An Improved Mechanism for Universal Sentence Representations Learnt from Natural Language Inference Data Using Bi-directional Information

Dian Jiao[1, a], Sheng Gao[b] and Baodong Zhang[2, c]

[1]Beijing University of Posts and Telecommunications, Beijing, CO 100876, China;

[2]Jinan Company in State Grid Corp. of China;

[a]jiaodian@bupt.edu.cn, [b]gaosheng@ bupt.edu.cn, [c]jnzbd@163.com

Abstract. BiLSTM with max pooling is adopted as a well-performed supervised universal sentence encoder. Max pooling is a common mechanism to get a fixed-size sentence representation. But we find that the max pooling for sentence encoder discards some useful backward and forward information at each time step and depends on a large number of parameters. In this paper, we propose an improved pooling mechanism based on max pooling for universal sentence encoder. The proposed model uses three kinds of methods to refine the backward and forward information at each time step, and then use a max-pooling layer or attention mechanism to obtain a fixed-size sentence representation from variable-length refined hidden states. Experiments conducted on Stanford Natural Language Inference (SNLI) Corpus, and we use it as a pretrained universal sentence encoder for transfer tasks. Experiments show that our model with less parameters performs better.

Keywords: Universal sentence encoder, supervised, SNLI, transfer tasks, pooling, attention.

1. Introduction

Distributed word embeddings [14] are widely used in semantic representation, and have an excellent performance to represent words. Similarly, we want to find a semantic representation mechanism to capture the relationship between words and phrases in a single vector. In other words, we want to obtain universal sentence representations.

There are two kinds of methods to train a model for universal sentence representations, supervised and unsupervised. Following unsupervised methods are inspired by word embeddings model, such as SkipThought [9] and FastSent [6]. And facebook proposes a supervised model [5] based on BiLSTM with max pooling and trained to solve natural language inference (NLI) task, inspired by computer vision, the model can be used as a pretrained model and it outperforms former unsupervised models on many transfer tasks.

In this paper, we focus on the supervised model. Sentence encoder is the most important module of it. Many neural networks of different architectures can represent a sentence as a fixed-size vector. Facebook compares 7 different architectures: Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) as a standard recurrent encoders, concatenation of last hidden states of forward pooling, self-attentive network [12] and hierarchical convolutional networks [17]. Bi-directional LSTM with max pooling gains the best performance on NLI task and has a remarkable performance on transfer tasks.

BiLSTM can obtain backward and forward information. Pooling and attention mechanism are feature selection methods to decrease the dimension of sentence to a fixed-size vector. Pooling mechanism generates value of each dimension of BiLSTM hidden units over time steps by some specific strategy, and facebook former experiments show that max pooling performs best. Attention mechanism sums values of every dimension over time steps with different weights, it performs well on specific task but maybe not suit transfer tasks. On transfer tasks, some sentence information may relate to both bidirectional hidden states, moreover each hidden state contains unequal backward or forward information, random noises affect correct selections. So direct selection (max pooling) may lose useful information of a sentence.

In order to overcome this shortcoming, we consider to leverage the fact that forward and backward hidden units at each time step contain the information of the whole sentence altogether. In this paper,

we propose two kinds of BF-max pooling methods to let every dimension obtain balanced and useful global information, one uses a weight matrix to combine forward and backward information named BF-combined max pooling, the other one adjusts the strength of two direction according to their amount of information named BF-balanced max pooling, finally we obtain a fixed-size vector by selecting the maximum value over each dimension. Meanwhile, we propose a method using attention mechanism after the combination of forward and backward information named BF-attention. With refined sentence representations, it uses less parameters to get remarkable performances.

We train a sentence representation model on the Stanford Natural Language Inference (SNLI) dataset [2], and compare it with other models on SNLI task. Moreover, we use our model as a universal sentence encoder on transfer tasks. On SNLI task, our proposed mechanism helps us improve the accuracy of the NLI task classification with less parameters. Furthermore, it gains a better performance on transfer tasks and get state-of-art sentence embeddings compared with FastSent and SkipThought. Experiments on a broad and diverse set of transfer tasks reveal the ability of our model to capture more useful semantic information.

2. Approach

In this section, we introduce the baseline based on bi-directional LSTM with max pooling and the proposed model based on bi-directional LSTM with an improved pooling layer named BF-max pooling.

