Co-Driven Recognition of Chinese Semantic Entailment via the Fusion of Transformer and HowNet Sememes Knowledge*

No Author Given

No Institute Given

Abstract. Chinese semantic entailment recognition aims to detect and judge whether the semantics of two Chinese sentences are consistent and whether there is entailment relationship. However, the existing methods usually face the challenges of chinese synonyms, polysemy and difficulty to understand long texts. To solve the above problems, this paper proposes a co-driven Chinese semantic entailment recognition method based on the fusion of Transformer and HowNet sememes knowledge. The experimental results on financial and multi semantic interpretation data sets such as BQ, AFQMC, PAWSX, etc. show that, compared with lightweight models such as DSSM, MwAN, DRCN, and pretraining models such as ERNIE, this model can not only effectively improve the accuracy of Chinese semantic entailment recognition (by 2.19%, 5.57% and 6.51% compared with the DSSM, MWAN and the DRCN model on the BQ dataset), control the number of model parameters (about 16M), but also adapt to long text implication recognition scenarios of more than 50 words.

Keywords: chinese entailment recognition · sememes knowledge fusion · transformer · HowNet and entailment recognition.

1 Introduction

Text entailment recognition task is one of the important sub tasks of natural language processing, and can be applied to a large number of natural language processing. It seems to be a simple task, but it has a wide variety in different downstream tasks, such as information retrieval, question answering system, dialogue system, machine translation, etc. The input of this task is mainly a sentence pair, and the final result is to judge the logical relationship (neutral, entailment, contradiction) between the sentence pairs or whether the sentence pairs are similar.

The task of Chinese text entailment recognition is one of the more difficult tasks, because the characteristics of polysemy and synonymy in Chinese are more prominent than those in other languages, which leads to the situation that "words fail to express meaning" in daily communication. However, traditional algorithms cannot obtain semantic information in Chinese, such as Bow, VSM, TF-IDF, BM25, Jaccord, SimHash and other classical algorithms, which mainly solve the matching problem at the lexical

* Supported by organization x.

level, or the similarity problem at the lexical level. However, the matching algorithm based on lexical coincidence has great limitations, for the following reasons:

1.The meaning of a word is limited. The same word can express different meanings in different contexts.

2.The structure is limited. For a chinese word, the meaning expressed in turn order will be completely different, such as "cow" and "milk".

3.Knowledge limitation: a sentence is wrong in combination with common sense, but it is correct in terms of morphology and syntax.

In order to solve the above problems, this paper proposes a text entailment recognition method based on Transformer [1] and HowNet, which expands the research on sentence semantic information acquisition. First, we use Transformer to conduct multi-level coding and data-driven on the semantic information of the internal structure of Chinese sentences, and introduce an external knowledge base, HowNet, to conduct knowledge driven modeling of semantic knowledge association between vocabularies. Then, we use Soft Attention to calculate mutual attention and achieve knowledge fusion with the semantic matrix. Finally, we use BiLSTM to further encode the semantic information of the conceptual level of text and infer and judge semantic consistency and entailment relationship. A number of experiments show that, compared with the lightweight model and the pre training model, the reasonable use of HowNet's semantic knowledge and Transformer's advantages for long texts can improve the accuracy of the model to a certain extent.

The model proposed in this paper has the following innovations:

- In order to solve the problem of Chinese synonyms, polysemy and difficult to understand long texts in Chinese semantic entailment recognition, a dual drive Chinese semantic entailment recognition method based on the fusion of Transformer and HowNet semantic primitives is proposed. In addition to using data samples to drive Transformer coding and model pre training, HowNet semantic primitives are used to drive knowledge to enhance synonyms The deep semantic understanding of polysemy.
- A technical means of semantic knowledge fusion is proposed. The internal structure semantic information of Chinese sentences is encoded in multiple levels through Transformer, and the external knowledge base HowNet is introduced to model the semantic knowledge base between vocabularies. At last, soft attitude is used to calculate interactive attention and realize knowledge fusion in the semantic matrix.
- The proposed model has been applied to ant finance, banking finance industry and multi semantic application scenarios, and compared with lightweight models such as DSSM [2], MwAN [3], DRCN [4] and pre training models such as ERNIE. It can not only effectively improve the accuracy of Chinese semantic entailment recognition, control the model parameters, but also adapt to long text entailment recognition scenarios of less than 50 characters. The code will be released at https://github.com/Platanus-hy/simese_codriven_entialment.

