Influence of Noise on Transfer Learning in Chinese Sentiment Classification using GRU

Mingjun Dai
College of Information Engineering
Shenzhen University
Shenzhen, China

Shansong Huang
College of Information Engineering
Shenzhen University
Shenzhen, China

Junpei Zhong
Artificial Intelligence Research Center
AIST
Aomi, Tokyo, Japan

Chenguang Yang
College of Engineering, Swansea University
Swansea, UK

Shiwei Yang
School of Management, Guangdong University of Technology
Guangzhou, China

***Abstract*—Sentiment classification for product reviews is of great significance for business feedback for manufactures, sellers and users. However, since a large amount of training data for a specific product domain is not always available, transfer learning is often utilized to do sentiment analysis applications. Specifically, after a pre-training of the large Chinese corpus by a word-embedding method, a larger size of training data for a specific domain was trained using a Gated Recurrent Unit. And then the trained model was used for testing the sentiment classification for a smaller amount of product reviews. The performances of this transfer learning method was also examined, especially to testify different factors affecting the performance of the transfer learning. The experimental results showed that different wording in the review domain (which we call it "noise") will have a greater impact on transfer learning. We also calculate the difference of the wording to verify our hypothesis. According to these results, we have explored the impacts of the dataset wording, while we are doing Chinese text sentiment classification. We also shed a light in optimizing the transfer learning effect in general.**

***Index Terms*—sentiment classification, neural network, Gated Recurrent Unit, transfer learning**

I. Introduction

Defined as the process of classifying the sentimental opinions of the customers about specific domains, sentiment analysis provides important resources for the manufacturers and retailers in the e-commence websites to measure the social data. This technique, understanding the quantitative sentimental measurements from the customers towards various products, would help to reveals the buying trends, product flaws and underlying customer sentiment. On the other hand, an automatic quantitative measurement (i.e. grading) of the product reviews can be used as quantitative reference for other users to purchases. Therefore, besides of academia research, modern e-commerce websites, such as Amazon or Taobao, are also keen to develop automatic sentiment classification techniques. The most straightforward input-output mapping for sentiment classification is to confirm the the sentiment polarity (positive, negative, or neutral) of sentence or phrase in the product review text.

Corresponding to the methods of Natural language processing (NLP), there are basically three kinds of sentiment classification methods: lexicon-based, statistical-based and neural-based methods. However, one critical problem of these three methods is that such sentiment classification techniques alone depend on a large number of annotated data-sets, which provides a constraints for practical use of such techniques. Thus, transfer learning (or pre-training) is often utilized while the size of the training samples is not enough. While utilized transfer learning in a correct way, people can train learning models with a larger and more easily obtained data sets from source domains, but doing forecasting with a more difficult obtained data set from target domains [1]. Such techniques have been widely used in practical applications such as commercial image classification, NLP, etc.

In practical projects, e-commerce websites contain multiple categories of products. However, there may be cases where the the number of comments toward specific products is too small to train the model, which will probably result in overfitting during the training process [2]. Therefore, using transfer learning to train the learning model with a larger dataset, and a smaller dataset will be practical and beneficial for the engineers. However, direct transfer of the knowledge domain sometimes is not accurate and efficient.

In this paper, we try to investigate how the performance of transfer learning is applied in Chinese sentiment classification with Gated Recurrent Unit (GRU) and how the properties of the datasets affect the transfer learning results. We choose three different domains of product comments dataset, which have annotated into positive and negative categories.

The organization of this paper is as follows: we will firstly introduce related models in sentiment classification and transfer learning at the next section. Then our learning method based on GRU will be proposed at the third section. At the fourth section, we will demonstrate the experimental data and do analysis on the performance of transfer learning.

Corresponding author: Junpei Zhong, E-mail: joni.zhong@aist.go.jp; Shansong Huang, E-mail: 1357182413@qq.com

II. Related Work

A. Sentiment Classification

In general, studies on sentiment classification adopt various kinds of NLP methods to analyze the text. For instance, corresponding to the lexicon-based NLP, the lexicon-based sentiment classification (e.g. [3]–[6]) uses the vocabulary from an annotated dictionary to calculate the polarity and strength of a sentence or a text. This is based on a dedicated weighting mapping from the wording space to the text space, in which the mapping is learnt by classification methods such as support vector machine (SVM) [7], Naive Bayes and Maximum Entropy model, etc. Hence, this method can calculate the sentiment polarity in a precious way. But since its mapping is dedicated designed for a specific domain, it cannot express the meaning of longer phrases from a different domain in a principled way.

