Human Performance on Chinese Spatial Semantic Judgment

Anonymous ACL submission

Abstract

With the emergence of ChatGPT, large-scale language models seem to possess cognitive abilities similar to humans. This paper mainly focuses on the comparative analysis of human-machine testing on the task of judging the correctness and incorrectness of Chinese spatial semantics, including the process of human testing, the source, scale, and text characteristics of human testing data, and the comparison of human-machine testing accuracy, etc. By summarizing the typical features presented by human-machine on spatial semantic topics, this paper tries to analyze whether the machine has human-like spatial language understanding ability.

1 Introduction

Space category is an important basic category in human cognition, which mainly includes entity category, position category, and displacement category. The analysis of spatial semantics in the text has attracted much attention both in the field of linguistics and natural language processing, and the task of spatial semantic understanding is also one of the important contents of natural language processing evaluation. The Chinese spatial semantic understanding evaluation task is based on the first Chinese Spatial Semantic Understanding and Evaluation Task 2021 (Weidong et al., 2022). Human-machine testing is to evaluate humans and machines in roughly the same way for a specific task. Most of the testing tasks aim to compare the accuracy of humans and machines. However, few studies have further analyzed the results of the Human-machine test. For instance, CLUE (Xu et al., 2020), SuperGLUE (Nangia and Bowman, 2019), OCNLI (Hu et al., 2020), CMRC 2018 (Cui et al., 2018), SWAG (Zellers et al., 2018), ChID (Zheng et al., 2019), have compared human-machine performances, but only Chaz Firestone identified three factors contributing to the species-fairness of human–machine comparisons, extracted from his recent work, to encourage "species-fair" comparisons between humans and machines by the distinction between performance and competence (Firestone, 2020). This paper primarily focuses on the analysis of human-machine testing for the Chinese spatial semantic true-or-false judgment task[1]. Firstly, it provides an introduction to the process of human testing. Secondly, it presents the data source, scale, and text characteristics of human testing. The third section provides a detailed illustration of the comparison of human-machine performance. The fourth section analyzes the test questions and results through observations of the human-machine test. Finally, the fifth section presents the conclusion and prospects.

2 Human testing process and method

This test aimed to assess the correctness of Chinese spatial semantics, specifically whether there are any spatial semantic anomalies in given Chinese texts. Seven participants were recruited for this test, representing different grades and majors. All the annotators were undergraduate students. With a rate of 2 RMB per question, there were 100 questions in total, so each person spent an average of 200 RMB. The total cost was 1400 RMB. The marking process primarily took place through an online marking platform [2], and the testing procedure consisted of a training phase and a formal testing phase. Each test question, both during the

[1] We have enriched sentences containing spatial orientation information by replacing orientation words and other methods. These sentences include both correct and incorrect spatial information. We require machines and humans to judge whether anomalies exist in the spatial orientation of entities within the sentences. For instance, in Chinese, the phrase "跳进山洞外(jump outside of the cave), "jump into" must be paired with a component that expresses the interior location of a space, such as "in the cave, inside the cave". It cannot be paired with "outside the cave".

[2] http://www.nlp2030.com/

training and testing periods, was assigned to all seven participants simultaneously.

During the training stage, the labeling specifications were explained, and questions were addressed through video conferences. The training phase was divided into four rounds, continuing until the individual audit pass rate exceeded 80 percent. Once this criterion was met, participants were considered to have passed the training and were eligible to participate in the formal test. In the formal testing phase, each participant was required to complete 100 questions.

2.1 The Source and Proportion of Human Test Data

The human test set data comes from the SpaCE2022 test set [3], which contains 3152 sentences, including 1695 positive examples and 1457 negative examples. The human test task selected 100 sentences from the SpaCE2022 test set, including 50 positive and 50 negative examples. The corpus types covered eight types of corpus from many fields, including primary and secondary schools, sports training, human body movements, research papers, literature, People's Daily, Encyclopedia of Geography, traffic, and driving texts, as shown in Table 1.

