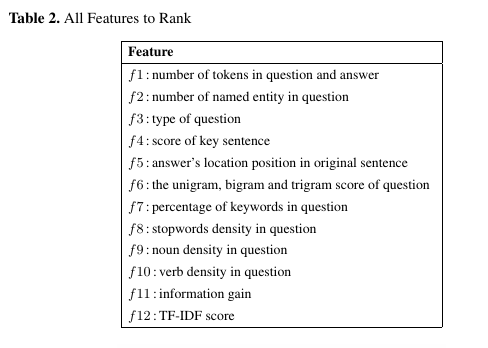
Responses to the Reviewer Comments

Dear Editors and Reviewers,

Thank you for your letter and for the reviewers’ comments concerning our manuscript entitled “A Novel Framework for Automatic Chinese Question Generation Based on Multi-Feature Neural Network Model” (ID: 7075). Those comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches.

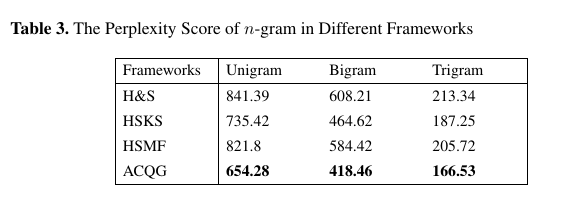
Specifically, we use templates and multi-feature to generate more meaningful Chinese questions. Compared to previous version, we add one *color* template and two more features. One is information gain and the other one is TF-IDF score. Information gain is a probability data, which represents the uncertainty of sentence. TF-IDF score is a numerical statistic measure that is a popular sentence-weighting scheme. Those help represent the sentence well. From the results, the perplexity score decreases 2.4% (Unigram), 2.62% (Bigram) and 3.59% (Trigram), and the human rated score increases 2.15% than previous version. Those features really improve the ranking results. And the detail description is shown in Section 3.3, Page 6, Paragraph 24 as follows:



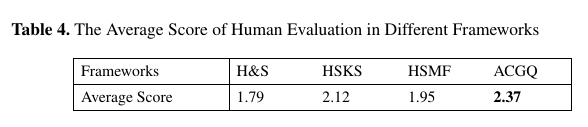
As shown in Table.2, we select twelve features about generated questions and original text. Additionally, the features f1~f5 are basic information of question and answer. f6 are the *n*-gram scores in question sequence, which provide rich contextual information. f7 is used to compute the keywords percentage in question. f8~f10 are different tags of word distribution in question, which indicate the structure of question sequence. f11 is a probability data, which represents the uncertainty of sentence. f12 is a numerical statistic measure that is a popular sentence-weighting scheme.

The analysis of results is shown in Section 4.2, Page 10 as follows:

Table.3 shows the perplexity score of *n*-gram in each framework. Unigram is n=1, Bigram is n=2 and Trigram is n=3. ACQG outperforms other frameworks, whose perplexity scores are 654.28 (Unigram), 418.46 (Bigram) and 166.53 (Trigram). H&S uses the over-generated questions and ranks them to get the output questions. The full text is used to generate questions, thereby the outputs are redundant and the scores are 841.39 (Unigram), 608.21 (Bigram) and 213.34 (Trigram). HSKS filters unmeaning sentences in texts, which is able to generate more targeted questions. Thus the scores decrease dramatically. Referring to HSMF system, HSMF is superior to H&S slightly. However the score of unigram is greater than H&S. The reason is that they both construct more absurd questions and the ranking method cannot improve this issue effectively. When we improve both key sentence extracting and question ranking, the performance of ACQG achieves best.



Furthermore, we introduce human evaluation into this work. The average marking scores are shown in Table.4, the detailed score distribution in each topic is shown in Fig. 3.



From Table.4, we can see ACGQ has the highest score 2.37, which indicates that the generated questions are around *Good* and *Borderline*. The questions are more likely to be acceptable. Because of the bad input and fragile ranking method, H&S has a poor performance, whose human evaluation score is only 1.79. For HSKS and HSMF, whose human evaluation score are 2.12 and 1.95 respectively. They both show better performance than H&S, but perform worse than ACGQ. Thus, the results of human evaluation are consistent with the above automatic evaluation.

