Benchmarks: An Industry Perspective

Hua Wu and Jing Liu

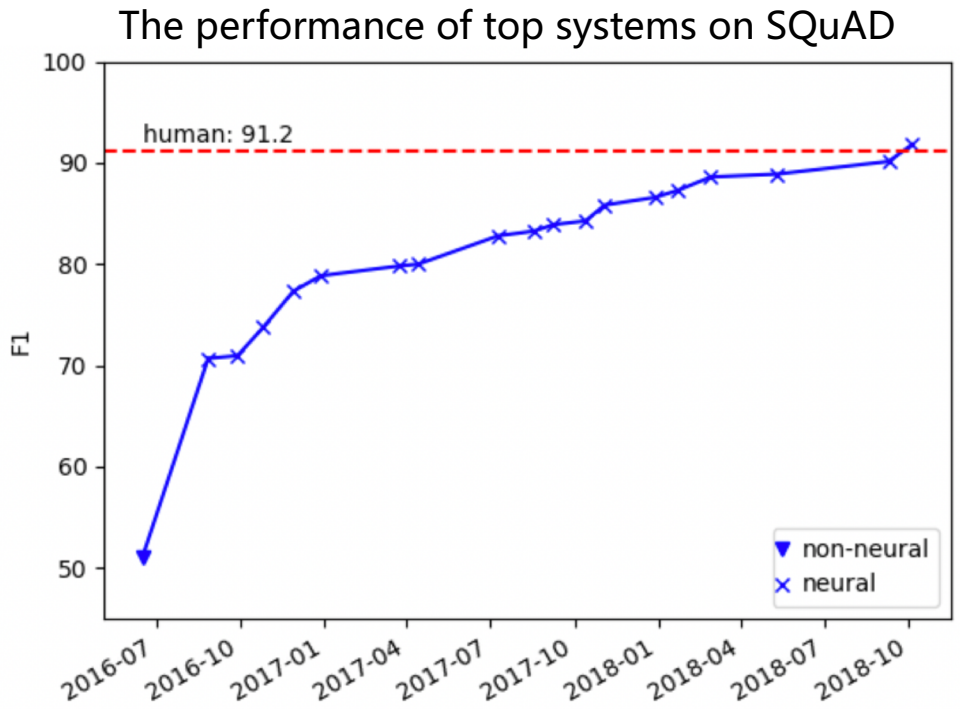
Baidu

Outline

- **Introduction**
- Our Open Datasets for Challenges in Real-world Setting
- LUGE: an Open-Source Project of Chinese NLP Benchmarks
- Conclusions

Introduction

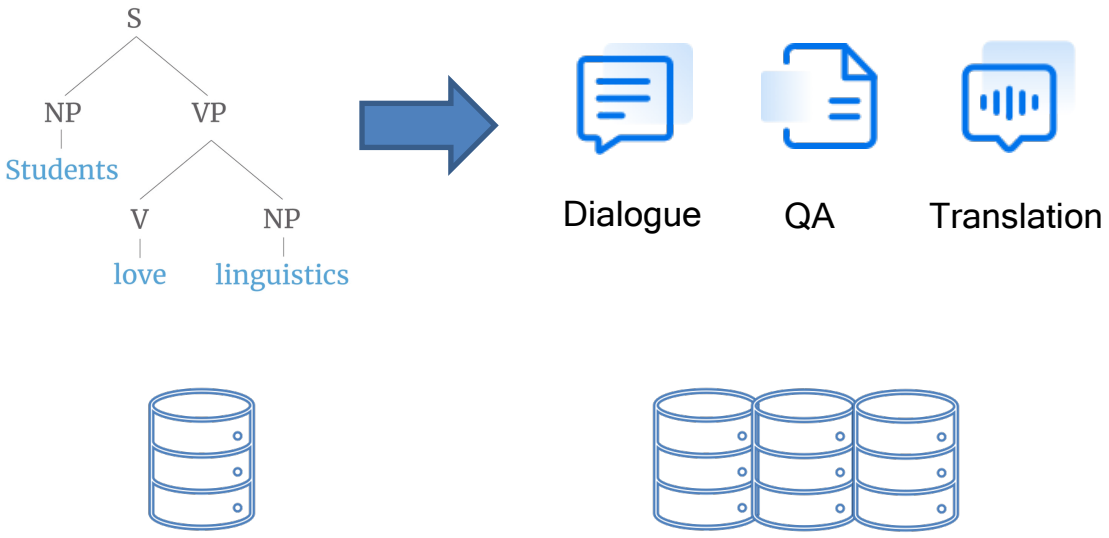
- In recent years, the researchers from academia created large-scale datasets mainly in a crowdsourcing way, e.g. SQuAD, that accelerate the development of NLP technology.



* Picture credit: Danqi's thesis

The Trending of Benchmarks

From linguistics to applications



From single-task to multi-task
From single-domain to multi-domain

GLUE

MRQA

The Challenges in Real-world Settings

- However, the datasets created in a crowd-sourcing way might present different distributions and different challenges from those in real-world applications.

Naturality **Robustness**

Knowledge **Faithfulness**

Multi-modal **Multi-task and Multi-domain**

…

* Picture credit: www.pararecruit.com

Outline

- Introduction
- **Our Open Datasets for Challenges in Real-world Setting**
  - **Naturality in Question Answering**
  - **Knowledge in Dialogue**
  - **Multi-Modal in Simultaneous Translation**
- LUGE: an Open-Source Project of Chinese NLP Benchmarks
- Conclusions

Our Open Datasets

Question Answering

DuReader Machine Reading Comprehension Dataset 300K questions, 1.5M documents	**DuSQL** Text-to-SQL Dataset 23K pairs of questions and SQL queries, 200 databases	**DuQM** Robust Question Matching Dataset 11K question pairs

Dialogue

DuConv Knowledge Driven Dialogue Dataset 120K turns of conversion	**DuRecDail** Conversational Recommendation Dataset 14K turns of conversation	**DuPersona** Persona Based Dialog Dataset 23k turns of conversations

Sentiment Analysis

DuTrustSenti Trustworthy Sentiment Analysis Dataset 3k sentences with labeled explanations	**DuMMSenti** Multimodal Sentiment Classification Dataset 8.6k videos and 11 sentiment classes

Machine Translation

BSTC Baidu Speech Translation Corpus 68 hours of audio, transcriptions and translations	**BLTC** Baidu Low-resource Translation Corpus 50K bilingual sentences and 200K monolingual sentences in low-resource languages

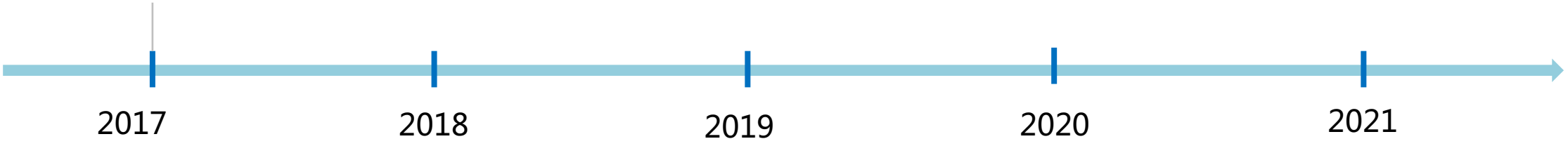
Information Extraction

DuIE Relation Extraction Dataset 210K sentences and 48 schemas	**DuEE** Event Extraction Dataset 17K sentences and 65 event schemas	**DuEE-fin** Doc-level Event Extraction Dataset in Finance domain 11.5k documents and 13 event schemas	**DuEL** Entity Linking Dataset 100K annotated short text

Our efforts on building question answering benchmarks in real-world settings

Natural questions and docs from Baidu search

DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications



(Extractive) Machine Reading Comprehension

- Given a passage and a question, the machine should extract the answer to the question that is a segment of text from the corresponding reading passage

Q How long is Qin Dynasty?

A 14 years

P Qin Dynasty was the first unified, multi-national and power-centralized state in the Chinese history. It lasted from 221 BC to 207 BC. Although surviving only 14 years, the dynasty held an important role in Chinese history.