2 Related works

In recent years, with the rapid development of deep learning, machine learning methods for text entailment recognition have been proposed in large numbers. In terms of semantic information acquisition, the classic short text matching model DSSM solves the problem of dictionary explosion in LSA, LDA and other methods, but also loses context information due to the use of the word bag model. The ESIM [5] model, proposed in 2016, comprehensively applies the BiLSTM and attention mechanism, and makes interaction between two sentences in local reasoning for the first time. The DIIN [6] model proposed in 2018 uses CNN and LSTM for feature extraction, but the author uses both word vectors and local vectors in its input layer, inputs some additional syntactic features, and uses DenseNet for feature extraction. The DRCN model proposed in 2018 draws on DenseNet [7]'s intensive connection operation in image recognition. Through intensive connection to RNN, it retains the most original information of the text. Through multiple cycles, it continuously adds interactive information to the matrix vector, and finally outputs the full connection layer. The KIM [8] model proposed in 2018 uses the external knowledge base WordNet [9] to judge the logical relationship between two sentences and embed external prior knowledge into the similarity matrix. In the MwAN model proposed in 2018, the author uses a variety of attention mechanisms (splicing, bilinear, dot multiplication, subtraction) to fully capture the relationship between sentence pairs. Finally, multiple results are weighted and combined, and the final probability is output through GRU and full connection layers.

In terms of sentence structure, CT-LSTM [10], which was proposed in 2015, proposed a tree shaped LSTM to solve the problem that LSTM cannot extract the structural information of sentences, and discussed the sequence problem. Different from the commonly used RNN sequence modeling, it uses the dependency relationship of sentences as the input of LSTM, which also has some inspiration for future research.

With the proposal of BERT [11] model in 2018, a trend of pre training models swept the whole NLP world and ranked among the top in the major NLP lists. BERT has a complete Encoder Decoder framework. Its basic composition is Transformer, which is mainly composed of multi head attention. It is a model built with pure attention, which can solve the problem of long-distance dependence in RNN machine argumentation. That is, the attention mechanism has better memory and can remember information over a longer distance. The advantage of BERT model is that it can learn more grammatical and semantic information in sentences, make the output word vector more representative, and the larger number of parameters make it perform well in various downstream tasks. In order to solve the problem that the pre training model has a large number of parameters and is difficult to apply on consumer GPUs, Tim Dettmers [12] proposed LLM.Int8() for Transformer, which enables ordinary consumer GPUs to use very large models without lowering performance. In addition to Transform, Hanxiao Liu [13] proposed another simple, attention independent architecture, gMLP, which achieves the same level as Transformer in some indicators of pre training, and even outperforms Transformer in some downstream tasks.

Due to the special polysemy of Chinese words, there are certain difficulties in Chinese text matching. Different words often contain the same meaning, such as "China" and "Huaxia", which are semantically consistent, but not related in terms of grapheme.

In order to solve this problem, many researchers choose to use Chinese part of speech, dependency syntax and other information to calculate the similarity. For example, Yan Jiao [14] and others tried to label the text with part of speech, and only reserved nouns, verbs and adjectives. They obtained word pairs by combining dependency syntax analysis. With PageRank [15] and degree centrality as indicators, they established a grammar network for a large number of texts, and proposed a text similarity calculation method combining syntactic relations and lexical semantics.Huang Yan [16] proposed A text level text automatic generation model based on topic constraints. The synonymy of the keyword set is used to generate multiple article theme plans Li Lin [17] put forward the concept of concept vector space. They expressed the document as a set of concept words to build a vector space, and then calculated the semantic similarity through cosine similarity. The effect is better than the word bag model+Word2Vec [18].