2.2 Number of Replacement Pairs

We collected raw corpora from the aforementioned eight fields and performed data cleaning and spatial orientation word replacement to obtain a substantial amount of natural text corpora. We conducted a count of substitution pairs and selected 100 sentences, utilizing 80 substitution pairs with a total of 212 replacements. Among these sentences, 12 contained two substitution pairs, while 88 had one substitution pair. Table 2 presents the high-frequency substitution pairs along with their respective frequencies. Notably, the substitution pairs "上-下" (up-down), "里-中" (inside-middle), "当地-原地" (local-in place), "里-后" (inside-back), and "里-外" (inside-outside) exhibited higher substitution frequencies.

2.3 The characteristics of test questions

This section primarily analyzes the test questions based on the distribution of sentence lengths, stylistic balance, label balance, balance of replacement words, as well as the coverage and deviation of replacement words.

2.3.1 Sentence length distribution

Table 3 below displays the number of human test questions and their sentence length distribution for each corpus type. The maximum sentence length is 209, the minimum is 32, and the average sentence length is 115.

2.3.2 Balance analysis

The data balance analysis primarily encompasses style balance, label balance, and replacement word balance.

Figure 1 illustrates the distribution of part-of-speech (POS) tags for the replacement words in the human test task. It is observed that the proportion of original words and replacement words varies across different POS tags. In the human test task, orientation words (represented by letter "f") have the highest proportion, followed by locative words (represented by letter "f"). On the other hand, verbs and adverbs (represented by the letters "DV") have the lowest proportion.

2.4 Coverage of substitution word or substitution pair

The test questions for the SpaCE2022 spatial evaluation task were expanded from the original corpus using constructed substitution pairs. The substitution word list consists of a total of 705 pairs, with 425 pairs being utilized in the test set. For this human test task, there were 80 replacement pairs, 37 original words, and 42 replacement words, forming a total of 100 human test questions.

3 Comparison of the Human VS Machine Test

Table 4 shows the participating teams, their institutions, and the performance of their systems on the overall test set for this task.

3.1 Evaluation index

The data for the Chinese spatial semantic true-false judgment task consists of three components:

a. qid: test question number;

b. context: the content of the text material to be evaluated;

c. judge: the judgment result indicating the correctness of the spatial semantics in the material (0 represents a negative example indicating the presence of a spatial semantic anomaly, while

[3]Table12 is the total dataset scale in Appendix

Corpus type	Number of positive questions	Number of positive questions(Human-test)	Number of negative questions	Number of negative questions(Human-test)
A(chinesebook)	206	6	318	11
B(sports)	248	7	160	5
C(rmrb20-21)	478	14	541	19
D(literature)	202	6	189	6
E(geography)	95	3	32	1
F(traffic)	407	12	153	5
G(article)	17	1	16	1
H(973srl)	42	1	48	2
Total number	1695	50	1457	50

Table 1: Number of SpaCE2022 test sets and number and percentage of human tests

substitution pair	Quantity
上→ 下(Up → down)	8
里→ 中(Inside → middle)	6
当地→ 原地(Local → in situ)	6
里→ 后(Inside → behind)	6
里→ 外(Inside → outside	6
中→ 前(Middle → Front)	6
中→ 外(Middle → outer)	6
过来→ 过去(Comeover → over)	4
里→ 上(Inside → up)	4
后→ 下(Back → Down)	4
北→ 东(North → East)	4
北→ 西(North → West)	4
回来→ 过来(Comeback → come here)	4
前→ 外(Front → outside)	4
上→ 里(Up → inside)	4
上→ 中(Top → middle)	4
西→ 北(West → North)	4
中→ 上(Middle → upper)	4

Table 2: High-frequency substitution pairs

Corpus type	Quantity	Minimum sentencelength	Maximum sentencelength	Average sentencelength
A(chinesebook)	17	32	175	106
B(sports)	12	42	128	73
C(rmrb20-21)	33	40	209	111
D(literature)	12	48	235	149
E(geography)	4	125	227	171
F(traffic)	17	55	217	140
G(article)	2	22	43	33
H(973srl)	3	51	127	90

Table 3: Sentence Length Distribution of Human Test Questions in Different Corpus Types

