An Overview of the 2019 Language and Intelligence Challenge

Quan Wang, Wenquan Wu, Yabing Shi, Hongyu Li, Zhen Guo, Wei He, Hongyu Liu, Ying Chen, Yajuan Lyu, and Hua Wu

Baidu Inc., Beijing, China
{wangquan05,wuwenquan01,shiyabing01,lihongyu04,guozhenguozhen,hewei23, liuhongyu02,chenying04,lvyajuan,wu_hua}@baidu.com

Abstract. This paper provides an overview of the 2019 Language and Intelligence Challenge (LIC 2019), which assesses the ability of machines to understand language and use language to interact with humans. The challenge comprised three tasks: information extraction (IE), knowledge-driven dialogue, and machine reading comprehension (MRC), all providing large-scale Chinese datasets and open-source baseline systems. There were 2,376 teams that took part in the challenge, with a total of 6,212 system runs submitted. The participating systems performed quite well, offering a 21.65% increase over the baseline in IE, a 37.40% increase in the dialogue task, and a 34.09% increase in MRC.

Keywords: Language understanding and interaction · Information extraction · Knowledge-driven dialogue · Machine reading comprehension

1 Introduction

Language is the most important medium for communication in human life. Building machines that could understand language and use it to interact with humans is a central goal of artificial intelligence. Towards this goal, the China Computer Federation(CCF), Chinese Information Processing Society of China(CIPS), and Baidu Inc. jointly organized the 2019 Language and Intelligence Challenge (LIC 2019). The challenge assesses the ability of machines to understand natural language text, automatically extract knowledge from it, and use the learned knowledge to answer questions or hold conversations with humans.

LIC 2019 was set up with three tasks: (i) *Information Extraction* that requires systems to automatically extract structured knowledge from natural language text; (ii) *Knowledge-driven Dialogue* that requires systems to have conversations with humans based upon a given knowledge graph; and (iii) *Machine Reading Comprehension* that requires systems to read natural language text and answer questions about the given text. All the three tasks provided large-scale Chinese datasets, as well as baseline systems implemented in PaddlePaddle.[1]

There were 2,376 teams that took part in the challenge, with a total of 6,212 system runs submitted. About 60% of the participants came from universities

[1] http://paddlepaddle.org

and research institutes at home and abroad, and the other 40% came from over 300 enterprises. The results of participating systems are promising. Compared against the baselines, the top-1 system performs 21.65% better in the information extraction task, 37.40% better in the knowledge-driven dialogue task, and 34.09% better in the machine reading comprehension task.

LIC 2019 has greatly advanced the technical development of natural language understanding and interaction. The infrastructure, including the datasets, baseline systems, and evaluation mechanisms, has been made publicly available to provide a good basis for future research in related areas.[2,3,4]

2 Tasks

LIC 2019 comprised three tasks: Information Extraction (IE), Knowledge-driven Dialogue, and Machine Reading Comprehension (MRC), detailed as follows.

2.1 Information Extraction

Information Extraction is to let machines automatically extract structured knowledge such as entities, attributes and relations from unstructured or semi-structured text. It is an important foundation for artificial intelligent application such as information retrieval, intelligent question answering, and intelligent dialogue, and has been widely concerned by the industry. The task provided a large-scale high-quality manually annotated information extraction dataset, DuIE, which aims to promote the development of information extraction technology.

Task Definition Given a sentence *sent* and a list of pre-defined schemas which define relation P and its corresponding classes of subject S and object O. The participant system is required to output all correct triples mentioned in *sent* under the constraints of given schemas.

Dataset DuIE dataset is the largest open-domain Chinese information extraction dataset, containing more than 450,000 instances in over 210,000 real-world Chinese sentences, bounded by a pre-specified schema with 49 predicates.