(Extractive) Machine Reading Comprehension

- Questions: created in a crowdsourcing way
- Document: single passage
- Type: focus on factoid question

DuReader: a Chinese MRC Dataset from Real-world Applications

Real Data Sources

	Source
Questions	Baidu Search logs
Documents	Baidu Search (Web Doc.) Baidu Zhidao (UGC)

Rich Question Types

	Fact	Opinion
Entity	iphone x哪天发布 When will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are fire engines red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Can learning go develop intelligence

Multi-documents and Multi-answers

- 5 candidate documents per question
- 66% questions have multiple answers

Large Data Scale

The largest Chinese MRC dataset so far

#Que	#Doc	#Ans
300K	1.5M	660K

DuReader: a Chinese MRC Dataset from Real-world Applications

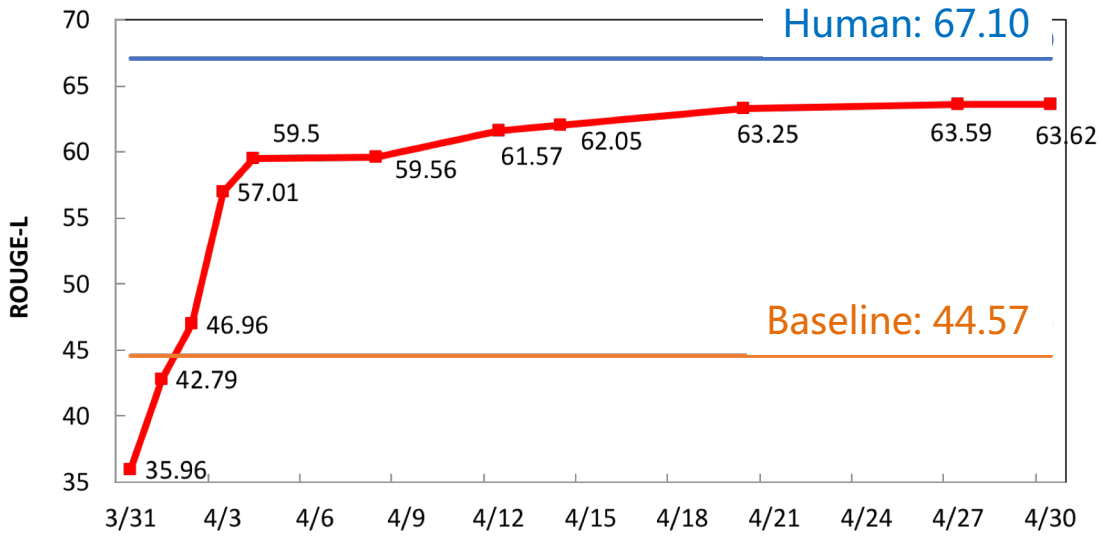
github.com/baidu/dureader

35,000+ downloads

2018 - 2021 Language and Intelligence Challenge

5,200+ registrations
12,000+ submissions

DUREADER 2.0	Multi-doc
$\text{DUREADER}_{\text{yesno}}$	Yes/no questions
$\text{DUREADER}_{\text{robust}}$	Robustness
$\text{DUREADER}_{\text{checklist}}$	Fine-grained robustness

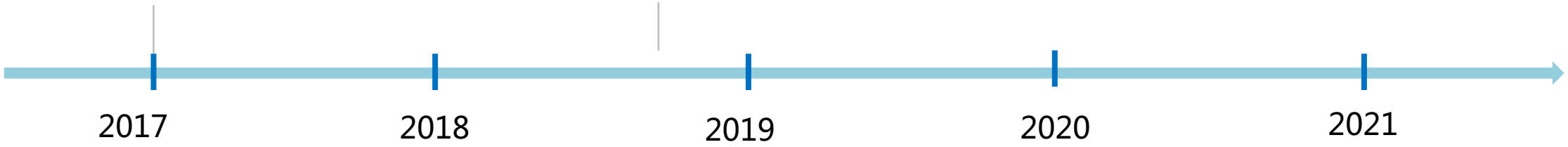


Our efforts on building dialogue benchmarks in real-world settings

Incorporating knowledge

DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications

DuConv: Proactive human-machine conversation with explicit conversation goal

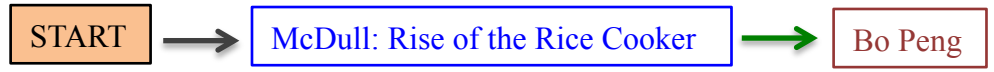


Proactive human-machine conversation with explicit conversation goal. 2019. In ACL.

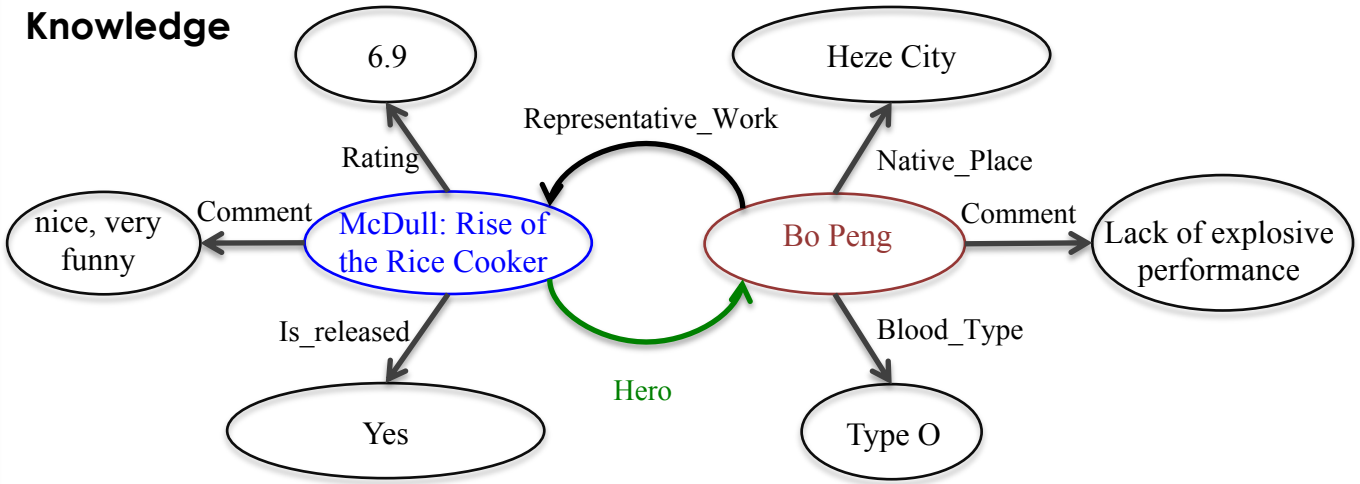
DuConv: Proactive Conversation with Explicit Conversation Goal

- **Goal:** enable chatbots to conduct knowledge-grounded conversation.
- **Task:** The chatbot proactively leads the conversation to complete the **goal** with related **knowledge**.