Statistical or neural based sentiment classification gain comparable performance with lexicon-based model in cross-domain training sets while a large number of training samples are available. Most importantly, with a large number of sample to do statistical learning, the hand-crafted annotation is no longer necessary. Although basically they calculate the probability distribution during learning, most of the statistical methods are modelled as an ordinal regression problem (e.g. [8], [9]). While the regression is learning by neural network representation, the learning is more straightforward in terms of its observability. For instance, using recursive auto-encoder [10], the vector space can represent multi-word phrases and can exploits the recursive nature of sentences. Convolutional network (CNN) was also employed to accomplish the semantic role labelling task [11], since it is able to extract the feature of each part of the sentences without a lot of preprocessing. Similar as CNN, recurrent neural networks (RNN) also learn fixed-length vectors for text of varying length, while the word order is strictly kept. Using the RNN as a learning method has been widely used in sentence generation from images [12], sentiment analysis [13] and language modelling [14]. Neural and statistical methods usually do not need to do a lot of manual pre-processing, but they are not as predictability and controllability as lexicon-based methods.

B. Transfer Learning

During the training of machine learning, the training and testing data sets generally have similar characteristics, but in practical applications this is not the case. That is why we have to investigate transfer learning methods. In many real-world situation, it is very expensive or impossible to re-collect the required training data and reconstruct the model. In such cases, knowledge transfers or transfer learning between task domains would be desirable [15]. Basically, traditional machine learning techniques try to learn each task from scratch, while transfer learning techniques try to transfer the knowledge from some specific domains to a target domain when the latter has fewer high-quality training data.

To solve the basic problem of transfer learning: although the source domain data cannot be reused directly, there are certain parts of the data that can still be reused together with a few labelled data in the target domain. To explore this possibility, Daume et al. [16] uses a specific Gaussian model to study the transfer learning issues in different domain. Dai et al. [17] provided a framework called TrAdaBoost, which extended boosting-based learning algorithms [18]. Similarly, instead of boosting methods, other heuristic methods (e.g. [19], [20]) were able to be adopted to learn to reuse data from the source domain maximally.

III. Models

A. Recurrent Neural Network

Recurrent neural network (RNN) is one of artificial neural networks in which the connections between the units form a directional cycle. It can be better approach the sequence issue especially storing temporal information in the loop. But in some cases, the conventional RNN cannot learn the temporal relation while the casual events happen in a long time range because of eliminating of the error in the back-propagation through time (BPTT) learning.

To overcome this so-called "long-term dependency" problem [21], more advanced RNN models employing the gating mechanism have been developed. Among them, the first attempt to solve this problem was done by Hochreiter and Schmidhuber [22], who proposed Long Short Term Memory networks (LSTM). Instead of the single activation function embedded in conventional neural units, a more adaptive gating mechanism is employed. With the gates operations which are also learnt by a large amount of data, LSTM could remember the long-term information. Recently, Chung et al. [23] also proposed Gated Recurrent Unit (GRU) (Fig. 1). GRU optimizes the LSTM' s internal structure and made the gating mechanisms simpler. With the help of such advanced designs, the applications of RNN have spanned in speech recognition, text abstraction [24], text translation, figure captioning and robot control [25], etc. As one of the state-of-the-art RNN, in this paper, we adopt GRU to build neural network model to text the sentiment classification of Chinese documents and examine the transfer learning performance.

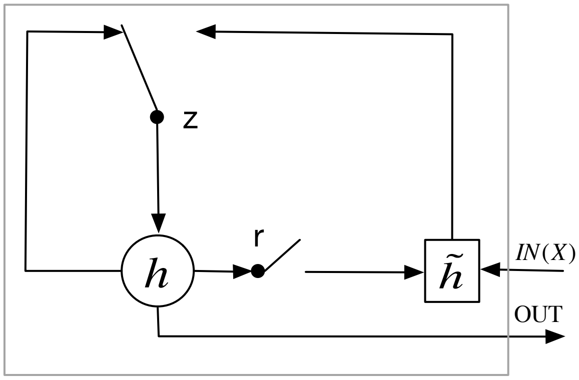

Fig. 1: Illustration of gated recurrent unit (GRU)