Also using HowNet for Chinese entailment recognition is the LET proposed by Boer Lyu [19] in 2021. They transformed the initial vectors of all the sememes under each Chinese word with graph attention, then obtained the sememe vectors of each word through attention pooling, and obtained the final word vector through the fusion of GRU and BERT word vectors. Although Chinese words often have multiple semantics, we often need to use only a few correct primitives to identify semantics, which will lead to the fact that the vector of primitives obtained by them cannot match the actual sentence well, and will contain redundant semantic information. In this paper, the semantic information is preliminarily screened and then incorporated into the interaction matrix to avoid adding redundant information.

3 Research Method

This section mainly introduces the dual drive Chinese entailment recognition model based on the fusion of Transformer and HowNet semantic primitives, and analyzes its main structure and functions. Its main structure is as follows. It is divided into 6 layers, namely Transformer layer, Attention layer, BiLSTM layer, average pooling layer and maximum pooling layer, and full connection layer.

3.1 Transformer Layer

The role of the Transformer layer is to enable the vectorized text to obtain deep semantic information through the neural network. The commonly used neural networks are CNN, LSTM, etc. This model uses Transformer architecture as the coding layer of the text to process sentence vectors. It is mainly composed of multi head attention mechanism and feedforward neural network, which can alleviate the problem of gradient disappearance. Its single head attention mechanism is calculated as follows.

$$Query = \mathrm{w}^{Q} X \tag{1}$$

$$Key = \mathrm{W}^{K} X \tag{2}$$

$$Value = \mathrm{W}^{V} X \tag{3}$$

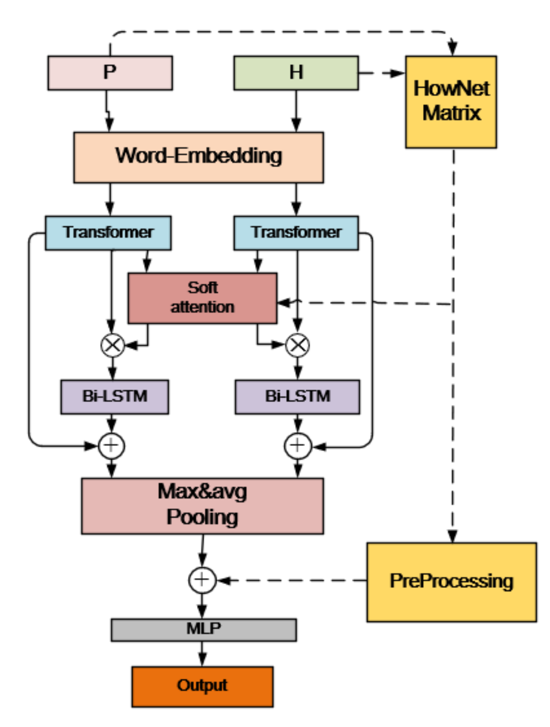

Fig. 1. Model structure.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

The dimension of the input vectors are $W^Q, W^K, W^V \in R^{d_{\text{model}} \times d_k}$

However, because it does not obtain the position information of the text sequence, it is necessary to add absolute position coding to the word vector. Its calculation method is shown in Formula (5) (6).

$$PE_{(p,2m)} = \sin\left(\frac{p}{10000^{2m/d}}\right) \tag{5}$$

$$PE_{(p,2m+1)} = \cos\left(\frac{p}{10000^{2m/d}}\right) \tag{6}$$

d represents the dimension of the current word vector, and p represents the position of the current word in the sentence.

3.2 Attention Layer

Attention layer is an important component of text entailment recognition model, which has the advantages of fast speed, good effect and few parameters. There are many different types, such as Soft attention, Hard attention, Self attention, etc. Haili Sun [20] also proposed the mechanism of focus and local attention. This paper adopts the commonly used soft attention mechanism, but adds the semantic matrix information generated based on HowNet to it, and adds the trainable weight.

HowNet is a common sense knowledge base based on Chinese and English, which explains the relationship between concepts and their attributes. This paper mainly uses the external knowledge base to obtain all the Chinese sememes of the corresponding two words in the sentence pair. If two words have the same meaning, the value of their corresponding position in the matrix will be set to 1, otherwise it will be set to 0. Figure 2 below takes two example sentences for semantic analysis.

