonstrate the superiority of our proposed model, we also use English as the source language and Dutch as the target language to evaluate the model of ***S+T(DU)***. On the four benchmark datasets of *MR*, *CR*, *Subj*, and *TREC*, our models of ***S+T(CH)*** and ***S+T(DU)*** have both achieved the best results at present.

Model	**Approach**	**Benchmark Dataset**				
		MR	***CR***	***Subj***	***TREC***	***SST*-1**
RNN-based Model	**Standard-RNN** (Socher et al., 2013)	-	-	-	-	43.2%
	MV-RNN (Socher et al., 2012)	-	-	-	-	44.4%
	RNTN (Socher et al., 2013)	-	-	-	-	45.7%
	DRNN (Irsoy and Cardie, 2014)	-	-	-	-	49.8%
LSTM-based Model	**bi-LSTM** (Tai et al., 2015)	-	-	-	-	49.1%
	SA-LSTM (Dai and Le, 2015)	80.7%	-	-	-	-
	Tree-LSTM (Tai et al., 2015)	-	-	-	-	**51.0%**
	Standard-LSTM (Tai et al., 2015)	-	-	-	-	45.8%
CNN-based Model	**DCNN** (Kalchbrenner et al., 2014b)	-	-	-	93.0%	48.5%
	CNN-Multichannel (Kim, 2014)	81.1%	85.0%	93.2%	85.0%	47.4%
	MVCNN (Yin and Schütze, 2015)	-	-	93.9%	-	49.6%
	Dep-CNN (Ma et al., 2015)	-	-	-	95.4%	49.5%
	MGNC-CNN (Zhang et al., 2016b)	-	-	94.1%	95.5%	-
	DSCNN (Zhang et al., 2016a)	82.2%	-	93.9%	**95.6%**	50.6%
Model based on Other Methods	**Combine-skip** (Kiros et al., 2015)	76.5%	80.1%	93.6%	92.2%	-
	CFSF (Li et al., 2017)	82.1%	86.0%	93.7%	93.7%	-
	GWS (Zhang et al., 2017)	81.9%	84.8%	-	-	-
Our Model	***Ours (S+T(CH))***	**87.6%**	**87.1%**	**95.0%**	**95.6%**	49.1%
	Ours (S+T(DU))	**82.8%**	**87.3%**	**95.3%**	**95.6%**	47.9%

Table 3. The comparison results between the state-of-the-art approaches and ours.

5 Conclusion and Future Work

In this paper, multilingual data augmentation is introduced to further improve sentence classification. A novel deep consensus learning model is established to fuse multilingual data and learn the language-share and language-specific knowledge. The related experimental results demonstrate the effectiveness of our proposed framework. In addition, our method requires no external data comparing to existing methods, which makes it very practical with good generalization abilities in real application scenarios. In the future, we will try to explore the performance of the model on larger sentence/document datasets. The linguistic features of different languages will be also considered when selecting the target language.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61976057, No. 61572140). Yanfei Wang and Yangdong Chen contributed equally to this work, and were co-first authors. Yuejie Zhang was the corresponding author.

References

Belainine Billal, Alexsandro Fonseca, Fatiha Sadat, and Hakim Lounis. 2017. Semi-supervised learning and social media text analysis towards multi-labeling categorization. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1907–1916. IEEE.

Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2590–2600.

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, 7(1):74–93.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Ruihai Dong, Michael P O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. 2013. Sentimental product recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 411–414.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Ozan Irsoy and Claire Cardie. 2014. Deep recursive neural networks for compositionality in language. In *Advances in neural information processing systems*, pages 2096–2104.

Mingyang Jiang, Yanchun Liang, Xiaoyue Feng, Xiaojing Fan, Zhili Pei, Yu Xue, and Renchu Guan. 2018. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, 29(1):61–70.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014a. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014b. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Tao Liu, Renfen Hu, and Xiaoyong Du. 2017. Initializing convolutional filters with semantic features for text classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1884–1889.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. pages 2873–2879.

Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10.

Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. 2015. Dependency-based convolutional neural networks for sentence embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 174–179.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

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

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Wenpeng Yin and Hinrich Schütze. 2015. Multichannel variable-size convolution for sentence classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 204–214.

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

Rui Zhang, Honglak Lee, and Dragomir Radev. 2016a. Dependency sensitive convolutional neural networks for modeling sentences and documents. In *Proceedings of NAACL-HLT*, pages 1512–1521.

Ye Zhang, Stephen Roller, and Byron C Wallace. 2016b. Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1522–1527.

Ye Zhang, Matthew Lease, and Byron C Wallace. 2017. Exploiting domain knowledge via grouped weight sharing with application to text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 155–160.

Tianyang Zhang, Minlie Huang, and Li Zhao. 2018a. Learning structured representation for text classification via reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018b. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.