MIntRec2.0: A Large-scale Benchmark Dataset for Multimodal Intent Recognition and Out-of-scope Detection in Conversations

Hanlei Zhang*, Xin Wang*, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, Yanting Chen

Task: Multimodal Intent Recognition

Leverage Multimodal Information (text, video, audio) to Recognize Human Intentions in Conversations

Fig 1. Examples from our MIntRec2.0 Dataset.

Challenges in the Literature

Disadvantages in Existing Multimodal Intent Resources:

a. The dataset scale is small.

b. There is a lack of multimodal context and multi-party information.

c. Out-of-scope (OOS) utterances in multimodal conversations are neglected.

Datasets	#I	#U	Conv. Scenes	Conv. Type	OOS	Multi-Party	T	V	A
ATIS (Tür et al., 2010)	17	6,371	✓	Single-turn	✗	✗	✓	✗	✗
Snips (Coucke et al., 2018)	7	14,484	✓	Single-turn	✗	✗	✓	✗	✗
CLINC150 (Larson et al., 2019)	150	23,700	✓	Single-turn	✓	✗	✓	✗	✗
MDID (Kruk et al., 2019)	7	1,299	✗	-	✗	✗	✓	✓	✗
Intentonomy (Jia et al., 2021)	28	14,455	✗	-	✗	✗	✗	✓	✗
MIntRec (Zhang et al., 2022a)	20	2,224	✓	Single-turn	✗	✗	✓	✓	✓
MIntRec2.0	30	15,040	✓	Multi-turn	✓	✓	✓	✓	✓

Tab 1. Comparison between Different Multimodal Intent Datasets.

The MIntRec2.0 Dataset

Key Features:

a. Large scale: MIntRec: 2K→MIntRec2.0: 15K.

b. Multiple speakers: 34 main characters from 3 TV series.

c. Multi-turn conversations: Involves 30 intents and one OOS tag.

d. Multimodal information: Text, video, and audio modalities.

# data sources	3
# intents classes	30
# dialogues	1,245
# utterances	15,040
# in-scope utterances	9,304
# out-of-scope utterances	5,736
# words in utterances	118,477
# unique words in utterances	9,524
Average length of utterances	7.9
Maximum length of utterances	46
Average video clip duration	3.0 (s)
Maximum video clip duration	19.9 (s)
Video hours	12.3 (h)

Tab 2. Statistics of MIntRec2.0.

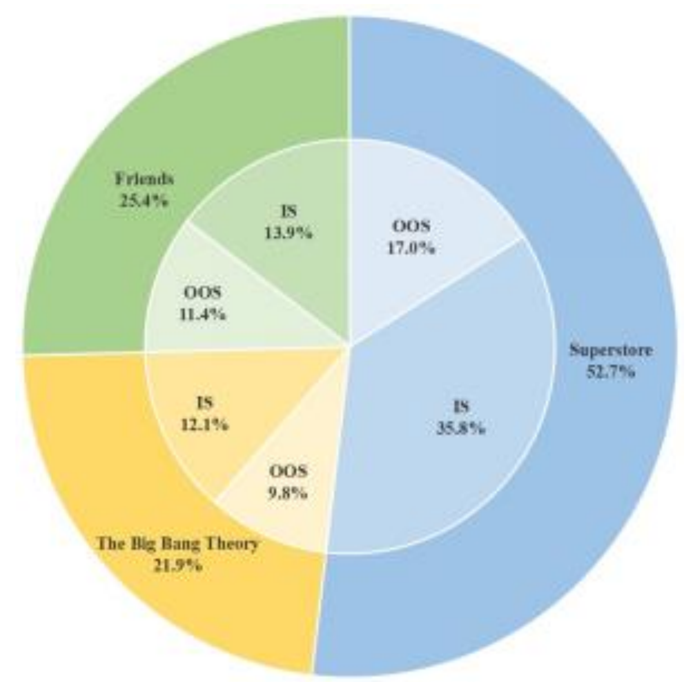

Fig 2. In-scope and Out-of-scope Data Distribution.