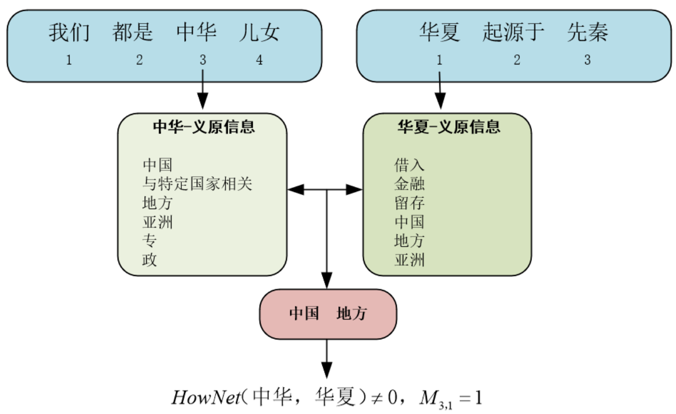

Fig. 2. Analysis of HowNet semantics.

The top box represents the result of sentence segmentation, the middle box represents the multiple source information corresponding to the current word, and the lower box represents the intersection of the source information of two words. It can be seen from the above diagram that "China" has Chinese, country related, local and Asian etymologies, and "Huaxia" has borrowing, finance, retention, China and country etymologies. The intersection of the two words is China, country and local, so at this situation.

$$M_{i,j} = \begin{cases} 1 & \text{HowNet}\left(P_i, H_j\right) \neq 0 \\ 0 & \text{How Net}\left(P_i, H_j\right) = 0 \end{cases} \tag{7}$$

$$M = \begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \end{bmatrix} \tag{8}$$

The generation formula of Attention matrix is shown in formula (9):

$$e = PH^T + \gamma \cdot M \tag{9}$$

Where γ is a trainable parameter. At this time, the Attention matrix e not only integrates the text information between sentences, but also obtains the semantic information of word pairs between sentences. The matrix change heat diagram is shown in Figure 4. After adding the semantic information, the weight of some positions is increased, which indicates that the position obtains the information of the semantic matrix. After obtaining the improved attention matrix, the soft attention is calculated as follows:

$$\hat{P} = \sum_{j=1}^{l_h} \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{l_h} \exp\left(e_{ik}\right)} P_{tf}, \forall i \in \left[1, 2, \cdots, l_p\right] \tag{10}$$

$$\hat{H} = \sum_{j=1}^{l_p} \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{l_p} \exp\left(e_{ik}\right)} H_{tf}, \forall i \in [1, 2, \cdots, l_h] \tag{11}$$

In Formula (10) (11), are matrix vectors of sentence pairs after being Transformer. It indicates the length of the sentence and the output after the soft attention mechanism.

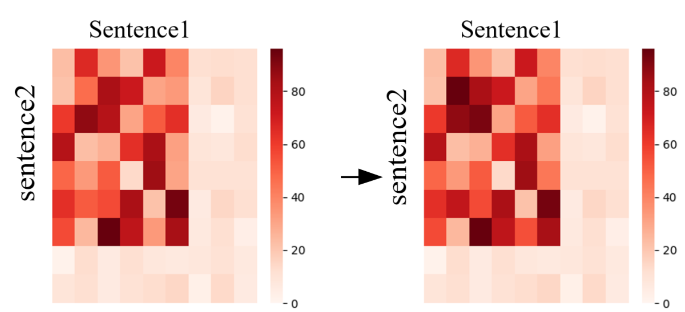

Fig. 3. Attention matrix changes.

3.3 BiLSTM Layer

This layer is used to output and process the Transformer layer through the soft attention mechanism. Through the BiLSTM, the forward coding and backward coding are spliced to further obtain the context information.

The input of the long and short term memory network is as follows:

$$P_{bi-lstm} = Bi\,\mathrm{LSTM}(\hat{P}) \tag{12}$$

Wherein, refers to the output vector of the sentence after the soft attention mechanism, and refers to the vector after the LSTM.