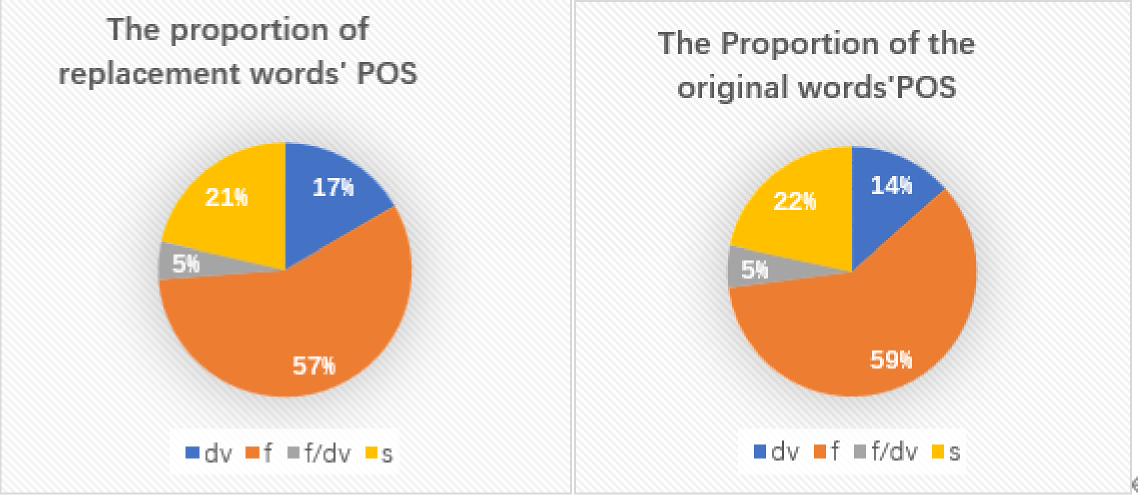

Figure 1: The proportion of the POS of the original words and replacement words

Team	Institution	Approach or Model	Accuracy(Acc)
Weihang	Suzhou University	Constructed a spatial word dictionary; Used Electra model (Clark et al., 2020) to judge the probability of each spatial word being a replacement word; Used the mean of the maximum spatial word replacement probability and [CLS] jointly as classification basis.	**0.7992**
CPIC	Taiji Technology	PTM->[CLS]->Binary classification with Sigmoid; Selected Electra model (Clark et al., 2020).	0.7985
NoMercy	Fudan University	[CLS] followed by classification head; Voting of 5 models	0.7865
Baseline	Peking University	BERT ->Linear classification layer (Kenton and Toutanova, 2019)	0.5864

Table 4: Participant Systems and Their Performance

1 represents a positive example indicating the absence of a spatial semantic anomaly).

The evaluation metrics for this task are as follows.

Accuracy(Acc) = number of questions with correct answers / total number of questions

3.2 Accuracy comparison

The comparison of human-machine test accuracy is illustrated in the following tables. Table 5 and Table 6 demonstrate that the average human accuracy for the task is 0.01 lower than that of the machine. The highest machine accuracy is 0.84, while the lowest is 0.71. On the other hand, the highest human accuracy is 0.95, while the lowest is 0.69. It is notable that the highest human accuracy surpasses the highest machine accuracy, while the lowest human accuracy falls below the lowest machine accuracy.

3.3 Comparison of consistency rate of human-machine test

In this paper, the Kappa value of human subjects was calculated through pairwise comparison of their answers, representing the human consistency rate. The results are presented in Table 7. Similarly, the Kappa value for the machine was obtained by

Machine UserID	Accuracy
M1	0.82
M2	0.82
M3	0.84
M4	0.71
Baseline	0.60
Average	0.80
Standard deviation	0.06

Table 5: Machine Accuracy

Human UserID	Accuracy
13	0.74
15	0.78
16	0.83
17	0.69
18	0.95
19	0.8
20	0.69
Average	0.78
Standard deviation	0.09

Table 6: Human Accuracy

comparing its answers in pairs, serving as the machine consistency rate. The corresponding results are displayed in Table 8.