2.1 BiLSTM with Max Pooling.

The simplest modules applied as the encoders in the sequence-to-sequence tasks [11] are LSTM [7] and GRU [3]. Given a sequence of words T = $\{w_0, w_1, \dots, w_t, w_{t+1}, \dots\}$as the input to the encoder, the encoder computes a set of T hidden states noted as $h_1, \dots, h_t$, while $\overrightarrow{h_t} = \overrightarrow{LSTM}_t(w_1, w_2, w_3, .. w_t)$, usually we use h_t to represent a sentence. But in order to capture more information, BiLSTM is proposed to represent a sentence as equation (1).

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{LSTM}_t(w_1, w_2, w_3, .. w_t) \\ \overleftarrow{\boldsymbol{h_t}} &= \overleftarrow{\boldsymbol{LSTM}}_{\boldsymbol{t}}(\boldsymbol{w_T}, \boldsymbol{w_{T-1}}, \boldsymbol{w_{T-2}}, .. \boldsymbol{w_t}) \\ \boldsymbol{h} &= [\overrightarrow{\boldsymbol{h_t}}, \overleftarrow{\boldsymbol{h_t}}] \end{aligned} \quad (1)$$

Different from LSTM, BiLSTM computes a set of h_t , and h_t is the concatenation of backward LSTM and forward LSTM. Every sentence has different length, In order to get a fixed-size vector to represent the sentence, we adopt a max pooling [4] layer after concatenation of two directional hidden units information. It selects the maximum over the same dimension, finally we obtain a fixed-size sentence representation. Facebook gains a remarkable performance using BiLSTM with max pooling as a universal sentence encoder. So we use it as our baseline.

2.2 BiLSTM with BF Mechanism.

For each hidden state h_t, it has different context, h_t consists of forward information $\overrightarrow{h_t}$ and backward information $\overleftarrow{h_t}$. If use max pooling to get fixed-size vectors, it selects the maximum over each dimension, but in the same dimension at different time steps, it contains unequal amount of information. For example, forward state $\overrightarrow{h_1}$ only has the information of w_1 , but $\overrightarrow{h_T}$ has the information of all words that consist the sentence. This situation also occurs on the backward information. It is irrational to select the maximum roughly over the same dimension, random noises affect selections and some useful information may relate mechanism leads to loss of useful information. In order to overcome the referred shortcomings, we propose an improved pooling mechanism named BF-max pooling as figure 1 shows.

At the very beginning, we propose an architecture as following, through BiLSTM we can get two directional hidden states at each time step, $\overrightarrow{h_t}, \overleftarrow{h_t}$. At each time step, the bi-directional representation contains all words information that consist the sentence. But two directions contain different amount of information at each time step, in order to make the important information stronger, we change the strength of $\overrightarrow{h_t}, \overleftarrow{h_t}$ according to their amount of information and concatenate them before pooling.