DuIE is generated by Baidu Baike and Baidu News Feeds as corpus to align with Baidu Baike Infobox as KB. Each sample in DuIE contains one sentence and a set of associated triples mentioned in the sentence. The dataset is divided into a training set (170k sentences), a development set (20k sentences) and a testing set (20k sentences). The training set and the development set were used for training and validating the model, the testing set was used for participants to submit the prediction result and used as the final evaluation for ranking. We further added 80k sentences as pseudo noises to rule out tuning against the test set. Table 1 provides the statistics of the dataset in detail.

[2] `http://ai.baidu.com/broad/subordinate?dataset=sked`
[3] `https://ai.baidu.com/broad/introduction?dataset=duconv`
[4] `https://ai.baidu.com//broad/introduction?dataset=dureader`

Table 1. Statistics of the information extraction dataset.

Dataset	Total Amount	Training Set	Dev Set	Testing Set
#Sentence	214,739	173,108	21,639	19,992
#Instance	458,184	364,218	45,577	48,389

Baseline System This task provided participants with an open source baseline system implemented in PaddlePaddle.[5] Baseline system separates this task into a pipelined architecture with relation classification and subject-object labeling, which significantly improved the performance of the extraction model.

Evaluation Metrics Standard *Precision*, *Recall* and $F1$ score were adopted as the basic evaluation metrics to evaluate the performance of participating systems, while the final grade is ranked according to the F1 value. A triple predicted by participant systems will be regarded as correct when its relation and two corresponding entities are both exactly matched with the true triple annotated on the testing set. In addition, considering the cases of alias in sentences, we used a dictionary of entity alias in Baidu Knowledge Graph in the evaluation.

2.2 Knowledge-driven Dialogue

Human-machine conversation is an important topic in AI and has received much attention in recent years. Currently dialogue system is still in its infancy, which usually converses passively and utters their words more as a matter of response rather than on their own initiatives, which is different from human-human conversation. Thus we set up a new conversation task, named knowledge-driven dialogue, where machines converse with humans based on a built knowledge graph (KG). It aims at testing machines' ability to conduct human-like conversations.

Task Definition Given a dialogue goal G and a set of topic-related background knowledge $M = f_1, f_2, \cdots, f_n$, a participating system is expected to output an utterance u_t for the current conversation $H = u_1, u_2, \cdots, u_{t-1}$, which keeps the conversation coherent and informative under the guidance of the given goal. During the dialogue, a participating system is required to proactively lead the conversation from one topic to another. The dialog goal G is given like this: "*[start]* $\rightarrow$ *topic_a* $\rightarrow$ *topic_b*", which means the machine should lead the conversation from any start state to "*topic_a*" and then to "*topic_b*". The given background knowledge includes knowledge related to "*topic_a*" and "*topic_b*", and the relations between these two topics.

Dataset We created a new dataset named DuConv [7]. The background knowledge provided in the dataset was collected from MTime.com[6], which records

[5] `https://github.com/baidu/information-extraction`

[6] `http://www.mtime.com`

the information of films and stars, such as box offices, directors, reviews, etc. We constructed a KG with collected knowledge organized as triplets {Subject, Predicate, Object}, where objects can be factoid facts and non-factoid sentences such as comments and synopsis. Table 2(a) lists the statistics of our KG.

Given the KG, we sampled some knowledge paths, used as conversation goals. Specifically, we focused on the simple but challenging scenario: naturally shifting the topics twice, i.e., from "*[start]*" state to "*topic_a*" then finally to "*topic_b*". We sampled two linked entities in our KG as '*topic_a*" and "*topic_b*" to construct the knowledge path. About 30k different knowledge paths were sampled and used as conversation goals for knowledge-driven conversation crowdsourcing.

Table 2. Overview of DuConv.