4 Analysis of human vs machine test characteristics based on test results

This paper primarily analyzed the test results from the following perspectives: the influence of corpus type, performance of substitution words, and characteristics of items with high or low human-machine consistency. Additionally, we conducted further analysis on the results of the human-machine test by establishing selection criteria to categorize the test questions. The test questions were classified into four situations: both humans and machines performed well, both humans and machines performed poorly, only machines performed well, and machines performed poorly while humans performed well. Table 11 illustrates these findings, revealing that a significant proportion of test questions demonstrated good performance by both humans and machines, while a small proportion exhibited poor performance.

4.1 The Relationship between Test Results and Corpus Type

By analyzing the proportions of the four types of questions in each corpus type, it is evident that the results of the human-machine test are influenced by the type of corpus. Examining the proportion of questions with good performance in each corpus type, it is observed that Chinese textbooks and H973srl have a relatively large proportion, indicating better performance by both humans and machines in these corpus types. Conversely, the sports action, geographical encyclopedia, traffic driving, and other corpus types have a relatively small proportion of good performance, suggesting poorer performance by both humans and machines in these types.

Further examination of questions with good machine performance and poor human performance reveals that two of them belong to the geographical encyclopedia category, accounting for 50% of all geographical encyclopedia questions tested. Additionally, eight questions fall under the traffic category, representing 57% of the 14 questions with good machine performance and poor human performance, and 47% of all traffic corpus tested. This indicates that humans do not perform well in the geographical encyclopedia and traffic corpus types.

Regarding questions with poor machine performance and good human performance, it is found that the machine performs poorly in the People's Daily and traffic corpus. In summary, both humans and machines demonstrate subpar performance in the sports action, geographical encyclopedia, and traffic domains, particularly in the geographical and traffic types. Additionally, the machine's performance in the People's Daily corpus is also unsatisfactory. The total number of human and machine test questions is 100. Out of these, there are 41 questions with an average score of both humans and machines above 0.8, indicating good performance by both humans and machines. These questions account for 41% of the total. Additionally, there are 25 questions where both humans and machines provided correct answers (with an average score of 1), accounting for 25% of the total. Among these questions, humans answered correctly in 31 cases, while machines answered correctly in 32 cases.

On the other hand, there are 7 questions where the average score of both humans and machines is below 0.6, indicating poor performance. These questions account for 7% of the total. Among them, there are 14 questions where the human average score is below 0.6 and the machine average score is higher than 0.8, indicating good machine performance but poor human performance. Additionally, there are 14 questions where the machine average score is below 0.6 and the human average score is higher than 0.8, indicating good human performance but poor machine performance.

When considering the agreement of both the average scores and the answers, the human agreement rate is slightly lower than that of the machine. By analyzing the proportion of human subjects with the same average score and the same answers, it is found that the consistency rate of humans is higher in Chinese textbooks, while it is lower in traffic driving, geographic encyclopedia corpus, and sports action texts.

Regarding the machine, it exhibits a high consistency rate in the geographical encyclopedia corpus but a poor consistency rate in literature, as observed by examining the proportion of questions with consistent machine answers in each corpus.

4.2 Regularity of replacement words

Observing the items in which both humans and machines perform well, the substitution pairs of "中-前(zhong-qian, middle-before)", "里-后(li-hou,inside-after)" and "里-外(li-wai,inside-

User	Human1	Human2	Human3	Human4	Human5	Human6	Human7
Human1		0.52	0.37	0.42	0.41	0.30	0.32
Human2	0.32		0.48	0.30	0.59	0.42	0.52
Human3	0.42	0.37		0.45	0.49	0.37	0.37
Human4	0.49	0.56	0.76		0.56	0.48	0.42
Human5	0.23	0.33	0.42	0.23			0.41
Human6	0.57	0.51	0.57	0.33	0.76	0.30	
Human7	0.53	0.53	0.51	0.42	0.45	0.59	0.30