3.4 Average Pooling and Max-Pooling Layer

In order to fuse the text information after Transformer and BiLSTM, this model splices multiple inputs through max-pooling and average pooling. The purpose of this layer is to change the vector dimension of two sentences from to facilitate subsequent input of full connection layer:

$$P_o = \text{ Concat } \left(\left[P_{tf}; P_{bi-lstm}\right]\right) \tag{13}$$

$$P_{rep} = [\text{Max}\,(P_o)\,; \text{Mean}\,(P_o)] \tag{14}$$

3.5 Fully Connected Layers

After obtaining the complete sentence vector expression of the sentence pair and , the common vector splicing method is direct splicing and inputting it into the multi-layer feedforward neural network to obtain the results. When the proposed model is spliced, the information in the HowNet matrix is considered, and the sum of two dimensions and in the HowNet matrix is obtained. And splice it with and . The final input H of the feedforward neural network is obtained.

$$\begin{gathered} HN_{\text{row}} = \text{sum}(M, axis = 0) \\ HN_{\text{line}} = \text{sum}(M, axis = 1) \\ H = \text{concat}\left(P_{rep}; H_{\text{rep}}; P_{rep} - H_{rep}; HN_{\text{col}}\,; HN_{\text{row}}\right) \end{gathered} \tag{15}$$

Here *sum*(), for example, HN_{row} represents the result of the sum of the HowNet matrix along the first dimension, that is, the HowNet information corresponding to sentence 1. Through vector splicing, the corresponding semantic information of the two sentences is obtained, in which, $P_{rep}-H_{rep}$ also represents the difference between the two sentence vectors.

3.6 Predicted Layers

After obtaining the final sentence vector expression of the sentence pair. The model uses a two-layer fully connected neural network to obtain sentence matching results. The loss function used is the cross entropy loss function, and its calculation method is shown in Formula (16)

$$output = FFN(H)Loss = \frac{1}{N}\sum_i -\left[y_i \times \log\left(p_i\right) + (1 - y_i) \times \log\left(1 - p_i\right)\right] \tag{16}$$

Where represents the label of the sample, the positive sample is 1, and the negative sample is 0. indicates the probability that the sample is predicted to be a positive sample.

In addition to the commonly used cross entropy loss function, we also tried the CoSent loss function, which makes the similarity of positive sample pairs greater than that of negative samples, and makes the distance between positive and negative samples in the vector space as far as possible. The experiment shows that using the CoSent loss function has certain effect on pre training methods such as BERT and Sentence-BERT [21], making the pre training model converge faster. But for the model proposed in this paper, In the case of non pre training, its effect is not as good as cross entropy loss function.

In the training phase, we used MultiStepLR to dynamically adjust the learning rate. In the 20th, 50th, 80th, 100th and 150th iterations of the experiment, we updated the learning rate with a decay rate of 0.5. By dynamically adjusting the learning rate as the number of iterations increases, the convergence speed of the model increases, and its change trend is shown in Figure 4 below.

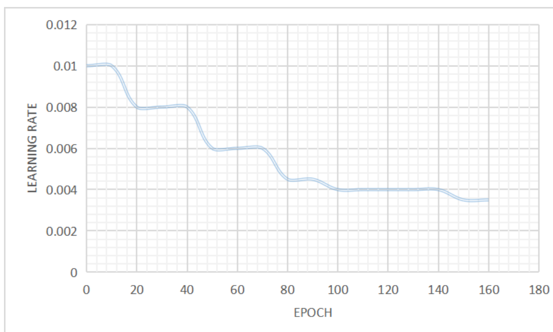

Fig. 4. Learning rate decay.

4 Experiments And Analysis

4.1 Datasets

Table 1. Dataset Size

Dataset Name	Training Set Size	Validation Set Size	Test Set Size
PAWSX	49401	2000	2000
AFQMC	34334	4316	3861
BQ Corpus	100000	10000	10000

In order to verify the effectiveness of the text entailment recognition model based on Transformer and HowNet dual drivers proposed in this paper, this paper conducts experiments on three open datasets respectively. The data sets are respectively PAWSX data set, AFQMC data set and BQ Corpus data set.