(a) Statistics of Knowledge

# entities	143627
# movies	91874
# person names	51753
# properties	45
# spo	3598246
average # spo per entity	25

(b) Statistics of Dialogues

# dialogs	29858
# utterances	270399
average # utterances per dialog	9.1
average # words per utterance	10.6
average # words per dialog	96.2
average # knowledge per dialog	17.1

Unlike using self-play in dataset construction [1], we collected lots of crowdsourced workers to generate the dialogues in DuConv [7]. For each given conversation goal, we assigned two workers different roles: 1) the conversation leader and 2) the follower. The leader was provided with the conversation goal and its related background knowledge in our knowledge graph, and then asked to naturally shift the conversation topic following the given conversation goal. The follower was provided with nothing but the dialogue history and only had to respond to the leader. The dialogue will not stop until the leader achieves the conversation goal. We recorded conversation utterances together with the related knowledge triplets and the knowledge path, to construct the whole dataset of DuConv. Table 2(b) summarizes the main information about DuConv. The data was divided into training, development, and test sets by 80%, 10%,10%.

Baseline System This task provided participants with two open-sourced baseline systems[8]: retrieval-based and generation-based systems, implemented by PaddlePaddle. To enable dialogue systems to converse with external background knowledge, the baseline systems were incorporated an external memory module for storing all related knowledge, making the models select appropriate knowledge to enable proactive conversations [7]. Our baseline systems can make full use of related knowledge to generate more diverse multi-turn conversations.

[7]The workers were collected from a Chinese crowdsourcing platform `http://test.baidu.com/`. The workers were paid 2.5 Chinese Yuan per conversation.

[8] `https://github.com/baidu/knowledge-driven-dialogue`

Evaluation Metrics The participating systems were tested under two settings: 1) automatic evaluation and 2) human evaluation. For automatic evaluation, in addition to BLEU1/2 and DISTINCT1/2 which measure the relevance and diversity, we also used F1 to measure the char-based F-score of output utterance against reference utterance. The total score of F1 and BLEU1/2 was used for ranking participating systems. DISTINCT1/2 were used as auxiliary metrics.

The top 10 systems in automatic evaluation phrase were further evaluated by human on dialogue-level goal completion and coherence. Firstly, each system was required to converse with human to generate multi-turn dialogue given a conversation goal and the related knowledge. For each system, 100 dialogues were generated. Then the generated dialogues were manually evaluated to measure the goal completion and coherence. Goal completion has three grades: "0" means that the goal is not achieved, "1" the goal is achieved by making minor use of knowledge, and "2" the goal is achieved with full use of knowledge. Coherence has four grades: bad(0), fair(1), good(2) and perfect(3). The total score of normalized goal completion and coherence was used for the final ranking.

2.3 Machine Reading Comprehension

This task requires machines to read natural language text and answer questions about the given text. It is a crucial task in language understanding and also an important component of human-machine interaction. Last year, CCF, CIPS, and Baidu Inc. jointly organized the 2018 NLP Challenge on MRC, and the winning systems could answer more than 75% of the questions correctly [4]. LIC 2019 continued to set up the task, focusing on difficult questions that current systems fail to answer correctly.

Task Definition Given a question Q and a set of documents $\mathcal{D} = \{d_1, d_2, \cdots, d_n\}$, the participating system is required to output an answer A that best answers Q based on knowledge from $\mathcal{D}$.

Dataset The data was collected from DuReader [2], a large-scale, open-domain Chinese MRC dataset. In DuReader, all questions are sampled from anonymized user queries submitted to Baidu Search. Each question gets five documents collected from search results of Baidu Search and Baidu Zhidao, from which answers are manually generated. Questions are further divided into three types: *Description*, *Entity*, and *YesNo*. Entity answers (a single entity or a list of entities) and opinion answers (affirmation or negation) are further provided for Entity and YesNo questions, respectively. See [2] for a detailed description of DuReader.