Table 7: Kappa values of human subjects

User	Machine1	Machine2	Machine3	Machine4	Machine5
Machine1		0.69	0.71	0.48	0.23
Machine2	0.69		0.44	0.41	0.34
Machine3	0.71	0.67		0.44	0.24
Machine4	0.48	0.41	0.23		0.25
Machine5	0.25	0.24	0.67	0.34	
Average	0.53	0.50	0.51	0.42	0.26

Table 8: Kappa Values of Machine Subjects

corpus type	Number of specific corpus	Select criteria
Both humans and machines perform well	41	The average value of humans and machines is above 0.8.
Both humans and machines perform poorly.	7	The average value of humans and machines is below 0.6.
Machines perform well, and humans perform poorly.	14	The human average is below 0.6, and the machines' average is higher than 0.8
Humans perform well, and machines perform poorly.	14	The machine average is below 0.6, and the humans' average is higher than 0.8

Table 9: Number of corpus types and the selection criteria

outside)" appeared three times respectively, among which "外(Wai, outside)" and "前(Qian, before)" appeared four times, "下(xia, below)" appeared three times, and "后(Hou, after)", "边(bian, side)", "外(Wai, outside)", "中(Zhong, middle)" and "里(Li, inside)" appeared two times respectively. Humans and machines performed well in these substitution pairs, so we evaluated the performance of the humans and machines in the substitution pairs of "中-前(zhong-qian, middle-before)", "里-后(li-hou, inside-after)" and "里-外(li-wai,inside-outside)", and found that the performance of both humans and machines was very good in these four groups of substitution pairs, and the machine performance was slightly better than that of the humans in the substitution pairs of "里-后(li-hou, inside-after)" and "里-外(li-wai,inside-outside)". By observing the substitution pairs of 14 questions in which the machine performed well and the human performed poorly, The substitution word "one side" appeared three times, and this substitution word appeared three times on all test questions (100 questions), indicating that the human did not perform well on this substitution word and was not as good as the machine.

4.3 Influencing factors of consistency rate

Out of the total 100 test questions, there are 25 questions where both humans and machines provide correct answers with an average score of 1. These questions account for 25% of the total. Among these questions, humans answered correctly in 34 cases, accounting for 34%, while machines answered correctly in 41 cases, accounting for 41

Furthermore, there are 25 questions where humans and machines have the same answers, accounting for 25% of the total. Among these questions, humans provided the same answers in 35 cases, accounting for 35%, while machines provided consistent answers in 43 cases, accounting for 43%.

By examining the 25 questions with consistent answers from both humans and machines, we found that all of these questions exhibit abnormal spatial semantics, as shown in Table 10. Notably, the syntactic distribution of "n + f + v" is prevalent among these questions. The proportion of "NP + VP + pp" is 36%, "NP + VP" accounts for 24%, and "pp + VP" represents 16% of the total, as shown in Table 10. This syntactic pattern is closely associated with spatial orientation and is easier for both humans and machines to capture.

When analyzing the questions with inconsistent answers, we speculate that humans may be influenced by the following factors in judging the abnormality of spatial semantics: ①Their existing cognitive experiences; ②Their judgment of spatial semantics abnormality based on constructed spatial scenes triggered by cognitive experiences, involving spatial entities and positions. In cases where there are multiple physically viable spatial entities that are contextually acceptable, it is more likely for human-machine answers to be inconsistent. Examples of questions with inconsistent human-machine answers are presented in Table 11.

4.4 Limitaions

This study has not yet considered the performance of large language models on this task. In the future, we will further evaluate the performance of large language models in comparison to humans based on this preliminary work.

5 Summary and Outlook

This paper presents an analysis of the human-machine test from the perspective of evaluating human-machine cognitive abilities. It provides a comprehensive summary of the Chinese spatial semantic true-false judgment and evaluation, including details on the human test process, test data source, test scale, and text features. The analysis of the test results explores the relationship between the results and corpus types, as well as the characteristics of replacement words. The findings reveal that machines achieve slightly higher accuracy than humans, and the performance in the human vs. machine test is influenced by the type of corpus.

Additionally, the study investigates the performance of humans and machines on different substitution pairs, highlighting that humans are less proficient than machines when dealing with "one side" substitution pairs. Based on these findings, it can be inferred that machines possess spatial language understanding abilities similar to humans.