B. GRU

GRU contains two gates: the reset gate r and the update gate z. We will bespeak the input continuous time sequence as $X = (x_1, x_2, ..., x_n)$, the each cell of hidden layer as $H = (h_1, h_2, ..., h_n)$. For time step t, GRU first calculates the reset gate r_t, which determines the amount of selection for the previous state, and update gate z_t determines how much the unit update. These gates are computed by the operations below:

$$r_t{=}\sigma(\mathrm{W}_r x_t + U_r h_{t-1}) \tag{1}$$

$$\mathrm{z}_t{=}\sigma(\mathrm{W}_z x_t + U_z h_{t-1}) \tag{2}$$

where σ is sigmoid function, W_* and U_* are weight matrixes The hidden node h_t and the candidate hidden node $\tilde{h_t}$ are computed by

$$\tilde{h}_t = \tanh(Wx_t + U(r_t \odot h_{t-1})) \tag{3}$$

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t \tag{4}$$

where $\odot$ is an Element-Wise Multiplication. See Fig.1 which displays the structure of GRU's memory block.

IV. Method and Data

The main target of our experiment was to use a simple and straightforward transfer learning method in Chinese sentiment classification to accomplish the transfer learning in different product domains. We hereby chose the GRU model as the main classification learning tool. Furthermore, we also used a Chinese word segmentation to do pre-processing and Word2vec [14] to do pre-training on a significantly larger data-set. In terms of the training data, three different domain of Chinese comment datasets, which have annotated with sentiment labels, were used. The flowchart as shown in Fig. 2.

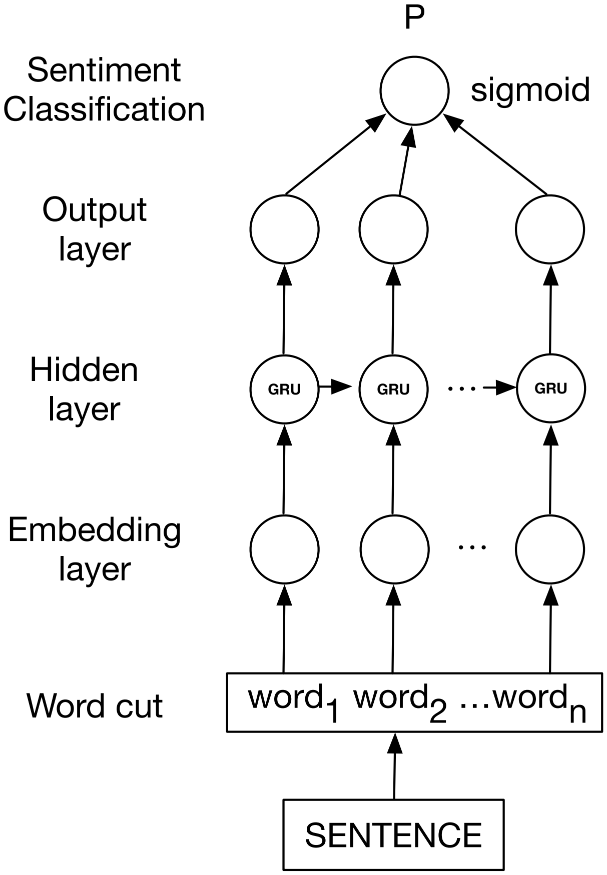

Fig. 2: Neural network for Chinese text sentiment classification

TABLE I: Information In Three Datasets

Dataset	Positive	Negative	Total
Book	2000	2000	4000
Hotel	2000	2000	4000
Laptop	2000	2000	4000

A. Datasets and System

We used the dataset [1]containing Chinese product reviews for training. The dataset consist of three categories of customer reviews in three domains: book, hotel and laptop. Each of the review containing a text-level sentiment polarity tag (positive and negative). The number of reviews in all of the three domains contains 2000 positive, and 2000 negative reviews (TAB. I).

B. Chinese Word Segmentation

The word segmentation is a process of re-assembling a sequence of consecutive words into a word-group according to certain meanings. Different from the English text, there is no clear delimiter between the Chinese words. For instance, the

[1] http://www.nlpir.org/?action-viewnews-itemid-77

sentence:“这个西瓜不大好吃”, when we can cut this sentence as “这个/西瓜/不大/好吃” means this watermelon is not big enough but good to eat, while “这个/西瓜/不大好吃” means this watermelon is not very tasty. Therefore, specifically in this project, we employed a word segmentation tool called Jieba [2] to do word segmentation and separate sentences of Chinese comments into separate words.