The data was divided into training, development, and test sets. The training set, consisting of all training examples from DuReader, is the same as that used in the 2018 contest. The development and test sets consist of difficult questions that the winning systems in the 2018 contest failed to answer correctly. Specifically, for each question, we compared the quality of its system answers against human answers. Questions on which the ROUGE-L [3] score of the former lags behind

that of the latter by 10 points or more were taken as difficult ones. For questions in the test set, we computed the average ROUGE-L of the top 6 winning systems in last year's contest. For questions in the development set where answers of the top 6 winning systems were not available, we picked an internal system used at Baidu which could rank among the top 3 in last year's contest and computed its ROUGE-L. In this way, we selected 3,311 questions from the original DuReader development set, and 8,996 questions from the original test set.

We further used manual annotation to judge whether the selected questions fit the bill, i.e., answerable by humans but difficult for machines. Given a question selected, we provided to an annotator a human answer and a system answer at the same time. The system answer was the one with the lowest ROUGE-L among the top 6 winning systems on the test set, and the one generated by the internal system on the development set. The annotator was then asked to judge whether there is a gap between the quality of the two answers, and to return one of the following labels: (i) the human answer is better; (ii) the system answer is better; and (iii) there is no gap between their quality. Each question was annotated by five annotators, and the majority was regarded as the final label. We selected questions with label "the human answer is better", resulting in 2,239 questions in the development set and 6,851 in the test set. We further added pseudo data to the test set to avoid exhaustive tuning. Table 3 lists the statistics of the data.

Table 3. Statistics of the machine reading comprehension dataset.

	Train	Dev	Test	Pseudo	Released Test
Baidu Search	135,000	1,179	3,959	56,041	60,000
Baidu Zhidao	135,000	1,060	2,892	57,108	60,000
Total	270,000	2,239	6,851	113,149	120,000

Baseline System This task provided participants with an open source baseline system based on BiDAF [6], implemented in PaddlePaddle.[9] BiDAF is a MRC model that achieved promising results on a variety of benchmarks.

Evaluation Metrics ROUGE-L [3] and BLEU-4 [5] were adopted as evaluation metrics, with the former used for ranking participating systems. We further made minor adaptations to the original metrics [8]. For Entity questions, correct entities mentioned in answers would receive additional reward. For YesNo questions, participants were expected to further predict the opinion of corresponding answers, and would receive additional bonus for correct predictions.

3 Organization & Participation

LIC 2019 took place between February and May 2019. The detailed schedule is:

[9] `https://github.com/baidu/DuReader`

- **Feb 25:** Registration opened, (partial) training and dev sets available;
- **Mar 31:** Registration closed, whole training and dev sets available, partial test set available, online evaluation opened;
- **May 13:** Whole test set available;
- **May 20:** Deadline of result submission, offline evaluation opened;
- **May 25:** Deadline of code submission for top 10 systems (only required for the knowledge-driven dialogue task);
- **May 31:** Notification of final rankings.

For all the tasks, test sets were released in two parts. The first part was released right after the registration deadline, used for online evaluation and ranking. The second part was released a week before the submission deadline. Performance on whole test sets was evaluated offline and used for final ranking. For the dialogue task, the top 10 systems were further required to submit their code for manual evaluation, by which final rankings and winners of this task were determined.

Overall, there were 2,376 teams that took part in the challenge, among which 1,836 participated in the IE task, 1,536 the dialogue task, and 1,553 the MRC task. About 60% of the participants came from universities, including 93 Project 211 universities at home and 28 universities abroad. The other 40% came from over 300 enterprises, e.g., NetEase, Kingsoft, and Samsung, etc.

During the evaluation phase (online and offline) 635 teams made valid submissions, with a total of 6,212 system runs submitted. The IE task got 3,367 submissions from 324 teams, the dialogue task 1,688 submissions from 178 teams, and the MRC task 1,157 submissions from 133 teams. Compared against the official baselines, the best performing system obtained an improvement of 21.65%, 37.40%, and 34.09% respectively on each task.