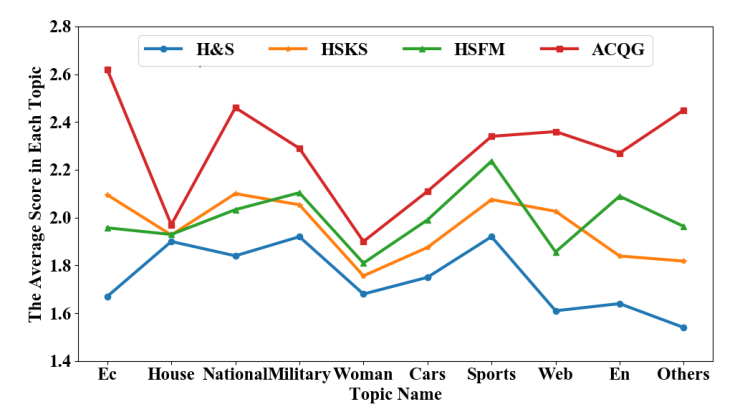


Fig.3. The Human Evaluation in Each Topic

Fig.3 shows the detailed human evaluation score distribution in each topic. We can observe that all frameworks have a low satisfaction score in the topic of House. The reason is that the collected news is almost advertisements in House, which cannot generate meaningful questions. For other topics, H&S always shows a lower score in each topic, while ACQG always shows a higher score in each topic. For other frameworks, the results fluctuate within H&S and ACQG. In detail, the upper scores are 2.62 (Ec), 1.97(House), 2.46 (National), 2.29 (Military), 1.90 (Woman), 2.11 (Cars), 2.34 (Sports), 2.36 (Web), 2.27 (En), 2.45 (Others) respectively. HSKS pays attention to the main points of text, which filters some noisy sentences from verbose text, so the generated questions almost match texts well. HSMF takes enough features into account, which makes the ranking model effective, thus the top questions are more likely acceptable. ACGQ includes both improved components, so we can see the average score of ACQG by human evaluation is the highest. Therefore, ACQG works effectively by extracting key sentences instead of full text and improving ranking method.

We have studied the comments carefully and the responses to the reviewers’ comments are listed as follows:

**Reviewer 1#**

**Comment 1**: In Section 3.2, this paper lists some templates aiming at generating targeted questions. For example, there are *number* question; *rank* question; *if-cause* question; *relative-cause* question. The authors had better present some examples of each template for better understanding.

**Answer 1：**Based on the comment, we have listed some examples in Section 3.2, Page 6, Paragraph 2 as follows:

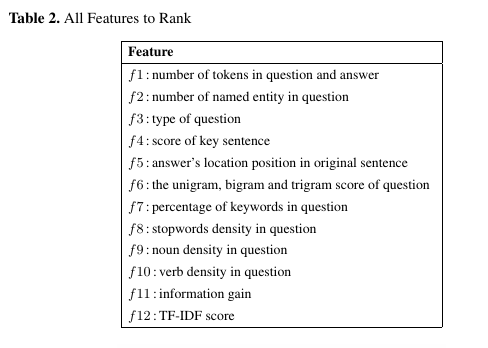
In order to have a better understanding of each template, we list some examples in Table.1.

Table 1. The Template Examples

|  |  |  |
| --- | --- | --- |
| Type | Sentence | Question |
| *who* | 刘佩英任安吉斯中国首席商务官。  (Peiying Liu acts as the China chief commercial officer of Angis.) | 谁任安吉斯中国首席商务官?  (Who acts as the China chief commercial officer of Angis?) |
| *when* | 2017年退休人员基本养老金上调。  (The basic pension of retirees will increase in 2017.) | 何时退休人员基本养老金上调？  (When will the basic pension of retires increase?) |
| *where* | 二战时日本曾在越南制造大饥荒。  (Japan had created a famine in Vietnam during World War II.) | 二战时日本曾在哪里制造大饥荒?  (Where had Japan created a famine during World War II?) |
| *rank* | 昆仑鸿星取关键一胜跃居东区第六位  (Kunlun Hongxing becomes the Sixth in the East by taking a key victory.) | 昆仑鸿星取关键一胜跃居东区第几位?  (What is the ranking of Kunlun Hongxing in the East by taking a key victory?) |
| *number* | 亚泰2000万镑交易震惊世界！  (The transaction of Yatay 20 million pound shocked the world.) | 亚泰多少镑交易震惊世界？  (What’s the amount of the transaction of Yatay shocked the world?) |
| *if-cause* | 因为人工智能快速发展，人们的生活方式得到巨大改变。  (Because of the rapid development of artificial intelligence, people’s lifestyle has been greatly changed.) | 人们的生活方式得到巨大改变，因为什么？  (What makes people's lifestyle have been change greatly?) |
| *relative-cause* | 春运将至，铁道部门虽然提前发售车票，但买票难的问题仍然存在。  (As Spring Festival is coming, the railway department has sold tickets in advance, but it is still difficult to buy a ticket. ) | 春运将至，铁道部门虽然提前发售车票，但怎么样？  (As Spring Festival is coming, what’s happened when the railway department has sold tickets in advance?) |
| *color* | 天气红色预警发布后,学校需要停课。 (After the weather red warning released, classes will be suspended.) | 什么颜色天气预警发布后，学校需要停课？(Classes will be suspended after which color warning released?) |