Goal



Knowledge



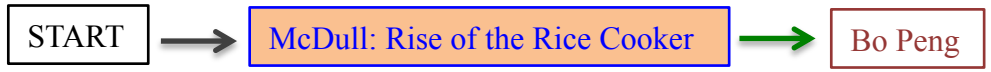
Conversation

Bot: 平时周末去看电影吗？
(Do you usually spend your weekend watching movies?)
User: 看呀，最近有好看的可以推荐一下，周末去看。
(Of course, could you recommend good films for me? I will watch it on this weekend.)
Bot: 麦兜·饭宝奇兵这部电影，大众评分在6.9呢，值得一看。
(You can watch Rise of the Rice Cooker, which is rated 6.9.)
User: 好的，一定去看看。(Sounds great, thanks.)
Bot: 这部电影还不错，挺搞笑的。你知道饭麦兜·饭宝奇兵的主演是谁吗?
(It is very nice and funny. Do you know who is the hero of this movie?)
User: 不知道，谁啊？(I don't know. Who is that?)
Bot: 彭博，网友们评论说他缺少爆发力太温柔。(It's Bo Peng. Do you know him? People said that he was a kind of lacking explosive Performance.)
User: 那我有时间去了解一下。（I will know more about him later.）

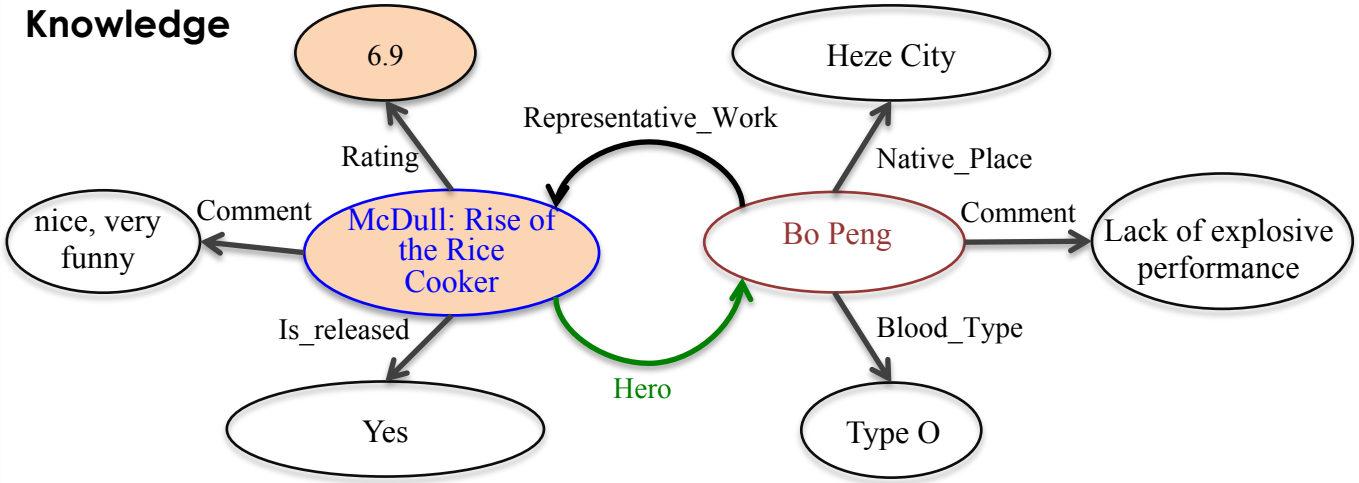
DuConv: Proactive Conversation with Explicit Conversation Goal

- **Goal:** enable chatbots to conduct knowledge-grounded conversation.
- **Task:** The chatbot proactively leads the conversation to complete the **goal** with related **knowledge**.

Goal



Knowledge

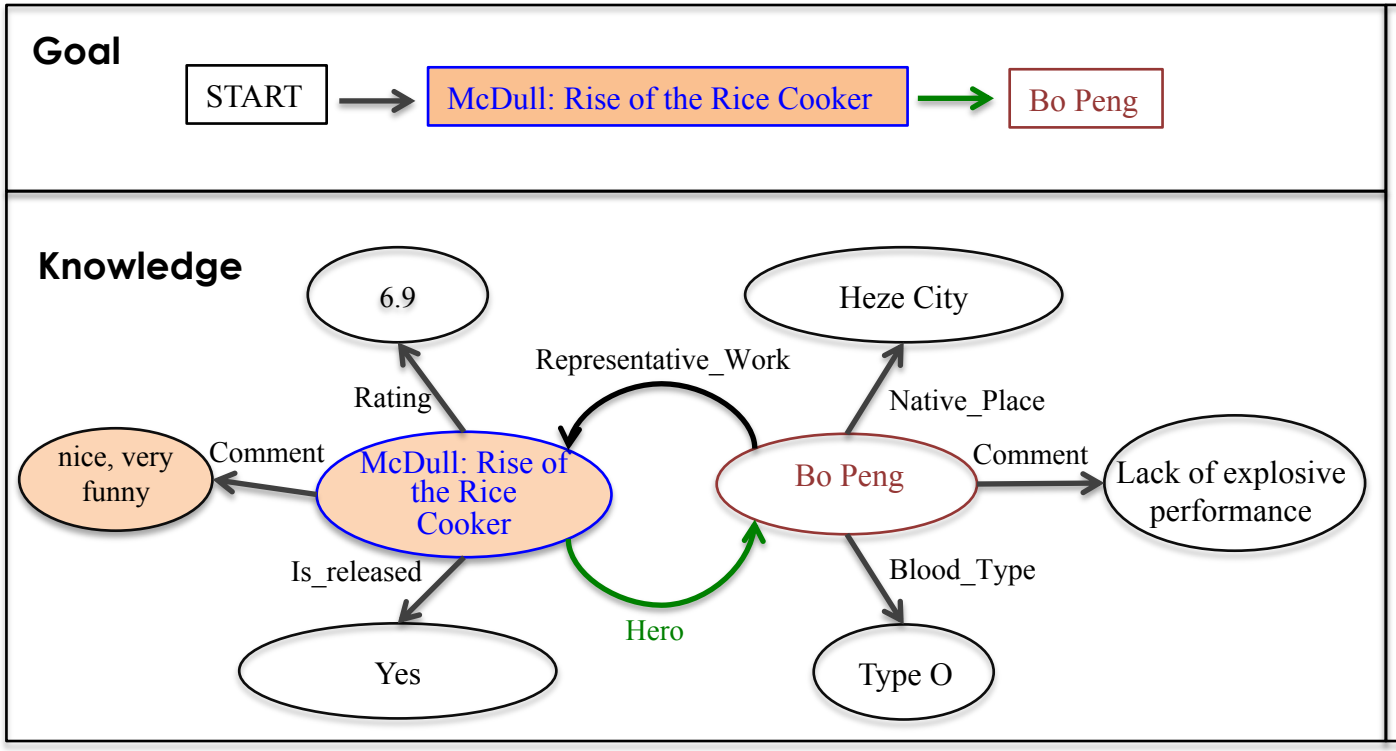


Conversation

Bot: 平时周末去看电影吗？
(Do you usually spend your weekend watching movies?)
User: 看呀，最近有好看的可以推荐一下，周末去看。
(Of course, could you recommend good films for me? I will watch it on this weekend.)
Bot: 麦兜·饭宝奇兵这部电影，大众评分在6.9呢，值得一看。
(You can watch Rise of the Rice Cooker, which is rated 6.9.)
User: 好的，一定去看看。(Sounds great, thanks.)
Bot: 这部电影还不错，挺搞笑的。你知道饭麦兜·饭宝奇兵的主演是谁吗?
(It is very nice and funny. Do you know who is the hero of this movie?)
User: 不知道，谁啊？(I don't know. Who is that?)
Bot: 彭博，网友们评论说他缺少爆发力太温柔。(It's Bo Peng. Do you know him? People said that he was a kind of lacking explosive Performance.)
User: 那我有时间去了解一下。（I will know more about him later.）

DuConv: Proactive Conversation with Explicit Conversation Goal

- **Goal:** enable chatbots to conduct knowledge-grounded conversation.
- **Task:** The chatbot proactively leads the conversation to complete the **goal** with related **knowledge**.