4 System Performance

LIC 2019 awarded one first prize, two second prizes and two third prizes for each task. This section presents the performance of these winning systems. Results of all participating systems are available on the official website.[10]

Information Extraction The overall evaluation results of the top 5 winning systems (S1-S5) are shown in Table 4. It can be seen that overall performances have been greatly improved, and the development of information technology has been promoted. With the F1 value, the top 1 system performance has a 21.65% (from 73.41% to 89.3%) improvement compared to the official baseline.

We further analyzed the evaluation results in different types of text sources. Table 5 presents the performance of the top 5 average and the top 10 average systems on encyclopedia and feed news separately. It can be seen that the performance of almost all participating systems on the encyclopedic text is better than the news text and the average F1 value of top 10 extraction systems on the encyclopedia text is 11.9% higher than that of the feed news. This shows that it is more difficult to extract news texts involved with diverse linguistic pattern.

[10] `http://lic2019.ccf.org.cn/`

Table 4. Evaluation results of the top 5 systems on information extraction.

System No.	Precision	Recall	F1
S1	89.75%	88.86%	89.3%
S2	89.62%	88.86%	89.24%
S3	89.76%	88.52%	89.14%
S4	89.48%	88.58%	89.03%
S5	89.24%	88.2%	88.72%
Baseline	**77.52%**	**69.72%**	**73.41%**

Table 5. Evaluation results in different sources of text.

	Encyclopedia			Feed News		
	Precision	Recall	F1	Precision	Recall	F1
Avg-top5	92.6%	92.3%	92.4%	82.4%	80.1%	81.2%
Avg-top10	92.2%	91.5%	91.9%	81.5%	78.6%	80.0%

Knowledge-driven Dialogue Table 6 and Table 7 list the automatic evaluation and human evaluation results of the top 10 systems. From the results, we can see that the performance on proactive conversation has been effectively improved. The score increases by 36.99% from 0.919 to 1.259 for automatic evaluation metrics, and 37.40% from 1.287 to 1.768 for human evaluation metrics.

Table 7 shows that the conversation goals have been completed very well, with an average score of 1.85, close to the maximum score 2.0. However the conversation coherence is far from perfect, whose score is only 2.59 (the maximum score is 4). It indicates that the systems could complete the given goal in most case, but with some sacrifice of multi-turn coherence.

Machine Reading Comprehension Table 8 lists the performance of the top 5 winning systems (S1-S5), where Baidu Search and Baidu Zhidao indicate the results on questions whose documents were collected from the two channels, and Total the results on the whole test set. We can see that all the winning systems perform substantially better than the official baseline, with the ROUGE-L score pushed from 47.08% to 63.13%, i.e., a relative improvement of 34.09%. Despite this significant improvement, there is still a big gap between system and human answers. Building machines that can conduct in-depth language understanding and answer difficult questions is still a challenge.

Table 9 further presents the performance of the winning systems on different query types. The results show that YesNo questions are more difficult for machines, while Entity questions are more difficult for humans.

5 Conclusion

The CCF, CIPS, and Baidu Inc. jointly organized the 2019 Language and Intelligence Challenge (LIC 2019), which comprised three tasks: Information Extraction, Knowledge-driven Dialogue, and Machine Reading Comprehension. There

Table 6. Automatic evaluation results of top 10 systems on the dialogue task.

rank	team	score	F1(%)	BLEU1/2	DISTINCT1/2
1	DLUT&Dicalab	**1.259**	**49.22**	**0.449/0.318**	0.118/0.299
2	iDeepWise	1.204	47.76	0.430/0.296	0.110/0.275
3	CUP_NLP	1.175	46.40	0.422/0.289	0.118/0.303
4	bangda	1.169	47.03	0.417/0.281	0.113/0.290
5	DH-Pretender	1.159	46.01	0.420/0.279	0.118/0.307
6	fxnlp	1.149	46.03	0.417/0.271	**0.129/0.318**
7	wholly	1.148	45.74	0.420/0.271	0.096/0.248
8	travel	1.143	44.84	0.412/0.283	0.119/0.293
9	DG	1.134	45.48	0.414/0.266	0.095/0.229
10	AI 小奶娃	1.132	45.25	0.411/0.268	0.122/0.308
baseline		0.919	37.69	0.347/0.198	0.057/0.155