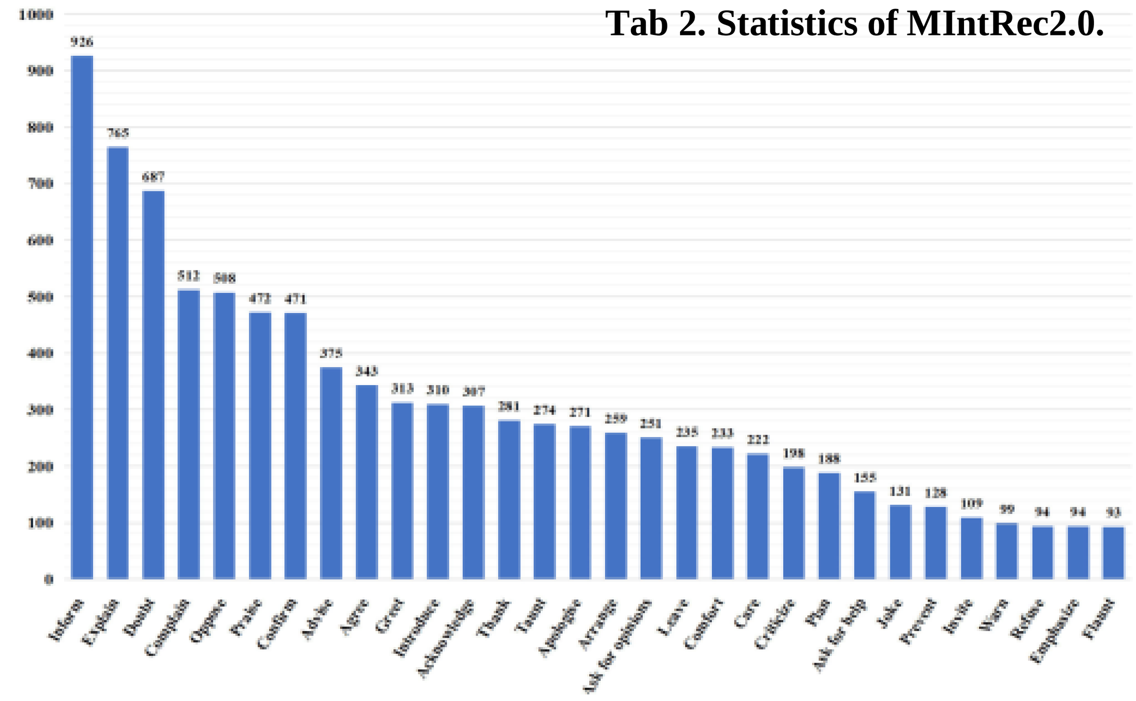

Fig 3. Intent Distribution.

Benchmark Framework

Supports:

a. Processing of multimodal information in both single-turn and multi-turn conversations.

b. Various multimodal fusion methods.

c. In-scope classification and out-of-scope detection.

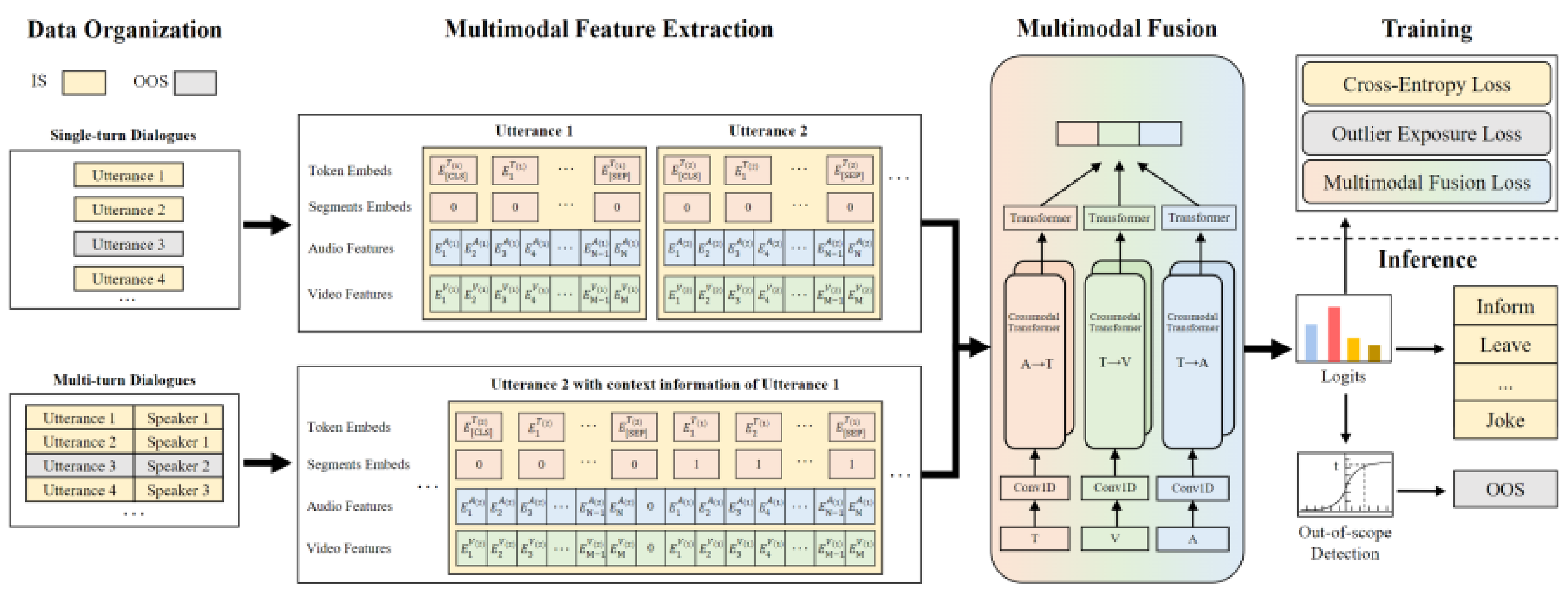

Fig 4. A Pipeline of our Algorithm Framework for Multimodal Intent Recognition.