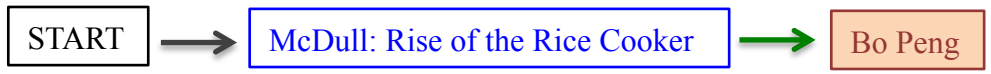
Conversation

Bot: 平时周末去看电影吗？
(Do you usually spend your weekend watching movies?)
User: 看呀，最近有好看的可以推荐一下，周末去看。
(Of course, could you recommend good films for me? I will watch it on this weekend.)
Bot: 麦兜·饭宝奇兵这部电影，大众评分在6.9呢，值得一看。
(You can watch Rise of the Rice Cooker, which is rated 6.9.)
User: 好的，一定去看看。(Sounds great, thanks.)
Bot: 这部电影还不错，挺搞笑的。你知道饭麦兜·饭宝奇兵的主演是谁吗?
(It is very nice and funny. Do you know who is the hero of this movie?)
User: 不知道，谁啊？(I don't know. Who is that?)
Bot: 彭博，网友们评论说他缺少爆发力太温柔。(It's Bo Peng. Do you know him? People said that he was a kind of lacking explosive Performance.)
User: 那我有时间去了解一下。（I will know more about him later.）

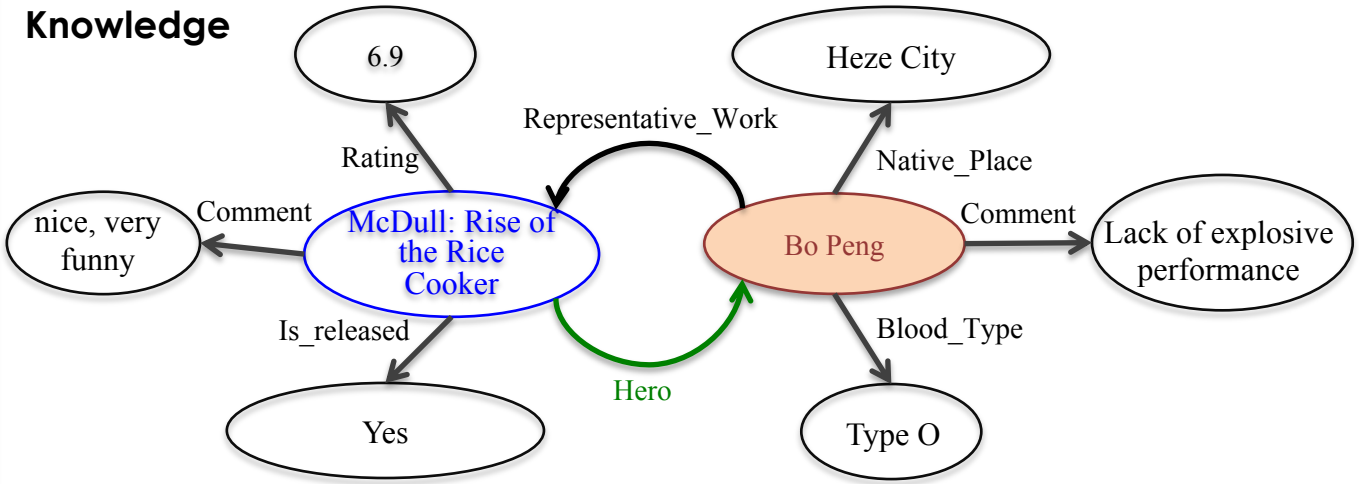
DuConv: Proactive Conversation with Explicit Conversation Goal

- **Goal:** enable chatbots to conduct knowledge-grounded conversation.
- **Task:** The chatbot proactively leads the conversation to complete the **goal** with related **knowledge**.

Goal



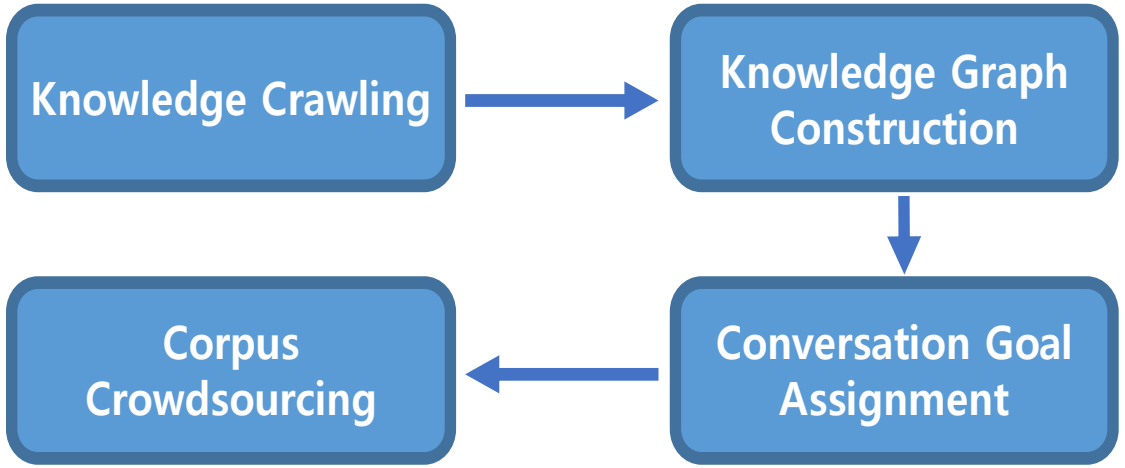
Knowledge



Conversation

Bot: 平时周末去看电影吗？
(Do you usually spend your weekend watching movies?)
User: 看呀，最近有好看的可以推荐一下，周末去看。
(Of course, could you recommend good films for me? I will watch it on this weekend.)
Bot: 麦兜·饭宝奇兵这部电影，大众评分在6.9呢，值得一看。
(You can watch Rise of the Rice Cooker, which is rated 6.9.)
User: 好的，一定去看看。(Sounds great, thanks.)
Bot: 这部电影还不错，挺搞笑的。你知道饭麦兜·饭宝奇兵的主演是谁吗?
(It is very nice and funny. Do you know who is the hero of this movie?)
User: 不知道，谁啊？(I don't know. Who is that?)
Bot: 彭博，网友们评论说他缺少爆发力太温柔。(It's Bo Peng. Do you know him? People said that he was a kind of lacking explosive Performance.)
User: 那我有时间去了解一下。（I will know more about him later.）

DuConv: Proactive Conversation with Explicit Conversation Goal



- Dialog proactivity:
  - Goal driven proactive dialogs
- Knowledge grounded dialogs
  - Multiple topics in a single session
  - Evaluation of information faithfulness in responses

DuConv: Proactive Conversation with Explicit Conversation Goal

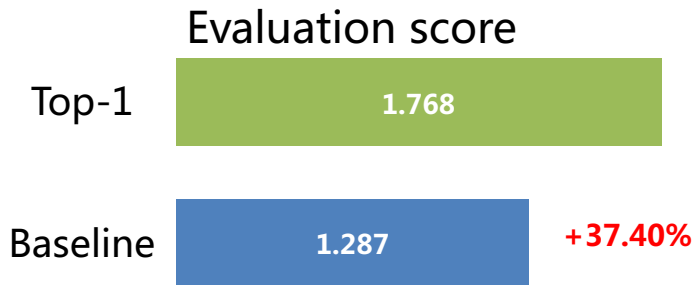
https://github.com/baidu/knowledge-driven-dialogue