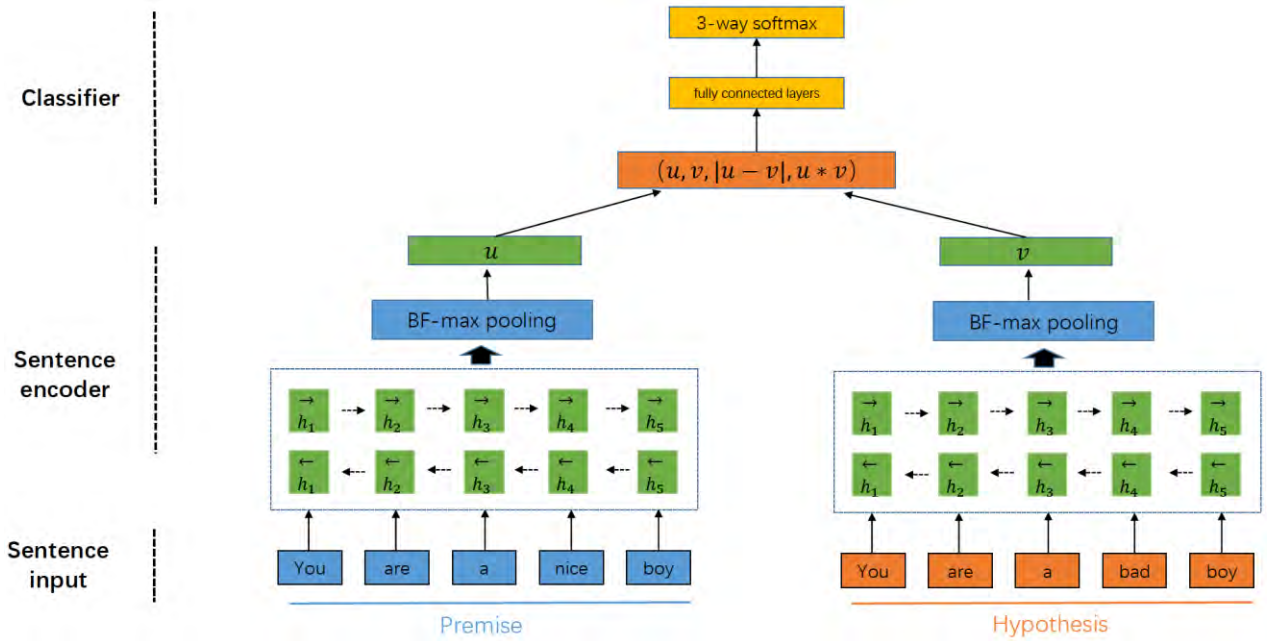

Fig. 1 Bi-LSTM BF-max pooling network

In the equation (2) W is a parameter matrix that measures the amount of information. For keeping the whole information stable, we use a scale multiply each direction hidden representation. Applying the pooling mechanism as it is mentioned, we can obtain a much more balanced sentence embedding S , therefore we call this mechanism as BF-balanced max pooling.

$$w_0 = \frac{\vec{I}}{\vec{I} + \overleftarrow{I}}$$

$$\boldsymbol{w_1} = \frac{\overleftarrow{\boldsymbol{I}}}{\vec{\boldsymbol{I}} + \overleftarrow{\boldsymbol{I}}} \qquad (2)$$

$$S = Maxpooling([2 * w_0 * \overrightarrow{h_t} + 2 * w_0 * \overleftarrow{h_t}])$$

Essentially, BF-balanced max pooling is a combination of bidirectional information. Therefore, in order to get a much more normal architecture, we simplify this mechanism as following equation (3), and call it as BF-combined max pooling.

$$H = [\vec{h}, \overleftarrow{h}]$$

$$\boldsymbol{A = WH + b} \qquad (3)$$

$$S = Maxpooling(A)$$

With the matrix W, at each time step, it combines bidirectional information, BF-combined max pooling avoids the situation that some sentence-level information relies on several dimensions but each of them is not the maximum, which means we have to discard a portion of information if select maximum directly. Different from self-attention mechanism, our proposed mechanism combines values at each time step in a confirmed way without different weights and selects the maximum over time steps, and in our experiment we compare our proposed model with a kind of self-attention model, which is named inner attention model. It seems can be replaced by a fully connected layer after a max pooling layer, but actually it is totally different, pooling layer may choose a different dimension if we exchange the order. Substantially BF-combined max pooling combines the bidirectional information in the sentence-level and select the maximum.

Attention mechanism is a valid method to get a fixed-size vector from a matrix. With the combination of backward information and forward information, we propose a model that adopts the BF-attention after BiLSTM. It is shown as following equation (4).

$$H = [\vec{h}, \overleftarrow{h}]$$

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{WH} + \boldsymbol{b} \\ \boldsymbol{S} &= \sum_{\boldsymbol{j=1}}^{\boldsymbol{T_x}} \boldsymbol{a_j w_j} \\ \boldsymbol{w_j} &= \frac{\mathbf{exp}(\boldsymbol{e_j})}{\sum_{\boldsymbol{k=1}}^{\boldsymbol{T_x}} \mathbf{exp}(\boldsymbol{e_k})} \\ \boldsymbol{e_j} &= \boldsymbol{u_w}^{\boldsymbol{T}} \boldsymbol{tanh}\left(\boldsymbol{W} \cdot \boldsymbol{a_j} + \boldsymbol{b}\right) \end{aligned} \qquad (4)$$

A is the output of the combination of backward and forward information, and w_j represents the weight of a_j, we adopt u_w to caculate all scores. Finally, we can get the sentence representation S. And this method is named as BF-attention.

2.2 SNLI Task Classification.

570k English sentence pairs of three categories: entailment, contradiction and neutral compose the SNLI dataset. The task is aimed to capture natural language inference, it is previously called as Recognizing Textual Entailment (RTE), the pair of sentences consists of a premise and a hypothesis. We acquire a pair of sentence representations using our proposed encoder, and put it into a classifier. As figure 1 shows, besides two sentence representations u, v, we also compute element wise product u*v, and element-wise difference |u-v|, then we concatenate them together, use a softmax layer after a fully connected layer to accomplish classification.

3. Experiments

The Experiments can be divided into two parts, we evaluate the performance of BF-max pooling on SNLI task, and the capability of our encoder as a universal sentence encoder on transfer tasks.

3.1 SNLI Parameter Setting.

We use SGD with a initial learning rate of 0.1 and 0.99 decay to train all models. The models are trained following the strategy that we divide the learning rate by 5 if dev accuracy decreases and stop learning if the learning rate decreases under 10^{-5} and use batch size 64 and pretrained Glove 300D embeddings. As for the dimension of LSTM encoder parameters, we respectively use 300, 2048 to do constrast experiments. For the classifier, 512 hidden units is adopted.