C. Word Embedding

Processing natural language by machine learning algorithm usually needs to encode word into representations. Instead of the straightforward one-hot representation, Bengio et.al. [26] proposed a distributed representation of high-dimensional vectors for text representation, termed as word embedding. In this word embedding, the logical meaning of words can be mapped in the embedding space so that words with similar meaning locate close to each other. Also it allows logical operation such as $vec(Queen) = vec(King) - vec(man) + vec(woman)$. By using the language neural network to map the dictionary space into word vector space, word embedding provides a state-of-the-art performance for capturing syntactic and semantic word similarities [27]. The word2vec is one of the word embedding methods by using simple feedforward neural learning.

Word2vec works as a predictive model to learn their co-occurrence vectors by a 3-layer feed-forward model[3]. The model learn the correlation between the target word and the context words in a simple way: it tries to capture the meaningful semantic regularities by this feed-forward training with the activation of the target word and its context words. And the hidden layer, as a “by-product” after training, represents such regularities between pairs of words.

In this project, we selected a python-based tool Gensim[4] to implement word2vec. For training word2vec, the greater size of the corpus, the better performance we get from the word embedding [28]. Therefore, we selected Chinese dataset of wikipedia for pre-training. It is also a crucial part for us to proceed to the transfer learning.

In the process of using word2vec to generate word embedding, we set the dimension of word vector embedding of each word to 128, so that the information is preserved but it did not cause the issue of curse of dimensionality [26]. TAB. II shows the number of different Chinese words covered in every dataset.

TABLE II: Number Of Words In Every Dataset

Dataset	Original	Processed
Book	248342	97758
Hotel	278834	105069
Laptop	117976	96541

V. Experimental Results

After the pre-processing with Jieba and pre-training with word2vec, we trained the GRU with the annotated review data. We used Adam [29] and binary_crossentropy[5] respectively as the training method and the loss function. By controlling the neuron disconnection ratio can effectively prevent overfitting and get superior results [30], thus we set the dropout to 0.3. We used cross-validation to train the datasets with the ratio of $0.7 : 0.3$ between the training and validation sets. We truncated each review and used only the first 30 words in each review. In this particular case, it reduces the training time without affecting much the sentiment polarity. The accuracies of different combination of three datasets are displayed in TAB. III.

TABLE III: Accuracy On Six Different Combination Of Datasets

Training and Testing Set	Train Accuracy	Test Accuracy
Book-Laptop	0.9889	0.5275
Book-Hotel	0.9793	0.5260
Hotel-Laptop	0.9668	0.7717
Hotel-Book	0.9761	0.5695
Laptop- Book	0.9807	0.6337
Laptop- Hotel	0.9879	0.7093

A. Single Domain Training

In the reviews of the product, the expression of the emotions include direct and implicit two ways, which can immediately judge from the comments of consumer emotional tendencies are called direct expression, For example, comment on laptop dataset: “Dell商务机做工不错，外观设计比以前的商务机好很多了。价格又便宜，这不错”(Dell Business notebook quality is good, the design is much better than the previous business notebook, the price is cheap, this is true), it contains a lot of emotional polarity vocabulary, very intuitive to show the consumer’s feelings on the purchase of laptop. However, the need to understand the emotional tendencies of consumers from comments that contain a small amount of emotional polarity, it is called indirect expression, such as comment on book dataset: “不是《庄子》里所有的文章都一定是庄子本人所做，大家要正确看待，读《庄子》要用心去领会，参考别人教的话，顶多是他人消化过的看法，我们要先有一定的理解，再结合各种观点去更好地领会其中的奥妙” (Not *Chuang Tzu* in all the articles must be written by Chuang Tzu himself, we must look at the correct, read *Chuang Tzu* to carefully understand, test others to teach, at most, is the idea of digesting others, we must first have a certain understand, and then combined with a variety of points to better understand the secret), we can not literally get the consumer’s emotions, but need to further semantic analysis to judge. In this paper, we called the comments that require indirect expression as noise. Through the analysis of the three datasets, we found out that the dataset of book contains the most noise, laptop contains the

[2] https://github.com/fxsjy/jieba

[3] The original report of Word2vec claimed that it is a 2-layer network without the input words. Here we regard it as a 3-layer network following the convention of feed-forward network including input vector itself.

[4] https://radimrehurek.com/gensim/models/word2vec.html

[5] https://keras.io/losses/#binary_crossentropy

least noise. In single domain training, by comparing Fig. 3 and 4, we could figure out that the effect of noise on the accuracy of the validation set.