PAWSX zh dataset is a multilingual definition pair dataset released by Google. It is characterized by highly overlapping vocabulary, which is helpful to further improve the model's judgment on difficult samples, and can test the model's judgment ability on similar samples. The AFQMC dataset is an ant financial similarity dataset, which contains 34,334 training data, 4316 validation data and 3861 test data. BQ Corpus is the problem matching data set in the banking and financial field, including the problem pairs extracted from the online banking system log of one year. It is the largest problem matching data in the banking field at present, including 10000 training data, 10000 verification data, and 10000 test data.

Table 2. Example of BQ Dataset

Sentence1	Sentence2	label
微信消费算吗	还有多少钱没还	0(contradiction)
下周有什么好产品?	元月份有哪些理财产品	1(entailment)
能查账单吗	可以查询账单	0
1无法借款	Qq有微粒代	0

4.2 Experimental Setup

The experiment of this paper adopts a 4-card GPU server with the model of RTX2080ti. The model training parameters and software version are shown in Table 3.

Table 3. Initial training parameters of the model

Parameter	Value
Word-embedding dim	300
Number of hidden layers	128
Maximum sequence len	100
batch_size	64
Number of Transformer layers	10
Optimizer	Adam
Initial learning rate	0.01
Software	Edition
python	3.6.13
torch	1.10.2
OpenHowNet	2.0
transformers	4.18.0

4.3 Comparison of experimental results

In order to verify the actual effect of the model proposed in this paper, three classical text matching models, DSSM, MwAN and DRCN, are selected as non pre training models. For the pre training model, we chose BERT wwm ext, BERT and Baidu ERNIE.

The selected data sets are PAWSX, AFQMC and BQ Corpus. In order to ensure the unity of the experiment, all models use the same jieba vocabulary for the same dataset, and the indicators of comparison are ACC and F1-score.

Table 4. Example of BQ Dataset

Model name	Pre-trained	Acc	F1
DSSM	✗	77.12	76.47
MwAN	✗	73.99	73.29
DRCN	✗	74.65	76.02
Ours	✗	78.81	76.62
Improvement	✗	+2.19%	+1.96%
BERT-wwm-ext	✔	84.71	83.94
BERT	✔	84.50	84.00
ERNIE	✔	84.67	84.20
Ours-BERT	✔	84.82	84.33
Improvement	✔	+0.177%	+0.464%

Table 5. Experimental Results of AFQMC Dataset

Model name	Pre-trained	Acc	F1
DSSM	✗	57.02	30.75
MwAN	✗	65.43	28.63
DRCN	✗	66.05	40.60
Ours	✗	66.62	42.93
Improvement	✗	+0.86%	+5.7%
BERT-wwm-ext	✔	81.76	80.62
BERT	✔	81.43	79.77
ERNIE	✔	81.54	80.81
Ours-BERT	✔	81.84	81.93
Improvement	✔	+0.097%	+1.38%

It can be seen from Table 4 that the accuracy of the proposed model in BQ dataset is higher than that of other models. As shown in Table 5, from the perspective of data sets, the results of the three models on AFQMC are not very good. Preliminary analysis shows that the language standardization of the sample data is poor, such as incomplete sentences, such as "可以帮我冻结花呗吗" and "里冻结花呗额度". The label is entailment, which leads to poor results in training set and test set. However, the model proposed in this paper integrates Transformer and Hownet's external knowledge base, which has a better effect. The results show that for non-standard text, the performance can be improved by obtaining the semantic information of some words and matching

Table 6. Experimental Results of PAWSX-zh Dataset

Model name	Pre-trained	Acc	F1
DSSM	✗	42.64	59.43
MwAN	✗	52.70	52.65
DRCN	✗	61.24	56.52
Ours	✗	62.55	59.72
Improvement	✗	+2.13%	+0.48%
BERT-wwm-ext	✔	77.23	76.52
BERT	✔	77.06	77.16
ERNIE	✔	78.02	77.59
Ours-BERT	✔	78.33	77.96
Improvement	✔	+0.397%	+0.476%

them. As shown in Table 6, for the PAWSX dataset where the data samples are difficult to sample, the traditional DSSM model cannot obtain interactive information and context information, so the effect is poor. For hard-negative samples, the highly similar sentence pairs lead to too similar semantic information. The HowNet matrix generated by each pair of sentences is almost identical, so the method of obtaining semantic knowledge is not good for judging the positive and negative samples of difficult samples.