**Comment 2:** The paper utilizes a multiple features neural network and selects ten features to rank these generated questions. The authors had better design more features for better ranking results in the future.

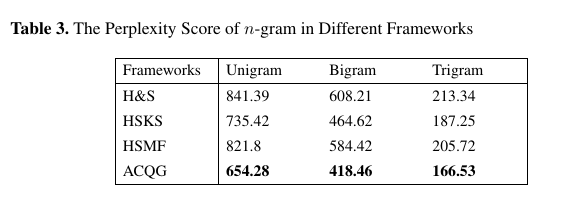
**Answer 2**: Based on the comment, we add two more features from text, one is information gain and the other one is TF-IDF score. Information gain is a probability data, which represents the uncertainty of sentence. TF-IDF score is a numerical statistic measure that is a popular sentence-weighting scheme. Those help represent the sentence well. From the results, the perplexity score decreases 2.4% (Unigram), 2.62% (Bigram) and 3.59% (Trigram), and the human rated score increases 2.15% than previous version. Those features really improve the ranking results. And the detail description is shown in Section 3.3, Page 6, Paragraph 24 as follows:



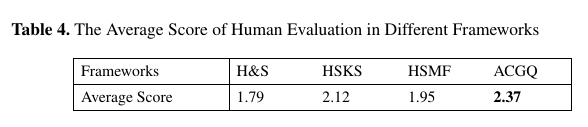
As shown in Table.2, we select twelve features about generated questions and original text. Additionally, the features f1~f5 are basic information of question and answer. f6 are the *n*-gram scores in question sequence, which provide rich contextual information. f7 is used to compute the keywords percentage in question. f8~f10 are different tags of word distribution in question, which indicate the structure of question sequence. f11 is a probability data, which represents the uncertainty of sentence. f12 is a numerical statistic measure that is a popular sentence-weighting scheme.

The analysis of results is shown in Section 4.2, Page 10 as follows:

Table.3 shows the perplexity score of *n*-gram in each framework. Unigram is n=1, Bigram is n=2 and Trigram is n=3. ACQG outperforms other frameworks, whose perplexity scores are 654.28 (Unigram), 418.46 (Bigram) and 166.53 (Trigram). H&S uses the over-generated questions and ranks them to get the output questions. The full text is used to generate questions, thereby the outputs are redundant and the scores are 841.39 (Unigram), 608.21 (Bigram) and 213.34 (Trigram). HSKS filters unmeaning sentences in texts, which is able to generate more targeted questions. Thus the scores decrease dramatically. Referring to HSMF system, HSMF is superior to H&S slightly. However the score of unigram is greater than H&S. The reason is that they both construct more absurd questions and the ranking method cannot improve this issue effectively. When we improve both key sentence extracting and question ranking, the performance of ACQG achieves best.



Furthermore, we introduce human evaluation into this work. The average marking scores are shown in Table.4, the detailed score distribution in each topic is shown in Fig. 3.



From Table.4, we can see ACGQ has the highest score 2.37, which indicates that the generated questions are around *Good* and *Borderline*. The questions are more likely to be acceptable. Because of the bad input and fragile ranking method, H&S has a poor performance, whose human evaluation score is only 1.79. For HSKS and HSMF, whose human evaluation score are 2.12 and 1.95 respectively. They both show better performance than H&S, but perform worse than ACGQ. Thus, the results of human evaluation are consistent with the above automatic evaluation.