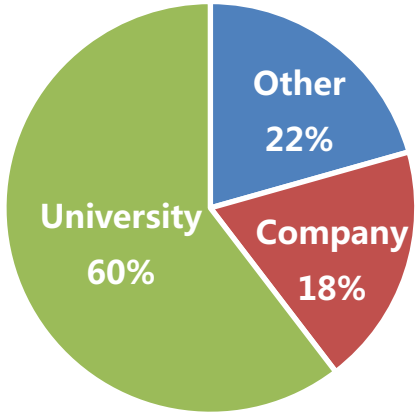
1970 downloads, 261 star, 85 fork

Statistics of Knowledge	
entities	143,627
knowledge triplets	3,598,246

Statistics of Dialogues	
dialogs	29,858
utterances	270,399

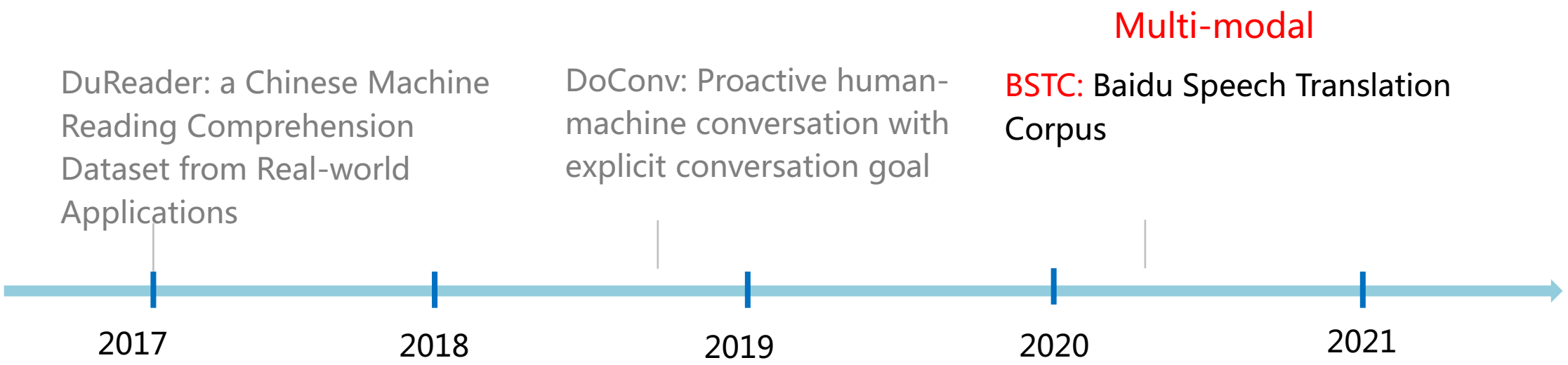
1,536 registrations
178 submissions





Distribution of registration

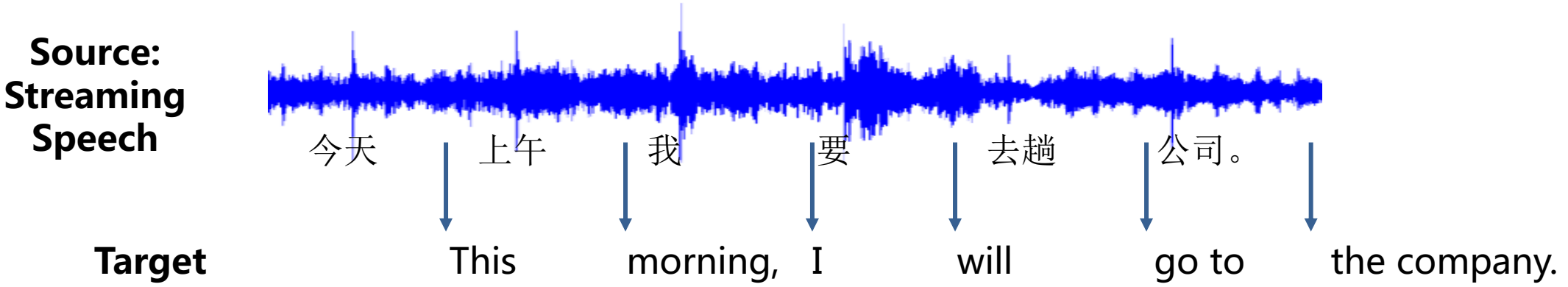
Our efforts on building simultaneous translation benchmarks in real-world settings



BSTC: A large-scale Chinese-English speech translation dataset. 2021. In NAACL-autosimtrans.

Simultaneous Translation

- **Main Challenges**
  - Lack of training data; balance of translation quality and latency; evaluation metrics
- **Motivation**
  - Speech-to-text simultaneous translation
  - Chinese-English simultaneous translation in real-world settings
  - New evaluation metrics for both translation quality and latency



BSTC: A Large-Scale Chinese-English Speech Translation Dataset

- Collected from Chinese talks similar to TED
- Domains: technology, economy, culture, art, etc.
- Almost 68 hours

	talks	sentences	Characters / words		Hours
			Chinese	English	
Training set	215	37901	1,028,538	524,395	64.71
Dev set	16	956	26,059	13,277	1.58
Test set	6	975	25,832	12,724	1.46

https://ai.baidu.com/broad/introduction?dataset=bstc

Comparison of BSTC and Others

	Speech	Translation	Data Size
TED-LIUM	English	Multiple	118 hours
Fisher-CALLHOME	Spanish	English	38 hours
Augmented LibriSpeech	English	French	100 hours
MSLT (Microsoft)	Multiple	Multiple	19 hours
BSTC	Chinese	English	68 hours

The Largest Chinese to English Speech to Text Corpus

A Sample of BSTC

Field	Content
Audio	
ASR	那么我们今天呢，就希望从一个20年的AI工作者来说，如何从专业的角度去解读一下，我们现在究竟发生了什么事情？他的权势金生。
Transcript	那么我们今天呢就希望，从一个二十年的AI工作者来说，如何从专业的角度去解读一下，我们现在究竟发生了什么事情，它的前世今生。
Translation	So today, as one who has been working on AI for twenty years, I wish I could give you a professional interpretation of what exactly is going on, its origin, history, characteristic, and where it is going.

Benchmark for Simultaneous Translation

2020-2021 Automatic Simultaneous Translation Shared Task

https://aistudio.baidu.com/aistudio/competition/detail/44

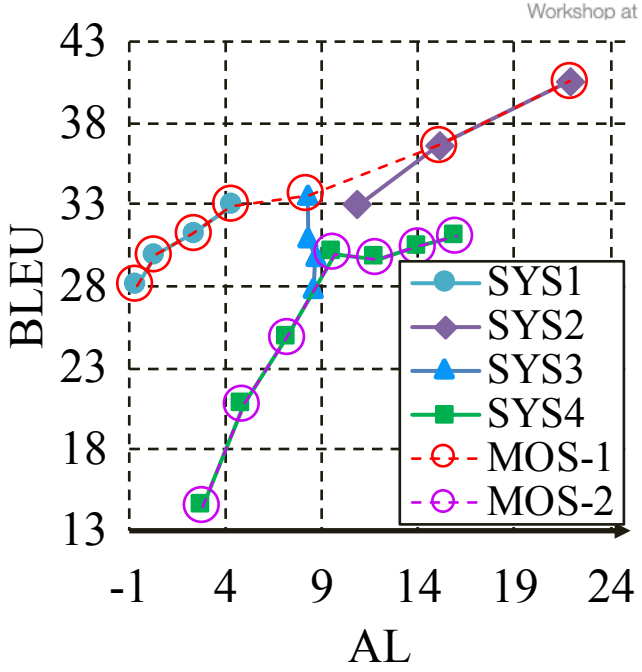
843 registrations
10 submissions

- Task
  - Chinese to English : speech to text, or text to text
- Evaluation metrics:
  - BLEU & Average Lagging
- New methods to rank AST systems
  - Iterative Monotonic Optimal Sequence

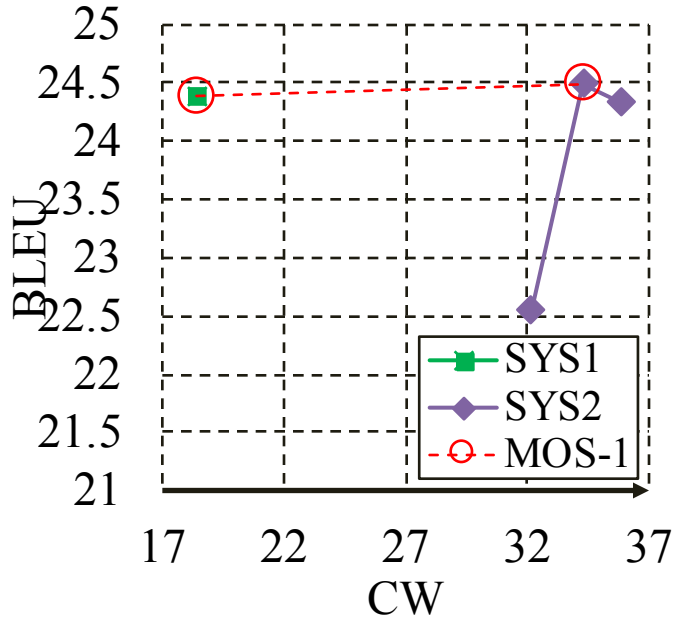
The 2nd Workshop on Automatic Simultaneous Translation