From the perspective of error analysis, because we directly used jieba word segmentation to preprocess the text, the word segmentation error produced by it has different degrees of influence on the experimental results. Although there are segmentation errors, all models use the same vocabulary for the same dataset, and the model proposed in this paper is more effective than others.

4.4 Ablation Study

In this section, in order to understand the relative importance and effectiveness of each part of the model, ablation studies were conducted on different structures of the model proposed in this paper. There are two experiments, both of which use BQ dataset. Experiment 1 evaluates the impact of using different word segmentation tools and whether to use HowNet, an external knowledge base, on the experimental results; Experiment 2 evaluated the impact of different text lengths on HowNet in BQ dataset; Experiment 3 changed the number of Transformer layers and evaluated the impact of the number of Transformer coding layers on the experimental results.

It can be seen from the experimental results in Table 7 that using HowNet can improve the performance of the model to some extent. Compared with not using HowNet, the accuracy of various word segmentation tools is improved. If there are some words with complex semantics in the data sample, the introduction of external knowledge base can significantly improve the sensitivity of the model to polysemy and synonym, and can significantly improve the performance of the model.

According to the results in Table 8, HowNet can effectively improve the performance of both texts with a length of less than 15 and texts with a length of more than

Table 7. Results of ablation experiment 1

Tokenizer	HowNet	Acc	F1
Jieba	✔	**0.7881**	**0.7662**
	✗	0.7783	0.7624
PKUseg	✔	**0.7869**	**0.7653**
	✗	0.7792	0.7611
HanLP	✔	**0.7853**	**0.7599**
	✗	0.7735	0.7512

Table 8. Results of ablation experiment 2

Sequence length	HowNet	Acc	F1
1~15	✔	**0.7869**	**0.7662**
	✗	0.7763	0.7521
15~50	✔	**0.7884**	**0.7684**
	✗	0.7792	0.7545

15 and less than 50, and can obtain more valid semantic information in longer texts to achieve better results. After the longest text segment experiment in the dataset, the longest text length that this model can handle is 50 on the basis of ensuring the experimental effect.

Table 9. Results of ablation experiment 3

Number of Transformer layers	Acc	F1
2	0.7433	0.7353
4	0.7648	0.7572
6	0.7752	0.7731
8	0.7853	0.7658
10	0.7881	0.7662

From the result data in Table 9, the higher the number of Transformer layers, the better the effect of the model. By stacking the Transformer coding layer, the performance of the model can be improved to a certain extent, but at the same time, the number of parameters of the model and the training time of the model will be significantly increased, and the convergence speed will also be significantly slower. We take the model with 6 coding layers as the optimal model, and its parameter quantity is 16M, while the DRCN model with the best effect in the non pre-trained model is 19M. During training, the changes of trainable parameters are as follows:/par

It can be seen from Figure 5 that during the experiment, by observing the changes of the trainable parameters of the attention matrix, the weight of the primitive information matrix obtained by HowNet in the attention matrix is gradually increased with the

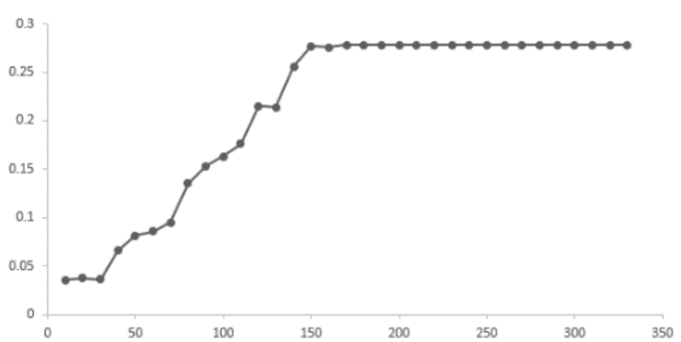

Fig. 5. Variation diagram of trainable parameters

increase of the number of iterations, which indicates that the HowNet primitive information generated by the original text has a positive role in improving the model effect.