The research also suggests that when judging the abnormality of spatial semantics, humans may be influenced by factors such as their existing cognitive experiences and the construction of spatial scenes triggered by spatial entities and positions described in the text. Future research will focus on further validating these hypotheses and exploring additional factors that may affect the construction of spatial scenes by humans and machines using textual information.

References

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.

Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Nikita Nangia and Samuel R Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. *arXiv preprint arXiv:1905.10425*.

Zhan Weidong, Sun Chunhui, Qin Ziwei, Yue Pengxue, and Tang Gantong. 2022. Development of space2021 data set: a new idea for task design of spatial semantic understanding assessment. *Language Application*, (2):12.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*.

qid	Replacement pairs	Syntactic distribution	Examples
1-test-12615	里-前(Inside-Front)	N+f+vp	食堂前吃了早饭I had breakfast in front of the canteen.
1-test-12852	里-外(Inside-outside)	N+f+vp	口袋外都寻了个遍I searched all over the outside of my pocket.
1-test-12862	后-上(Back-Up)	Np+v+f	四持黑红棍者在上People who hold black and red sticks are on the top.
1-test-12893	上-下(Up-down)	把+n+v+p+n+f Put+n+V+p+n+f	把头靠在门下Put your head under the door.
1-test-13022	里-外(Inside-outside)	N+v+p+n+f	村子都浸在桂花的香气外The village is immersed in the fragrance of osmanthus.
1-test-13076	里-后(Inside-behind)	N+v+p+n+f	信应该丢在邮筒后The letter should be left behind the mailbox.
1-test-13151	上-旁(Top-Side)	N+n+p+n+f	头部躯干在一条直线上Head and torso in a straight line
1-test-13188	右侧-左侧(Right-left)	N+v+f+n	双臂伸直左侧手臂Straighten your arms and left arm.
1-test-13258	侧面-里面(Side-Inside)	N+v+p+n+f	双臂分别放置于耳朵里面Place your arms inside your ears.
1-test-13607	后-边(Back-side)	null	达到入库条件边Reach the edge of warehousing condition
1-test-13726	中-下(Middle-lower)	N+p+n+f+v	利息预先在本金下扣除的Interest is deducted from the principal in advance.
1-test-13890	内-外 上-后(Inner-outer upper-rear)	N+p+v+np	店外为了答谢附近居民的支持Outside the store to thank nearby residents for their support.
1-test-13939	中-前(Middle-Front)	N+v+p+np	我沉睡在乡音的呓语I sleep in the babble of the local accent.
1-test-14226	中-外前-内(Middle-outer front-inner)	N+p+f+v	我往内跑I ran inside.
1-test-14263	中-旁(Middle-Side)	null	为群众办实事旁深化学习成果Do practical things for the masses and deepen the learning results
1-test-14351	中-前(Middle-Front)	N+vp+p+v+f+v	老人靠抓住铁门在水前保命The old man saved his life by holding on to the iron gate before the water.
1-test-14469	前面-下面(Front-below)	N+p+n+vp	妻子在下面拉车The wife is pulling the cart below.
1-test-14510	东-北(East-North)	N+p+f+p+f	太阳从北到西The sun goes from north to west.
1-test-14714	里-后(Inside-behind)	N+f+的+n+vp + n + VP of N + f +	饭店后的人全扭头看着The people behind the restaurant all turned to look.
1-test-14781	外面-里面(Outside-Inside)	N+n+v+p+f	衣服口袋全挂在里面The pockets of the clothes are all hanging inside.
1-test-14986	中-前(Middle-Front)	N+v+n+f	河流很快没入塔克拉玛干沙漠前The river soon sank into the Taklimakan Desert.
1-test-15639	左边-两边(Left-both sides)	N+p+n的+f + f of N + p + n	人民大会堂在我的两边The Great Hall of the People is on either side of me.
1-test-15681	上-下(Up-down)	P+n+f+vp	在鞋上套下塑料套Put the plastic cover on the shoe.
1-test-15737	后-中 上-前(Back-upper middle-front)	P+n+vp	在御碟摆出肉花Put out the meat flower on the imperial dish.