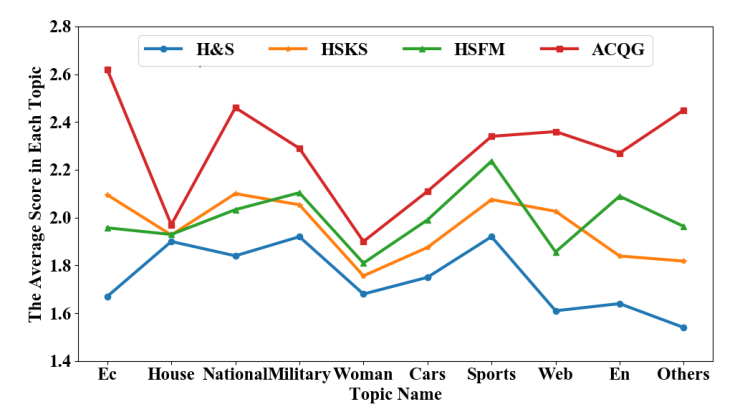


Fig.3. The Human Evaluation in Each Topic

Fig.3 shows the detailed human evaluation score distribution in each topic. We can observe that all frameworks have a low satisfaction score in the topic of House. The reason is that the collected news is almost advertisements in House, which cannot generate meaningful questions. For other topics, H&S always shows a lower score in each topic, while ACQG always shows a higher score in each topic. For other frameworks, the results fluctuate within H&S and ACQG. In detail, the upper scores are 2.62 (Ec), 1.97(House), 2.46 (National), 2.29 (Military), 1.90 (Woman), 2.11 (Cars), 2.34 (Sports), 2.36 (Web), 2.27 (En), 2.45 (Others) respectively. HSKS pays attention to the main points of text, which filters some noisy sentences from verbose text, so the generated questions almost match texts well. HSMF takes enough features into account, which makes the ranking model effective, thus the top questions are more likely acceptable. ACGQ includes both improved components, so we can see the average score of ACQG by human evaluation is the highest. Therefore, ACQG works effectively by extracting key sentences instead of full text and improving ranking method.

**Comment 3:** In Section 4, there is an important baseline method named H&S, which is the state-of-the-art in the area of Question Generation. The authors should describe more about the framework developed by Hellman and Smith for comparison in the related work.

**Answer 3:** Based on the comment, we have revised the related work in Section 2, Page 3, Paragraph 1 as follows:

Heilman and Smith [8] introduced an overgenerate-and-rank approach. Their framework can be viewed as a two-step process for question generation. In the first step, it transforms the input sentence into a simpler sentence, which is transformed into a more succinct question. In the second step, the declarative sentence is transformed into sets of questions by a sequence of well-defined syntactic and lexical transformations. It identifies the answer phrases which may be targets for WH-movement and converts them into question phrases.

**Reviewer 2#**

**Comment 1:** There are some grammar problems in this paper. For example, “reading comprehensive” should be “reading comprehension” on page 2.“comparing with the state-of-the-art systems” should be “compared to the state-of-the-art systems”. The authors should correct the syntax problems.

**Answer 1:** Based on the comment, we have fixed the grammar problems in this paper as follows:

In Section 2, Page 3, Paragraph 1: Mostow et.al [17] proposed a self-questioning strategy for reading comprehension, in this strategy, three templates are firstly built to generate questions about *what, how, why.*

In Section 1, Page 2, Paragraph 3: We conduct extensive experiments and the results show that our framework achieves a better performance compared to the state-of-the-art systems in terms of perplexity and human evaluation measurement.

**Comment 2:** In the future work, the authors had better design more templates to generate more meaningful questions

**Answer 2:** Based on the comment, we learn *color* question template to generate more targeted questions. A simple example is from sentence “天气红色预警发布后,学校需要停课。 (After the weather red warning released, classes will be suspended.)” to question “什么颜色天气预警发布后，学校需要停课？ (Classes will be suspended after which color warning released?)”. The total number of generated questions increases from 606 to 647 by 6.76%. The detail description is shown in Section 3.2, Page 6, Paragraph 1 as follows:

Secondly, we design more templates aiming at generating targeted questions. Parser tree describes the relationship among verb, subject, object, and so on. We utilize it to analyze each sentence, and edit some templates learning from text to generate targeted questions. The templates are listed as follows.