Challenges, Recent Advances, and Future Directions

Workshop at NAACL 2021, Mexico City, June 10, 2021

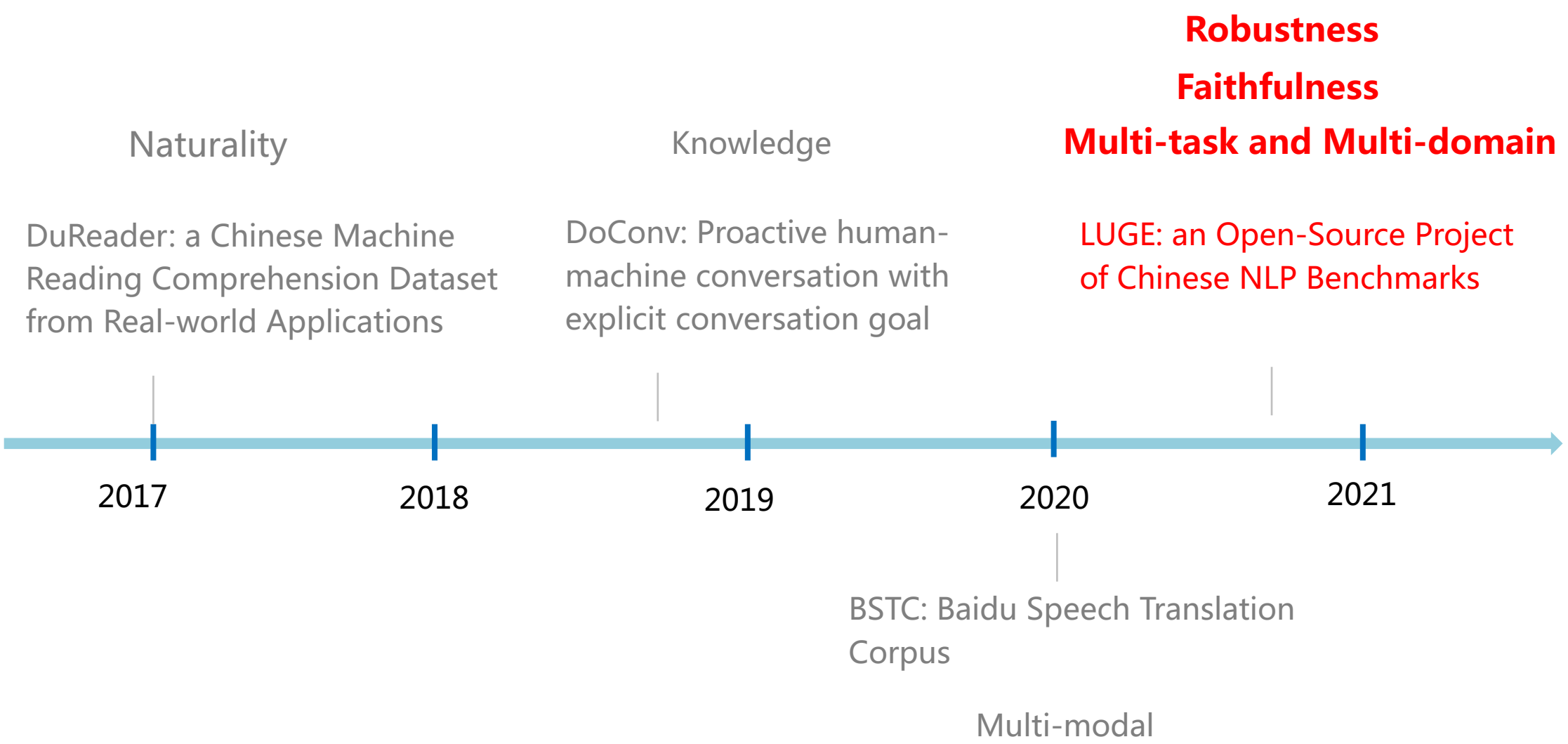


Track1. Text-to-Text AST



Track2. Speech-to-Text AST

Can we work together to make more progress?



Outline

- Introduction
- Our Open Datasets for Challenges in Real-world Setting
- **LUGE: an Open-Source Project of Chinese NLP Benchmarks**
  - **Robustness in Question Matching**
  - **Faithfulness in Text Generation**
  - **Multi-task and Multi-domain in Sentiment Analysis**
- Conclusions

LUGE: an Open-Source Project of Chinese NLP Benchmarks

- It was initiated by Baidu, CCF and CIPSC in Aug 2020.
- Comprehensive evaluations: robustness, faithfulness, generalization etc.
- Rich types of tasks: natural language understanding, natural language generation and multi-modal tasks, etc.

The Contributors of LUGE from 12 Organizations

Prof. Qingcai Chen et al.
Harbin Institute of Technology (Shenzhen)

Dr. Shuming Shi et al.
Tencent

Dr. Ming Zhou et al.
Sinovation Ventures*

Prof. Minlie Huang et al.
Tsinghua University

Dr Lifeng Shang et al.
Huawei

Prof. Yanyan Zhao et al.
Harbin Institute of Technology

Dr. Yunfeng Liu et al.
Zhuiyi Technology

Dr. Songbo Tan et al.
Lenovo*

Prof. Yue Zhang et al.
Westlake University

Prof. Tingwen Liu et al.
IIE Chinese Academy of Sciences

Dr. Hua Wu et al.
Baidu Inc.

Dr. Jinggang Wang et al.
Meituan Inc.

Prof. Baotian Hu et al.
Harbin Institute of Technology (Shenzhen)

* The data set was built by the author in Microsoft and ICTCAS

The Current Progress of LUGE

10 tasks

36 datasets

unified data formats
and evaluation scripts

baselines

The 10 Tasks in LUGE

- Dialogue
- Sentiment Analysis
- Simultaneous Translation
- Low Resource Machine Translation
- Machine Reading Comprehension
- Semantic Parser
- Question Matching
- Entity Linking
- Information Extraction
- Text Generation

Can we work together to make more progress?

Question Matching

Robustness

Text Generation

Faithfulness

Sentiment Analysis

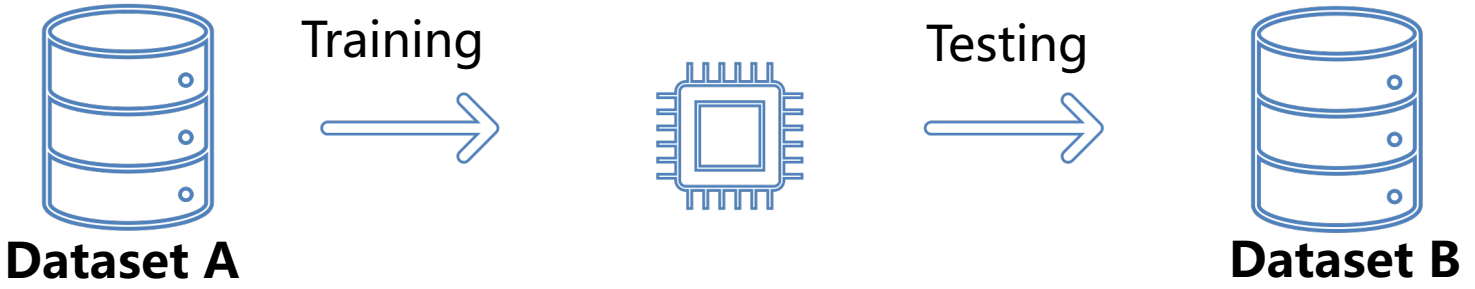
Multi-task、Multi-domain、Multi-modal

The Robustness of Question Matching

- **Question Matching**: identify question pairs that have the same intent

Question 1	*Question 2*	*Same Intent ?*

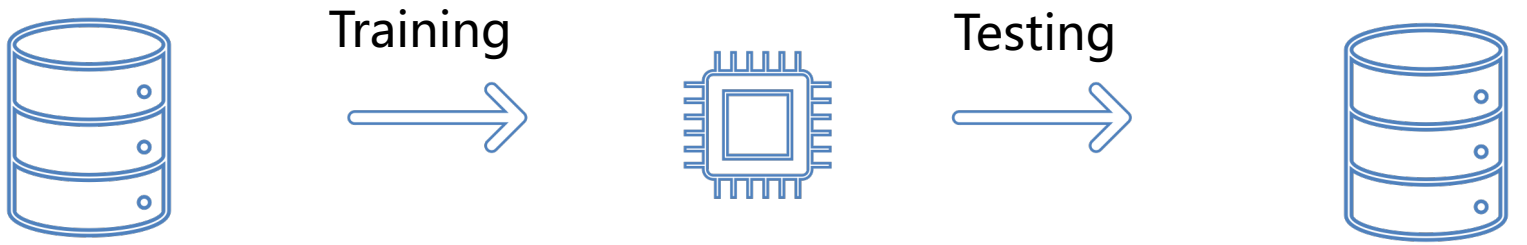
- The neural models learn **shortcuts** from training set, and suffer from the **robustness** issues in real-world settings



