We conducted two human evaluations on our model:

1. Human evaluation experiment 1

In human evaluation 1, we focus on evaluate the quality of the generated court view by our model and some compared algorithms. We use two metrics for this evaluation. Following many previous studies on text generation, we evaluated the fluency of generated court opinion. Because court view also needs to express the rationale correctly and cover the necessary facts in the fact description section, so we also evaluate the accuracy of the generated view. Both metrics use a score scale of 1 to 5, with 5 as the highest score.

We randomly select 200 cases from our test data set, and recruited four law school graduate students as the e valuators. The results are presented in Table 1, which shows the average scores of the four evaluators. We can see that our approach outperformed the compared ones.

Method	**accuracy**	**fluency**
BM25 + charge & law	3.84	4.92
Seq2seq-LSTM-attention + charge & law	4.17	4.11
SUMEXTABS	4.29	4.25
Our approach	**4.41**	**4.34**

Table 1. Human evaluation on accuracy and fluency of generated court opinion, based on the scale of 1-5, where 5 is the best

2. Human evaluation experiment 2

In this evaluation, we asked the users of the deployed system to give their experiences and opinions about the court view generation application. This survey is from the real users (individuals seeking legal advices, judges) that have used our tool for at least 1 month.

There are many questions in this survey, Table 2 show some result closely related to our model. Some observations from the original survey are: 91% of the users consider this tool useful (answers with score 4 & 5), and 62% of them consider it as very useful (score 5). 82% of them would recommend it to their friends (score 4 & 5).

Question dimension	**score (1-5)**
usefulness	4.62
easy to use	4.24
recommen it to your friends	4.38
content accuracy	4.32
content fluency	4.21

Table 2. some summary result from the user survey