3.2 Evaluation Method.

As our aim is to obtain a universal sentence encoder, which means that embeddings generated from this encoder can cover a big set of sentence tasks, in order to evaluate the performance of a universal encoder, we use our pretrained model as encoder in 12 transfer tasks, and our evaluation is shown as follow. We adopt the open source tool to evaluate the transfer performance of our model. The tool used Adam [8] to t a logistic regression with batch size 64. In order to compare models on transfer tasks quanticationally, we adopt wo averages of development set (dev) results on transfer tasks whose metrics is accuracy, "micro" and "macro". "macro" aggregates score that corresponds to the classical average of dev accuracies, and the "micro" score is a sum of the dev accuracies, weighted by the number of dev samples.

3.3 Transfer Datasets.

To explore whether BF-max pooling can increase the performance as a pretrained sentence encoder on these following transfer tasks. These transfer tasks can mainly be divided into three types:

Table 1 Classification tasks

name	datasets	Class number	Task field	example
MR	11k	2	movies	”Too slow for a younger crowd , too shallow for an older one.” (neg)
CR	4k	2	product reviews	”We tried it out christmas night and it worked great .” (pos)
SUBJ	10k	2	subjectivity/objectivity	”A movie that doesn’t aim too high , but doesn’t need to.” (subj)
MPQA	11k	2	opinion polar	”don’t want”; ”would like to tell”; (neg, pos)
TREC	6k	6	question type	”What are the twin cities ?” (LOC:city)
SST	70k	2	movies	”Audrey Tautou has a knack for picking roles that magnify her [..]” (pos)

Binary and multi-class classification. As table 1 shows, it includes sentiment analysis (MR, SST), question-type (TREC), product reviews (CR), subjectivity/ objectivity (SUBJ) and opinion polarity (MPQA).

Entailment and semantic relatedness. We apply the existing method [16] to predict the probability distribution of relatedness scores. Pearson correlation is the measurement on SICK.

Semantic Textual Similarity . STS14 is a set of unsupervised SemEval tasks [1], and this dataset measures similarity with score from 0 to 5. We evaluate it by Pearson and Spearman correlations.

3.4 Result on SNLI and Transfer Tasks.

Table 2 Models performances on SNLI and transfer tasks

Model	Dim	Dev	Test
LSTM	2048	81.9	80.7
GRU	4096	82.4	81.8
BiGRU-last	4096	81.3	80.9
BiLSTM-mean	4096	79.0	78.2
Inner-attention	4096	82.3	82.5
HConvnet	4096	83.7	83.4
BiLSTM-intra-attention	600	84.5	84.2
BiLSTM-Max	600	84.3	84.0
BiLSTM-Max	1024	84.5	84.2
BiLSTM-BF-attention	600	84.7	84.3
BiLSTM-BF-balanced	600	84.5	84.3
BiLSTM-BF-combined	600	84.6	84.4
BiLSTM-Max	4096	85.0	84.5
BiLSTM-BF-attention	4096	85.2	84.8
BiLSTM-BF-combined	4096	85.1	84.9

Table 2 shows the performances on SNLI task. With the same size of BiLSTM hidden units, BiLSTM-BF encoders and intra-attention encoders perform well, but the latter one and BF-attention has poor a performance on transfer tasks. It reveals that attention models have ability to catch useful information from parts of a sentence for the training task, instead of general sematic information. Comparing with a self-attention model, our model outperforms inner-attention model which adopts a kind of self-attention mechanism. It shows better general adaptability on transfer tasks using BF-max pooling. Therefore, we select BiLSTM-Max-based models to generate a universal sentence encoder.

Comparing 600D BiLSTM-BF encoder with 1024D BiLSTM-Max encoder, model with BF-max pooling can decline much more parameters and gets a better performance. With less parameters it gets better performance, it reveals that only max pooling is too rough to get a sentence representation from hidden units of BiLSTM, backward information and forward information at different time steps

contain unbalanced information, combination of these features together can obtain more refined useful information. And we have more detailed comparations in next subsection on transfer tasks.

3.5 Result on Transfer Tasks.

Table 3 shows the performances on transfer tasks. Comparison with unsupervised unordered model. Besides untrained BiLSTM-BF model, this part includes bags-of-words models such as word2vec, Unigram-TFIDF, ParagraphVec [11], SIF model, they are all trained on Toronto book corpus. For introduction of ordered information, our model obtains better performances on most classification tasks, entailment and semantic relatedness tasks.