Question Pair	Translation	Label	Model
婴儿吃什么蔬菜好 婴儿吃什么 绿色 蔬菜好	*What vegetables are good for baby* *What green vegetables are good for baby*	N	Y

The Datasets for Robustness Evaluation of Question Matching

Dataset	Source	Domain	Training	Testing
LCQMC	Baidu Zhidao	General Domain	Y	
BQ Corpus	Bank	Bank Domain	Y	
DuQM	Baidu Search	General Domain		Y



LCQMC
BQ Corpus

DuQM

DuQM: Linguistically Perturbed Natural Questions for Robustness Evaluation of Question Matching

- Fine-grained types with linguistic perturbation
- Natural questions from Baidu search

Cross-grained Category	Fine-grained Category	Example	Label
Lexical Semantics	Word & Phrase	下蹲膝盖疼 / 下跪膝盖疼 knee pain when squatting / knee pain when kneeling	N
	Named Entity	iphone 6多少钱 / iphone6x多少钱 how much is iphone 6 / how much is iphone6x	N
	Synonym	猕猴桃的功效是什么 / 奇异果的功效是什么 what are the health benefits of Chinese gooseberry / what are the health benefits of Kiwi	Y
	Antonym	只吃蔬菜会让皮肤好吗 / 只吃蔬菜会让皮肤差吗 does only eating vegetables make my skin better / does only eating vegetables make my skin worse	N
	Negation	为什么宝宝哭 / 为什么宝宝不哭 why baby cry / why baby doesn’ t cry	N
	Temporal	我一边吃饭一边看电视 / 我吃完饭要看电视 I eat while watching TV / I will watch TV after eat	N
Syntactic Structure	Symmetry	鱼和鸡蛋能一起吃吗 / 鸡蛋和鱼能一起吃吗 can I eat fish with egg / can I eat egg with fish	Y
	Asymmetry	北京飞上海航班 / 上海飞北京航班 Beijing to Shanghai flights / Shanghai to Beijing flights	N
	Negative Asymmetry	男人比女人更高吗 / 女人比男人更矮吗 are men taller than women/ are women shorter than men	Y
	Active/Passive	我撞了别人怎么办 / 我被别人撞了怎么办 what should I do if I hit someone / what should I do if I get hit by someone	N

Can we work together to make more progress?

Question Matching

Robustness

Text Generation

Faithfulness

Sentiment Analysis

Multi-task、Multi-domain、Multi-modal

Text Generation

Task: Generate natural language description based on input data, including various tasks, such as data-to-text, summarization generation and question generation.

Data-to-Text

City	Beijing	
Today's weather	min	-6
	max	5
Tomorrow's weather	min	-8
	max	3
Alerts	Strong wind	

The temperature in Beijing tomorrow is - 8 ℃ ~ 3 ℃. It ‘s a little colder than yesterday. It’ s windy. Remember to keep warm.

Summarization

Research institutions have previously issued a report saying that the long-term gold bull market in 2013 may end, and lowered the three-month, six-month and 12-month gold price estimates to 1,825 US dollars, 1,805 US dollars, and 1,800 US dollars, respectively. Recently, the research report of the commodity analyst Damien Courvalin further predicts that the international gold price may fall to US$1,200 per ounce by 2018...

The price of gold may fall to US$1,200 per ounce in 2018

Question Generation

Context: This mechanism is still the leading theory today; however, a second theory suggests that most cpdna is actually linear and replicates through homologous recombination.

Answer: homologous recombination

Question: How does the second theory say most cpdna replicates ?

Text Generation：From Fluency to Faithfulness

Input：Research institutions have previously issued a report saying that the long-term gold bull market in 2013 may end, and lowered the three-month, six-month and 12-month gold price estimates to 1,825 US dollars, 1,805 US dollars, and 1,800 US dollars, respectively. Recently, the research report of the commodity analyst Damien Courvalin further predicts that the international gold price may fall to US$1,200 per ounce by 2018...

Reference：The price of gold may fall to US$1,200 per ounce in 2018

Prediction_1(wrong): Analyst: The price of gold may fall to $1,800 in 2018 **Wrong fact**

Prediction_2(correct): Research Institutes Expect International Gold Prices to Fall to $1,200 **Correct fact**

Text Generation in LUGE

Dataset	Task	Domain
Advertising Text Generation	Data to Text	E-commerce
LCSTS	Summarization Generation	News
DuReader	Question Generation	General Domain

- **Factual faithfulness** is very important for the usability of text generation
  - Traditional text generation effect measurement methods (BLEU, ROUGE, etc.) cannot fully reflect availability
  - Incorporating both natural language inference and manual evaluation for faithfulness evaluation

Can we work together to make more progress?

Question Matching

Robustness

Text Generation

Faithfulness

Sentiment Analysis

Multi-task、Multi-domain、Multi-modal

Sentiment Analysis

Task: identify opinion, sentiment, affect/emotion and mood expressed in text, video, and audio.

Definition: Given an opinion doc, mine all quintuples.

(entity, aspect, sentiment, holder, time)

Example:

Id: John on 5-1-2008 – "I bought an iPhone yesterday. The touch screen is really cool. The voice quality is great"

(iPhone, touch_screen, +, John, 6-2-2008)

From Bing Liu's talk CCAC 2021

Multi-task, Multi-domain and Multi-modal in Sentiment Analysis

Multi-task

Sent-level Sentiment Classification

Input: This is amazing technology that is far more advanced than those of the real world
Output: 🙂

Aspect-level Sentiment Classification

Input: The taste is very good, but the service attitude is poor
Output: (taste , 🙂)

Opinion Extraction

Input: The service is good and the products are comprehensive
Output: (service , good) 、 (product , comprehensive)

Multi-domain

- **Book review** :

After reading “Young China”, I love my motherland even more -> positive

- **Phone review** :

The weight of the real machine is much heavier than that of the previous generation, and the operation is not smooth-> negative

Multi-modal

Title：Moved [Cry] [Cry]