1. *Number* question: (QP < CD =number<CLP)
2. *rank* question: (QP < OD=number)
3. *if-cause* question : ((IP| PP=reason << 由于(because of) | 因为(because)) ..(IP| PP|VP)|<< (IP| PP |VP<<所以(so) |于是(so that)))
4. *relative-cause* question: ((( IP | PP=front <<虽然(though)) .IP=however )|<< (IP | PP=front(IP | PP | VP=however <<但是(but)|但(but))))
5. *color* question : (QP < NP=JJ < NP)

IP, PP, and VP represent different tag of words. A dot means subsequent follow and a left shaped arrow means an immediate subtree relation. These templates guide to search possible answer phrases on the parser tree. Once a subtree matches any template, a question will be constructed. In order to have a better understanding of each template, we list some examples in Table.1.

Table 1. The Template Examples

|  |  |  |
| --- | --- | --- |
| Type | Sentence | Question |
| *who* | 刘佩英任安吉斯中国首席商务官。  (Peiying Liu acts as the China chief commercial officer of Angis.) | 谁任安吉斯中国首席商务官?  (Who acts as the China chief commercial officer of Angis?) |
| *when* | 2017年退休人员基本养老金上调。  (The basic pension of retirees will increase in 2017.) | 何时退休人员基本养老金上调？  (When will the basic pension of retires increase?) |
| *where* | 二战时日本曾在越南制造大饥荒。  (Japan had created a famine in Vietnam during World War II.) | 二战时日本曾在哪里制造大饥荒?  (Where had Japan created a famine during World War II?) |
| *rank* | 昆仑鸿星取关键一胜跃居东区第六位  (Kunlun Hongxing becomes the Sixth in the East by taking a key victory.) | 昆仑鸿星取关键一胜跃居东区第几位?  (What is the ranking of Kunlun Hongxing in the East by taking a key victory?) |
| *number* | 亚泰2000万镑交易震惊世界！  (The transaction of Yatay 20 million pound shocked the world.) | 亚泰多少镑交易震惊世界？  (What’s the amount of the transaction of Yatay shocked the world?) |
| *if-cause* | 因为人工智能快速发展，人们的生活方式得到巨大改变。  (Because of the rapid development of artificial intelligence, people’s lifestyle has been greatly changed.) | 人们的生活方式得到巨大改变，因为什么？  (What makes people's lifestyle have been change greatly?) |
| *relative-cause* | 春运将至，铁道部门虽然提前发售车票，但买票难的问题仍然存在。  (As Spring Festival is coming, the railway department has sold tickets in advance, but it is still difficult to buy a ticket. ) | 春运将至，铁道部门虽然提前发售车票，但怎么样？  (As Spring Festival is coming, what’s happened when the railway department has sold tickets in advance?) |
| *color* | 天气红色预警发布后,学校需要停课。 (After the weather red warning released, classes will be suspended.) | 什么颜色天气预警发布后，学校需要停课？(Classes will be suspended after which color warning released?) |

**Comment 3:** The author should check the whole manuscript via professional English proof reading to fix the grammar issues.

**Answer 3:** Based on the comment, we have fixed the grammar issues in this paper.

**Comment 4:** The references should be cited as per the journal style.

**Answer 4:** Based on the comment, we have changed the references into the journal style.

**Comment 5：** Performance Comparison with recent studies and recent 2-3 methods in the context of the proposed work would be added.

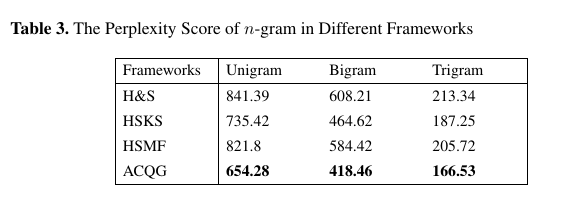
**Answer 5:** Based on the comment, in Section 4.1, Page 9, Paragraph 4, we introduce three recent methods including H&S, HSKS and HSMF for comparison as follows:

We introduce three recent methods including H&S, HSKS and HSMF for comparison to evaluate the performance of each framework. The compared methods are listed as follows:

* H&S: H&S transforms the input sentence into a simpler sentence and the simpler sentence is transformed into sets of questions by a sequence of well-defined syntactic and lexical transformations [8].
* HSKS: HSKS extracts the key sentences from texts and then constructs questions from the key sentences using templates.
* HSMF: HSMF applies a multi-feature model to rank all generated questions to make the top questions more useful.