5 Conclusion

The text entailment recognition model based on Transformer and HowNet is proposed in this paper, which uses the HowNet external knowledge base to perform semantic matching for the word pairs in sentence pairs. At the same time, Transformer based on multiple heads of attention is used to obtain text information, and BiLSTM, which can better obtain sequence information, is used to fuse multiple information in the model, making the model more sensitive to conceptual information. The experiment shows that, compared with DSSM, MwAN, DRCN models and pre training models without semantic information, this model has a certain improvement on BQ, AFQMC and PAWSX datasets. By stacking Transformer layers, the experimental results can be effectively improved, but at the same time, the number of parameters also increases. In contrast, the model proposed in this paper has fewer parameters and better effect. Compared with LET, which also uses HowNet as the external information base, this paper only filters the relevant semantic information in the use of HowNet semantic information, avoiding the impact of other semantic elements on the results, which is more accurate and more intuitive. Through theoretical innovation, technological innovation and application innovation, this paper proposes a research method of Chinese semantic entailment driven by Transformer and HowNet semantic primitive knowledge, and applies the model to financial and other specific application fields. Experiments show that the model proposed in this paper can effectively use semantic knowledge, effectively improve the accuracy of Chinese entailment recognition, and can also adapt to long text entailment recognition scenarios within 50 characters, Whether in the lightweight model or in the pre training model, there is a certain improvement effect.

In the future research work, we will continue to improve the model, add its original knowledge to the internal structure of Transformer, and further improve the effect on Chinese text. In addition, more work is needed to supplement the external knowledge base of Chinese.

References

1. Vaswani A, Shazeer N, Parmar N , et al. Attention Is All You Need. arXiv, 2017.
2. Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data，Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013
3. Tan C, Wei F, Wang W, et al. Multiway attention networks for modeling sentence pairs, Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI-18. 2018.
4. Kim S, Kang I, Kwak N. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. 2018.
5. Chen Q, hu X, Ling Z, et al. Enhanced LSTM for Natural Language Inference. 2016.
6. Yichen G, Heng L, Jian Z, et al. Natural Language Inference over Interaction Space, ICLR 2018.
7. Huang G, Liu Z, Laurens V, et al. Densely Connected Convolutional Networks, IEEE Computer Society. IEEE Computer Society, 2016.
8. Qian C, Zhu X, Ling Z H, et al. Neural Natural Language Inference Models Enhanced with External Knowledge，Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
9. Lin D. Review of: WordNet: An Electronic Lexical Database, computational linguistics, 2002.
10. Tai K S, Socher R, CD Manning. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, Computer Science, 2015, 5(1):: 36.
11. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2019.
12. Dettmers T, Lewis M, Belkada Y, et al. LLM.int8: 8-bit Matrix Multiplication for Transformers at Scale. arXiv e-prints, 2022.
13. Hanxiao Liu, Zihang Dai, David R. So, et al. Pay Attention to MLPs. NeurlPS 2021.
14. Jiao Y, Jing M, Kang F. Text Semantic similarity calculation in syntactic network based on Fusion co-occurrence distance, Modern library and information technology, 2019, 003(012):93-100.
15. Page, Lawrence, Brin, et al. The PageRank citation ranking: Bringing order to the web. stanford digital libraries working paper, 1998.
16. Yan H, Haili S, Ke X, at al. Text generation method based on topic constraint. Natural Science Edition of Peking University, 2020, 56(1): 9-15.
17. Lin L, Hui L. A text similarity calculation method based on concept vector space. Data analysis and knowledge discovery, 2018, 2(5): 11.
18. Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space. Computer Science, 2013.
19. Boer Lyu, Lu Chen, et al. Let:Linguistic Knowledge Enhanced Graph Transformer for Chinese Short Text Matching. Proceedings of the AAAI Conference on artificial Intelligence. 2021.
20. Haili Sun, Yan Huang, Songfeng Lu. Improving Fine-grained Text Sentiment Transfer for Diverse Review Generation. In: Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA 2020), Dalian, China, June 27-29, 2020. IEEE, 2020: 261-266
21. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.