Text：Neutral

Video：Sad

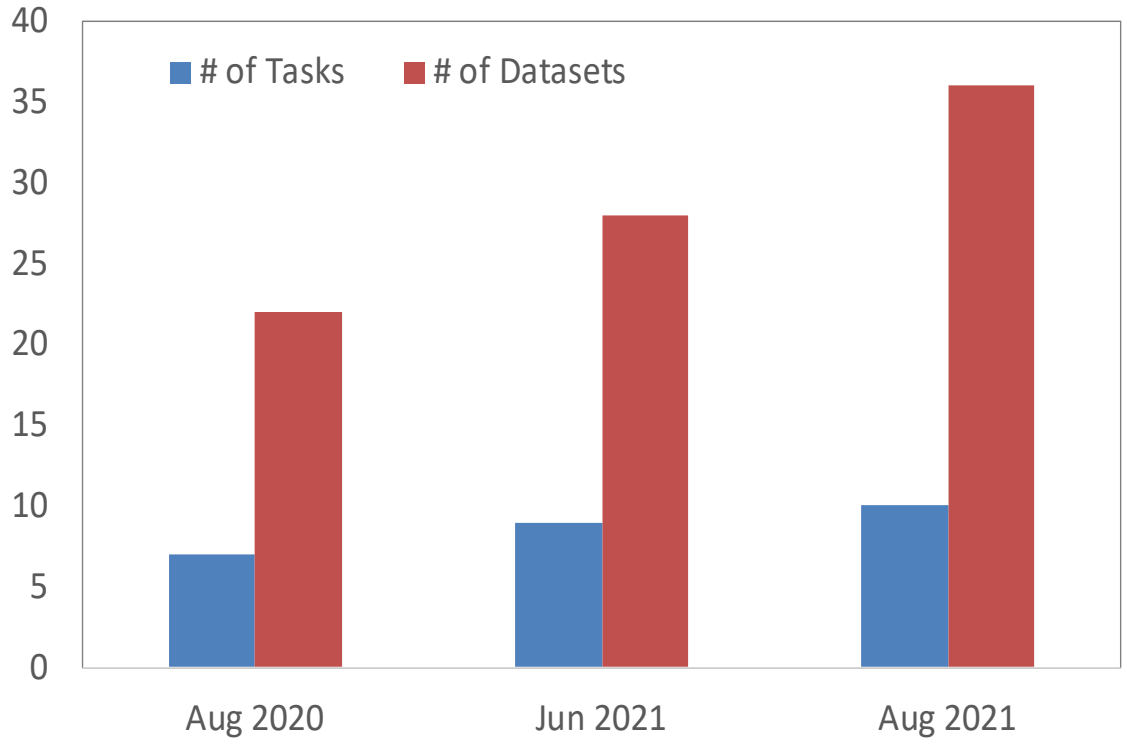
Emotion：Sad

Sentiment Analysis in LUGE

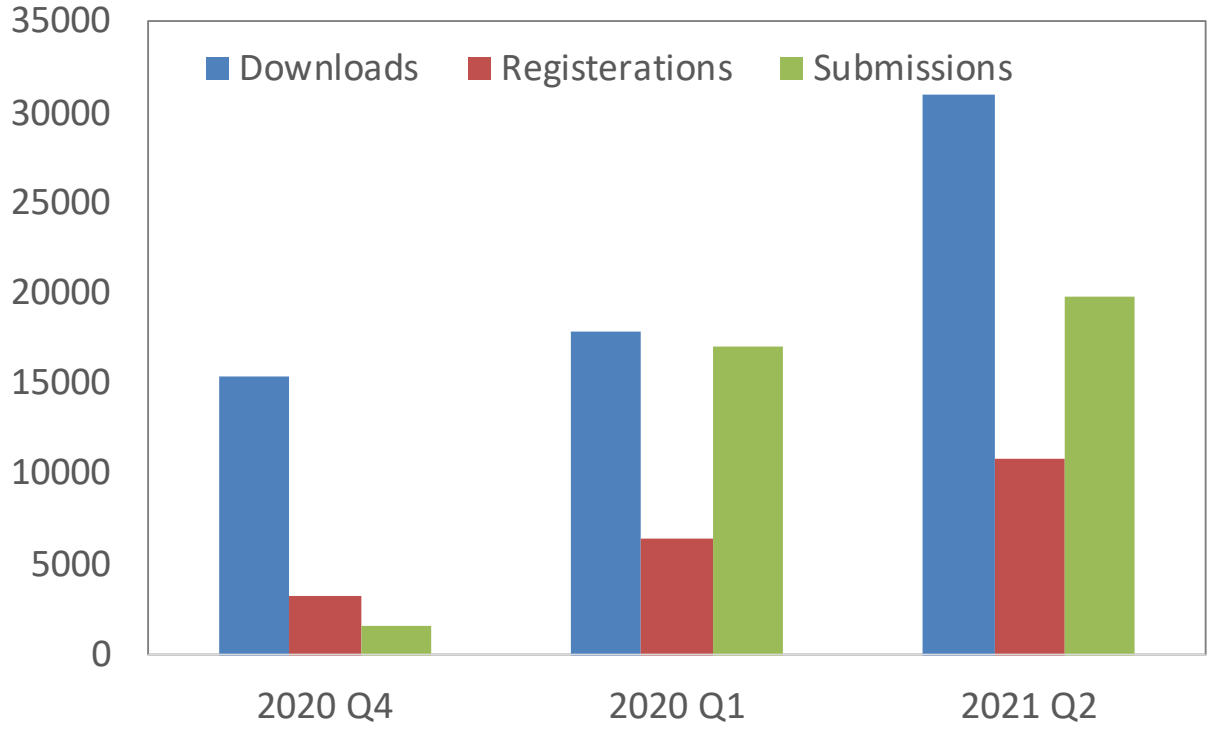
- 8 datasets, 4 tasks, 10 domains and 2 modals
- It aims to promote the unified modeling of multi-task, multi-domain and multi-modal sentiment analysis

Dataset	Task	Domain	Modal
ChnSentiCorp	Sent-level Sentiment Classification	Hotels, Laptops, Books	Text
NLPCC14-SC	Sent-level Sentiment Classification	Books, Electrical Appliances	Text
SE-ABSA16_PHNS	Aspect-level Sentiment Classification	Phone	Text
SE-ABSA16_CAME	Aspect-level Sentiment Classification	Digital Camera	Text
COTE-BD	Opinion Extraction	Travel, Restaurant, Movies, Books	Text
COTE-MFW	Opinion Extraction	Travel	Text
COTE-DP	Opinion Extraction	Restaurant	Text
DuMMSenti	Emotion Classification	Life	Video

The Statistics of LUGE



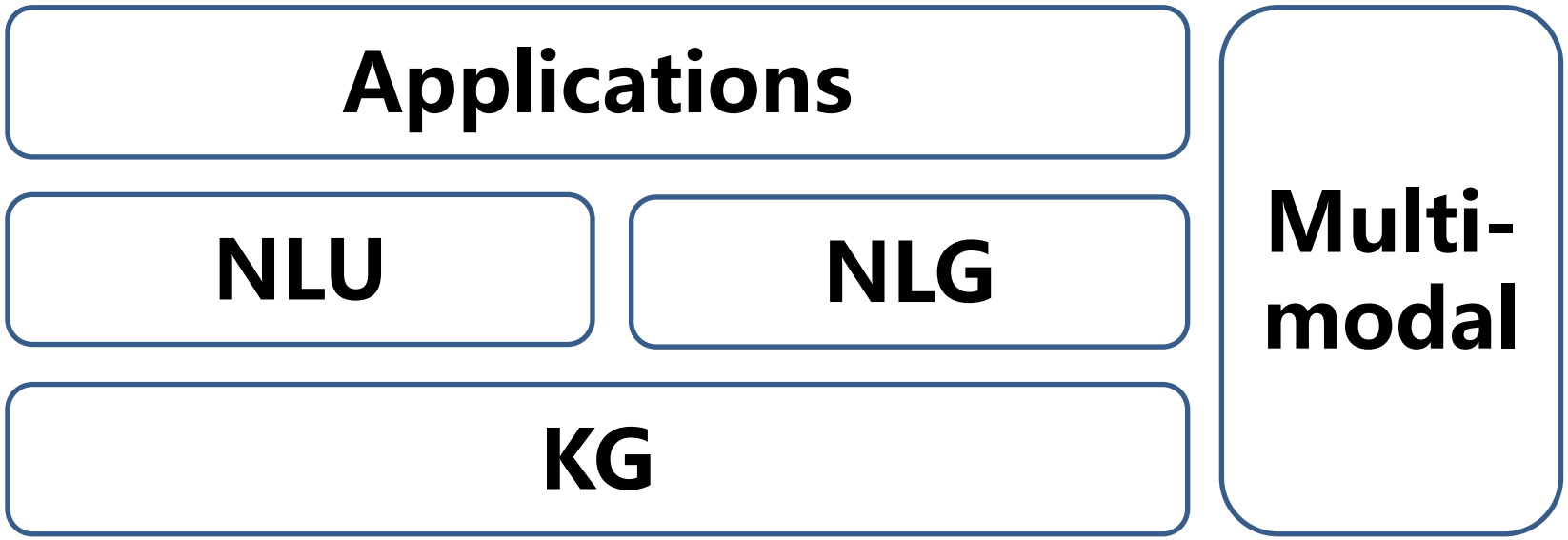
The statistics of tasks and datasets in LUGE



The statistics of downloads and submissions in LUGE

Invite More Authors to Build LUGE Together

3 years **20+** tasks **100+** Chinese NLP datasets



Conclusions and Future Work

- Efforts to deal with the real-world challenges
  - Benchmarks for dialogues, translations, question answering, etc
  - worked together with the community to create **LUGE** to including robustness, faithfulness etc.
- Future Work
  - Define new tasks existing in applications, but have rare researches in academia
  - More tasks for novel multi-modal scenarios

Thanks!

https://luge.ai