The compared results are shown in Section 4.2, Page 10, Paragraph 2 as follows:

Table.3 shows the perplexity score of *n*-gram in each framework. Unigram is n=1, Bigram is n=2 and Trigram is n=3. ACQG outperforms other frameworks, whose perplexity scores are 654.28 (Unigram), 418.46 (Bigram) and 166.53 (Trigram). H&S uses the over-generated questions and ranks them to get the output questions. The full text is used to generate questions, thereby the outputs are redundant and the scores are 841.39 (Unigram), 608.21 (Bigram) and 213.34 (Trigram). HSKS filters unmeaning sentences in texts, which is able to generate more targeted questions. Thus the scores decrease dramatically. Referring to HSMF system, HSMF is superior to H&S slightly. However the score of unigram is greater than H&S. The reason is that they both construct more absurd questions and the ranking method cannot improve this issue effectively. When we improve both key sentence extracting and ranking methods, the performance of ACQG achieves best.



**Comment 6:** The research gap, open issues, contribution needs to be justified in section 2.

**Answer 6:** Based on the comments, we introduce the research gap, open issues and contribution in Section 2, Page 4, Paragraph 2 as follows:

To conclude, most of the existing question generation models are based on full texts, leading to generate some redundant questions. In addition, the fragile ranking methods do not take enough features into account and many useless questions are generated. In order to improve the performance of Chinese question generation, ACQG identifies the key sentences in texts and utilizes multiple features for question ranking.

**Comment 7:** Limitation and future scope of the work should be defined in conclusion section.

**Answer 7:** Based on the comments, we introduce the limitation of ACQG detailed and clarified the future work of this paper in Section 5, Page 12, Paragraph 3 as follows:

There still exist limitations in ACGQ. For example，the types of generated questions are limited, which is caused by the limited templates. In the future work, we will design more useful templates from texts to generate more meaningful questions. In addition, we will construct enough pairs of text and questions to train an end-to-end neural model to generate high-quality Chinese questions directly.

**Reviewer 3#**

**Comment 1:** The paper uses a Chinese dataset to evaluate the effectiveness of their design. In fact, the dataset is rare in Chinese question generation field and I hope the authors could make the dataset public for other researchers.

**Answer 1:** Based on the above comments, we have uploaded the Chinese dataset on the web for other researches. In Section 4.1, Page 9, Paragraph 1, the URL given as <https://github.com/hanjx16/question-data.git.>

**Comment 2:** The figures should be of higher quality and good representation. For example, the Fig.2 in Section 3.3 seems to be squashed.

**Answer 2:** Based on the comment, in Section 3.3, Page 8, we have replaced it with a high quality of figure as follows:

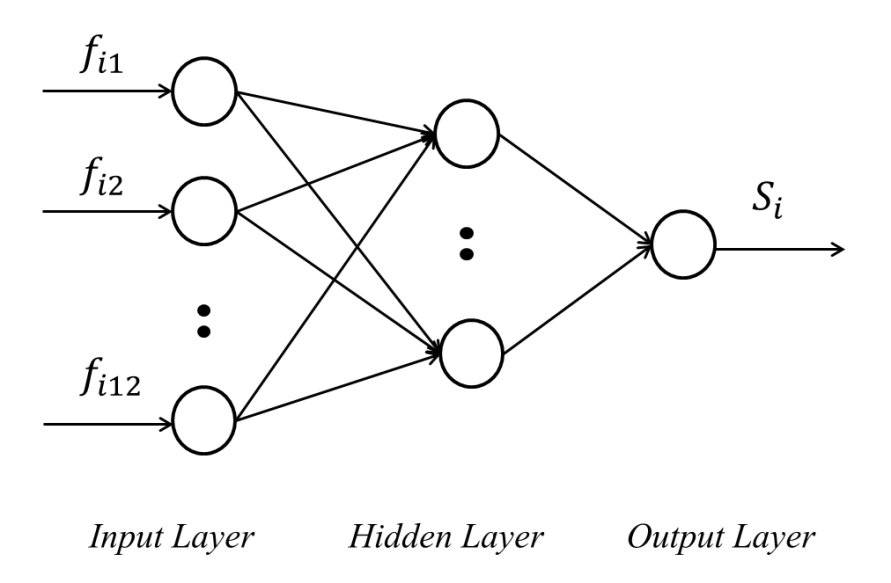


Fig.2. The Description of Multi-Feature Neural Ranking Model