Table 7. Human evaluation results of top 10 systems on the dialogue task.

rank	team	score	goal completion	coherence
1	bangda	**1.768**	1.81	**2.59**
2	DLUT&Dicalab	1.732	**1.85**	2.42
3	fxnlp	1.720	1.80	2.46
4	iDeepWise	1.715	1.73	2.55
5	CUP_NLP	1.662	1.79	2.30
6	travel	1.602	1.73	2.21
7	wholly	1.587	1.76	2.12
8	DG	1.515	1.53	2.25
9	DH-Pretender	1.513	1.66	2.05
10	AI 小奶娃	1.503	1.72	1.93
baseline		1.287	1.22	2.03

were 2,376 teams that participated in the challenge, with a total of 6,212 system runs submitted. The winning systems offered a 21.65% increase over the official baseline in information extraction, a 37.40% increase in knowledge-driven dialogue, and a 34.09% increase in machine reading comprehension. Although LIC 2019 has greatly advanced the technical development of natural language understanding and interaction, there are still many unsolved challenges, e.g., how to extract structured knowledge from news texts with diverse linguistic patterns, how to effectively evaluate the performance of a dialogue system, and how to conduct in-depth language understanding so as to answer difficult questions.

References

1. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A knowledge-grounded neural conversation model. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
2. He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H.: DuReader: A Chinese machine reading com-

Table 8. Results of top 5 winning systems on machine reading comprehension (%).

	Baidu Search		Baidu Zhidao		Total	
	ROUGE-L	BLEU-4	ROUGE-L	BLEU-4	ROUGE-L	BLEU-4
Human	**89.95**	**88.50**	**88.33**	**85.72**	**89.26**	**87.27**
S1	60.34	56.99	66.96	63.10	63.13	59.34
S2	56.54	58.01	64.31	61.15	59.82	59.34
S3	54.91	53.33	61.75	58.82	57.80	55.55
S4	53.84	54.89	61.32	61.08	57.00	57.30
S5	52.45	53.20	59.69	59.44	55.51	55.71
Baseline	**42.46**	**42.15**	**53.42**	**49.52**	**47.08**	**46.01**

Table 9. Results of top 5 winning systems on different query types (%).

	Description		Entity		YesNo	
	ROUGE-L	BLEU-4	ROUGE-L	BLEU-4	ROUGE-L	BLEU-4
Human	**91.50**	**89.65**	**84.75**	**76.83**	**87.26**	**87.89**
S1	64.08	61.34	62.97	51.68	56.87	56.59
S2	61.14	61.97	59.31	50.52	52.01	49.50
S3	58.78	58.47	58.25	46.31	49.46	47.42
S4	58.46	59.83	56.71	48.57	47.52	47.23
S5	56.37	58.42	56.25	47.74	47.11	45.78
Baseline	**47.85**	**47.87**	**47.18**	**37.40**	**41.30**	**30.80**

prehension dataset from real-world applications. In: Proceedings of the Workshop on Machine Reading for Question Answering. pp. 37–46 (2018)
3. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
4. Liu, K., Liu, L., Liu, J., Lyu, Y., She, Q., Zhang, Q., Shi, Y.: Overview of 2018 NLP challenge on machine reading comprehension (in Chinese). Journal of Chinese Information Processing **32**(10), 118–129
5. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318 (2002)
6. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In: International Conference on Learning Representations (2017)
7. Wu, W., Guo, Z., Zhou, X., Wu, H., Zhang, X., Lian, R., Wang, H.: Proactive human-machine conversation with explicit conversation goals. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2019)
8. Yang, A., Liu, K., Liu, J., Yajuan, L., Li, S.: Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In: Proceedings of the Workshop on Machine Reading for Question Answering. pp. 98–104 (2018)