Comparison with unsupervised ordered model. This part includes FastSent and SkipThought, they are also trained on Toronto booktabs corpus, we also compare the variant model FastSent+AE and SkipThought-LN that applies layer normalization.

As the result shows, compared with SkipThought, our model gains a remarkable performance on STS task, SkipThought just has 0.44 Pearson score compared to our 0.69. It illustrates our model supervised by NLI task captures real distant of sentences , using element-wise product and element-wise difference. Compared with unsupervised models, our model has a stronger ability to represent sentence distribution in the semantic space. Comparison with supervised representation model. In this part, we compare our model with DictRep (bow) and NMT En-to-Fr, result shows that our model outperforms them distinctly.

Comparison with BiLSTM-Max model. On all transfer tasks, our model outperforms BiLSTM-Max encoder. Moreover, we get remarkable performances on several tasks (SUBJ, MPQA, TREC, SICK) even it is trained on a smaller size dataset than BiLSTM-Max. It illustrates BF-max pooling can capture the combination relationship of bidirectional hidden units to optimize sentence representations accurately and does not lead to only improvement on the specific task.

We get the state-of-art performance on SICK task. We obtain a score of 0.888 on SICK-R task that is better than the previous model [16]. On SICK-E task, our model is better than the previous best hand-engineered models [10] that gets accuracy of 0.845. We suppose our sentence encoder can learn the in-domain information, what limits the performance is just the out-domain information.

Table 3 Transfer results for different architectures

Model	MR	CR	SUBJ	MPQA	SST	TREC	SICK-R	SICK-E	STS14
Unsupervised representation training (unordered sentence)									
Unigram-TFIDF	73.7	79.2	90.3	82.4	-	85.0	-	-	.58/.57
ParagraphVec	60.2	66.9	76.3	70.7	-	59.4	-	-	.42/.43
Glove BOW	78.7	78.5	91.6	87.6	79.8	83.6	0.800	78.6	.54/.56
word2vec BOW	77.7	79.8	90.9	88.3	79.7	83.6	0.803	78.7	.65/.64
untrainedBiLSTM-BF	77.6	81.4	89.6	88.6	80.8	85.7	0.860	83.2	.39/.48
Unsupervised representation training (ordered sentences)									
FastSent	70.8	78.4	88.7	80.6	-	76.8	-	-	.63/.64
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	-	-	.62/.62
SkipThought	76.5	80.1	93.6	87.1	82.0	92.2	0.858	82.3	29/.35
SkipThought-LN	79.4	83.1	93.7	89.3	82.9	88.4	0.858	79.5	.44/.45
Supervised representation training									
DictRep (bow)	76.7	78.7	90.7	87.2	-	81.0	-	-	.67/.70
NMT En-to-Fr	67.4	70.1	84.9	81.5	-	82.8	-	-	.43/.42
BiLSTM-Max	79.9	84.6	92.1	89.8	83.3	88.7	0.885	86.3	.68/.65
BiLSTM-BF	80.2	85.3	92.5	90.3	84.2	88.7	0.886	86.6	.69/.67

4. Conclusion

In this paper, we propose an improved pooling mechanism named BF-max pooling to obtain a well-performed universal sentence encoder. BF-max pooling mechanism improves sentence representations by combination of backward and forward hidden states of BiLSTM. Experiments demonstrate that BiLSTM with BF-max pooling outperforms other models on SNLI task, and obtains

better performances on transfer tasks, attention-based models have good ability to specific task but is not very suitable for sentence transfer tasks. It illustrates that our model is effective to improve the performance as a universal sentence encoder with less parameters.

References

[1]. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre,A., ... & Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantictextual similarity. In Proceedings of the 8th international workshop onsemantic evaluation (SemEval 2014) (pp. 81-91).

[2]. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

[3]. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[4]. Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp.160-167). ACM.

[5]. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017).Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

[6]. Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483.

[7]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[8]. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[9]. Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

[10]. Lai, A., & Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 329-334).

[11]. Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International Conference on Machine Learning (pp. 1188-1196).

[12]. Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

[13]. PLiu, Y., Sun, C., Lin, L., & Wang, X. (2016). Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv:1605.09090.

[14]. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013).Distributed representations of words and phrases and their compositionality.In Advances in neural information processing systems (pp. 3111-3119).

[15]. Arora, S., Liang, Y., & Ma, T. (2016). A simple but tough-to-beat baseline for sentence embeddings.

[16]. Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

[17]. Zhao, H., Lu, Z., & Poupart, P. (2015, July). Self-Adaptive Hierarchical Sentence Model. In IJCAI (pp. 4069-4076).