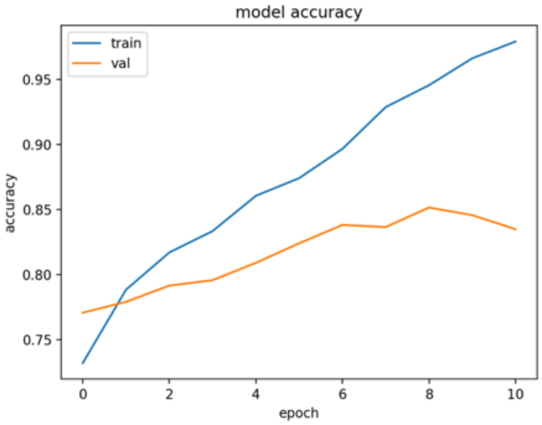

Fig. 3: Accuracies of the training set and the validation set (book)

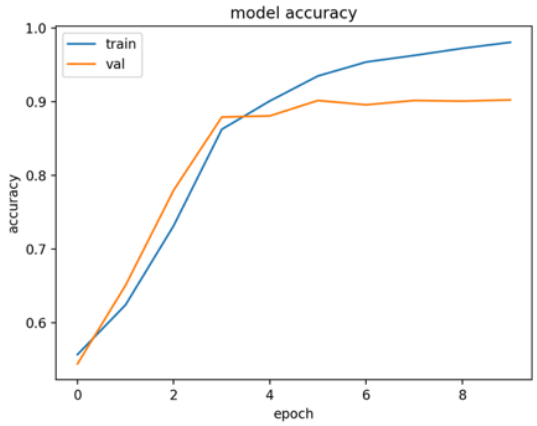

Fig. 4: Accuracies of the training set and the validation set (laptop)

B. Performance and Analysis Transfer Learning

From the experimental results we could figure out that when the book as a training set, the transfer learning performance is the worst, at this time, hotel and laptop as a testing set of accuracy were 52.75% and 52.60%. When the hotel as a training set, laptop and book as a testing set which accuracy were 77.17% and 56.95%. And the laptop as a training set, book and hotel as a testing set which accuracy were 63.37% and 70.93%. As show in TAB. III, it is observed that when the book as a training set, and the other two datasets as a testing set, the results are not very good. And when the book as a testing set, the effect is not good either, but probably contain some of the positive or negative vector, make it still work. Therefore, in the case where the quantity of the datasets and the complexity of the model are similar, when more noise exists in the training set or testing set, it will be large effect on transfer learning.

VI. Conclusions

As one of the promising machine learning techniques, transfer learning has been widely used in many machine learning fields. Focusing on the field of Chinese text sentiment classification based on GRU model, our research discussed how the different distributions in text space affect the performance of transfer learning between different product domains.

At the next stage, we will focus on the optimization of transfer learning in terms of adjusting the following parameters:

- the distribution of training and target domains.
- the parameters of the trained model.

Acknowledgment

MJ and SS were supported by research grant from Natural Science Foundation of China (61301182), from Specialized Research Fund for the Doctoral Program of Higher Education from The Ministry of Education (20134408120004), the Key Project of Department of Education of Guangdong Province (2015KTSCX121), from Foundation of Shenzhen City (KQCX20140509172609163), and from Natural Science Foundation of Shenzhen University (00002501). SW was partially supported by research grant from Natural Science Foundation of China (71673062)

References

[1] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, 2016.

[2] G. P. Zhang, “Neural networks for classification: a survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[3] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[5] X. Ding, B. Liu, and P. S. Yu, “A holistic lexicon-based approach to opinion mining,” in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 231–240.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 115–124.

[9] B. Snyder and R. Barzilay, “Multiple aspect ranking using the good grief algorithm.” in *HLT-NAACL*, 2007, pp. 300–307.

[10] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 151–161.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[13] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification." in *EMNLP*, 2015, pp. 1422–1432.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[16] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[17] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 193–200.

[18] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *Journal of the ACM (JACM)*, vol. 44, no. 3, pp. 427–485, 1997.

[19] P. Wu and T. G. Dietterich, "Improving svm accuracy by training on auxiliary data sources," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 110.

[20] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *ACL*, vol. 7, 2007, pp. 264–271.

[21] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] J. Zhong, A. Cangelosi, and T. Ogata, "Toward abstraction from multimodal data: Empirical studies on multiple time-scale recurrent models," *International Joint Conference on Artificial Neural Networks (IJCNN)*, 2017.

[25] J. Zhong, A. Cangelosi, T. Ogata, and C. Yang, "Understanding natural language sentences with word embedding and multi-modal interaction (submitted)," *Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2017 Joint IEEE International Conferences on*, 2017.

[26] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[27] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[28] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.