Experiments and Results

Item	#D	#U	# U (In-scope)	#U (Out-of-scope)
Total	1,245	15,040	9,304	5,736
Training	871	9,989	6,165	3,824
Validation	125	1,821	1,106	715
Testing	249	3,230	2,033	1,197

Tab 3. Data Splits of MIntRec2.0.

	In-scope			In-scope + Out-of-scope		
Methods	ACC	WF1	WP	ACC	F1-OOS	F1
MAG-BERT-10	9.82	11.58	13.34	34.28	50.57	3.75
ChatGPT-0	35.27	37.10	48.22	27.68	21.21	28.34
ChatGPT-10	34.53	36.39	49.27	29.72	27.85	28.41
Humans-10	64.34	67.82	72.80	60.43	62.83	57.83
Humans-100	**71.03**	**75.63**	**81.83**	**71.86**	**75.41**	**69.49**

Tab 4. Performance of ChatGPT and humans.

		In-scope Classification						In-scope + Out-of-scope Classification			
Train	Methods	F1	P	R	ACC	WF1	WP	F1-IS	ACC	F1-OOS	F1
w / o OOS	TEXT	51.60	55.47	51.31	59.30	58.01	58.85	43.37	43.24	30.40	42.96
	MAG-BERT	55.17	57.78	55.10	60.58	59.68	59.98	46.48	44.80	34.03	46.08
	Δ(MAG-BERT)	3.57↑	2.31↑	3.79↑	1.28↑	1.67↑	1.13↑	3.11↑	1.56↑	3.63↑	3.12↑
	MulT	54.12	58.02	53.77	60.66	59.55	60.12	45.65	46.14	38.57	45.42
	Δ(MulT)	2.52↑	2.55↑	2.46↑	1.36↑	1.54↑	1.27↑	2.28↑	2.90↑	8.17↑	2.46↑
w OOS	TEXT	52.08	54.57	52.11	59.99	58.62	58.65	45.83	55.61	61.54	46.34
	MAG-BERT	53.64	54.84	53.79	60.12	59.11	58.83	47.52	56.20	62.47	48.00
	Δ(MAG-BERT)	1.56↑	0.27↑	1.68↑	0.13↑	0.49↑	0.18↑	1.69↑	0.59↑	0.93↑	1.66↑
	MulT	52.72	56.45	52.56	60.18	58.82	59.38	46.88	56.00	61.66	47.35
	Δ(MulT)	0.64↑	1.88↑	0.45↑	0.19↑	0.20↑	0.73↑	1.05↑	0.39↑	0.12↑	1.01↑
w OOS	Context TEXT	53.61	54.46	54.10	59.04	58.69	59.27	46.42	56.12	63.56	46.98
	Context MAG-BERT	53.89	55.72	54.21	59.84	59.41	60.22	46.74	56.20	62.52	47.25
	Δ(Context MAG-BERT)	0.28↑	1.26↑	0.11↑	0.80↑	0.72↑	0.95↑	0.32↑	0.08↑	1.04↓	0.27↑
	Context MulT	53.96	54.91	54.15	59.48	59.33	60.04	46.45	56.07	62.93	46.98
	Δ(Context MulT)	0.35↑	0.45↑	0.05↑	0.44↑	0.64↑	0.77↑	0.03↑	0.05↓	0.63↓	0.00

Tab 5. Benchmark Baseline Results.

Findings:

a. Multimodal information can enhance in-scope classification and out-of-scope detection in single-turn conversations.

b. Existing methods struggle to fully leverage multimodal contexts and out-of-scope utterances as prior knowledge.

c. Humans can achieve significant improvements (over 30% accuracy) compared to ChatGPT in the same few-shot scenarios.

Resources and Contacts

Paper Link

Datasets and codes are Opensourced! Welcome stars and forks!

Github Link

Hi! I am Hanlei Zhang, a fourth-year PhD candidate at Tsinghua University in Beijing, China.

- My research goal is to understand human intentions in real open-world and multimodal scenarios. I have led pioneering work in open intent detection, new intent discovery, and multimodal intent analysis, including the TEXTOIR platform and the MIntRec and MIntRec2.0 datasets.
- I have published seven first-author papers in top-tier AI conferences and journals, including ICLR, IEEE TKDE, ACL, AAAI, ACM MM, and IEEE/ACM TASLP, which have collectively received over 320 citations on Google Scholar.

I'm currently on the job market (expecting to graduate in June 2025). Please do not hesitate to contact me via email at zhang-hl20@mails.tsinghua.edu.cn or through WeChat by scanning the QR code on the right.

WeChat

Homepage