Table 10: Syntactic Distribution of Questions with Consistent human-machine Answers

qid	Replacement pairs	Examples	The numbers of human-machine answers that are abnormal to normal
1-test-14785	下来-上去(Down-Up)	左手将宋钢的脑袋按上去(Press Song Gang’s head with your left hand.)	3：4
1-test-15516	北-西(North-West)	电动自行车向西左转弯穿越控江路(The electric bicycle turns left to the west and crosses Kongjiang Road.)	5：2
1-test-13837	里-下(Inside-down)	大楼下有周雪群等护理员24小时守候(Under the building, Zhou Xuequn and other nurses are waiting for 24 hours.)	4：3
1-test-13843	两侧-下侧(Both sides-lower side)	山谷下侧裸露着光秃秃灰里带红的岩石。(The underside of the valley was bare with reddish grey rocks.)	4：3
1-test-15633	后面-左面(Back-Left)	他们还在战车左面拖着伐下的树枝(They also dragged the fallen branches on the left side of the chariot.)	2：5
1-test-12621	上-中(Top-middle)	在无边的旷野中，在凛冽的天宇下(In the boundless wilderness, under the cold sky)	3: 4
1-test-12640	下面-上面(Below-above)	悬崖上面的大地越来越暗(The earth above the cliff is getting darker and darker.)	6：1
1-test-12744	回来-过去(Come back-past)	超声波遇到障碍物就反射过去(Ultrasonic waves bounce off obstacles.)	6：1
1-test-13308	后-下(Back-Down)	右腿顺势下滑(The right leg slides down)	2：5
1-test-13367	内-边(Inside-side)	后脑勺紧靠在十指相扣的双掌边(The back of the head is close to the side of the palm with the fingers interlocked.)	2：5
1-test-13563	身边-身后(Side-behind)	游客顺着浮毯滑到他的身后(The tourist slid down the floating carpet behind him.)	2：5
1-test-14013	前-外(Front-outside)	站在红火建设的新门诊楼外，李开文笑得开心。(Standing outside the new outpatient building, Li Kaiwen smiled happily.)	2：5
1-test-14036	当地-原地(Local-in situ)	据原地居民介绍(According to the local residents,)	5：2
1-test-14135	乡下-城里(Country-City)	城里的农具和民俗也随主人到了城里(The farm tools and folk customs in the city also came to the city with their masters.)	6：1
1-test-14176	上-下(Up-down)	那雾就围在中梁子山的脖子下(The fog is around the neck of Zhongliangzi Mountain.)	2：5
1-test-14231	下-前(Lower-Front)	唐宗秀坐在墙根前(Tang Zongxiu sat in front of the wall.)	3：4
1-test-14247	市区-乡下(City-country)	乡下居民步行15分钟就可到达一处公园(A park is a 15-minute walk for country dwellers.)	3：4
1-test-14442	去-来(Go-come)	初中毕业来江西读高中(Graduated from junior high school and came to Jiangxi to study in senior high school)	2：5
1-test-14546	过来-过去(Come over-over)	武莉开车过去后发现广场附近的公共场所入口处的拦车杆高高竖起(After Wu Li drove over, she found that the car stop pole at the entrance of the public place near the square was erected high.)	2：5
1-test-14736	前-外(Front-outside)	她跪在车站外的地上(She knelt on the ground outside the station.)	3：4
1-test-14752	回来-过来(Come back-come here)	宋钢在读着林红的纸条时一直害怕李光头会过来(Song Gang was always afraid that Li Baldy would come when he read Lin Hong’s note.)	3：4

Table 11: Examples of Some Questions with Inconsistent Human-machine Answers

[HTML]4472C4 **Set**	**Positive Numbers**	**Negative Numbers**	**Positive/Negative ratio**
Full dataset	5,077	10,670	0.48
Training set	2,677	8,316	0.32
Validation set	705	897	0.79
Test set	1,695	1,457	1.16

Table 12: The total dataset Scale