针对视觉基础对话任务、基于空间表达的语言学分析

0 摘要

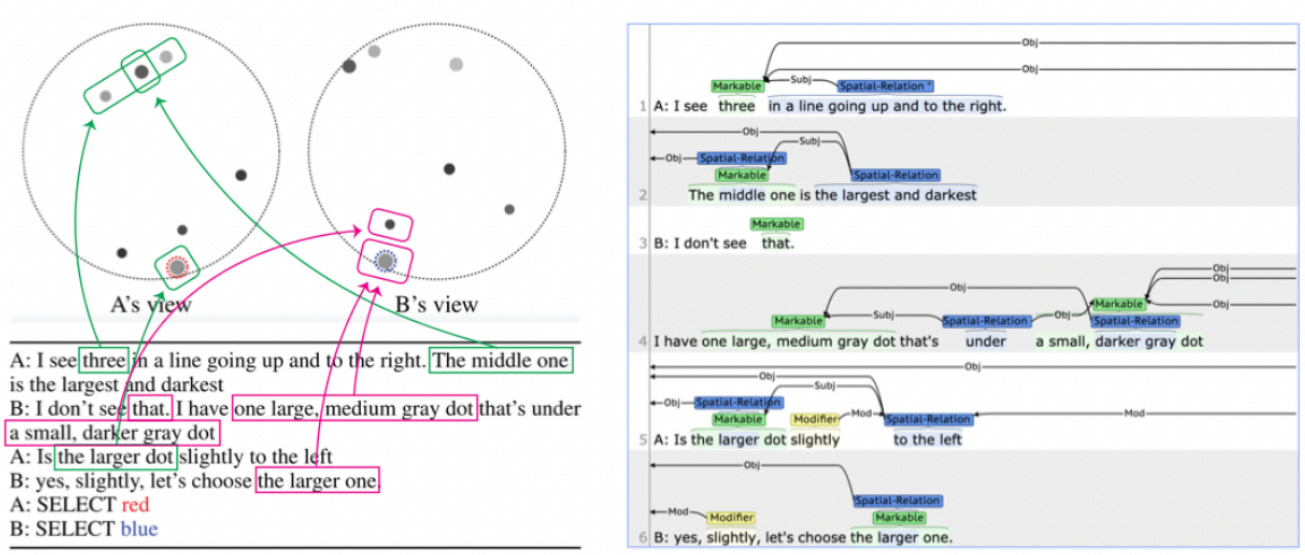
- 近来，机器模型在以视觉为基础的对话中取得了令人鼓舞的结果。然而，现有的数据集往往包含不良的偏差（biases），并且缺乏复杂的语言学分析，这使得我们很难了解当前模型对精确语言结构的识别程度。
- 为了解决这个问题，研究者设计了两个对策：首先，研究者关注OneCommonCorpus语料库（Udagawa和Aizawa，2019、2020），这是一个简单但具有挑战性的对话数据集，在设计上包含尽量小的偏差。其次，研究者基于空间表达（spatial expression）分析该语料库中的语言结构，并为600个对话提供全面可靠的标注。
- 研究者的标注捕捉到了重要的语言结构，包括谓语-论元结构、修饰和省略语。在研究者的实验中，研究者通过指代消解（reference resolution）来评估该模型对这些结构的理解。
- 结果证明，研究者的标注可以揭示基线模型在一些关键细节上的优势和劣势。总的来说，研究者提出了一个新的框架和资源，用于研究视觉基础对话中细粒度的语言理解。

1 引言

- 视觉对话（Visual Dialogue）是在视觉背景（visual context）下进行自然的、通常是目标导向的对话的任务。

    - VisDial数据集（Das等，2017）

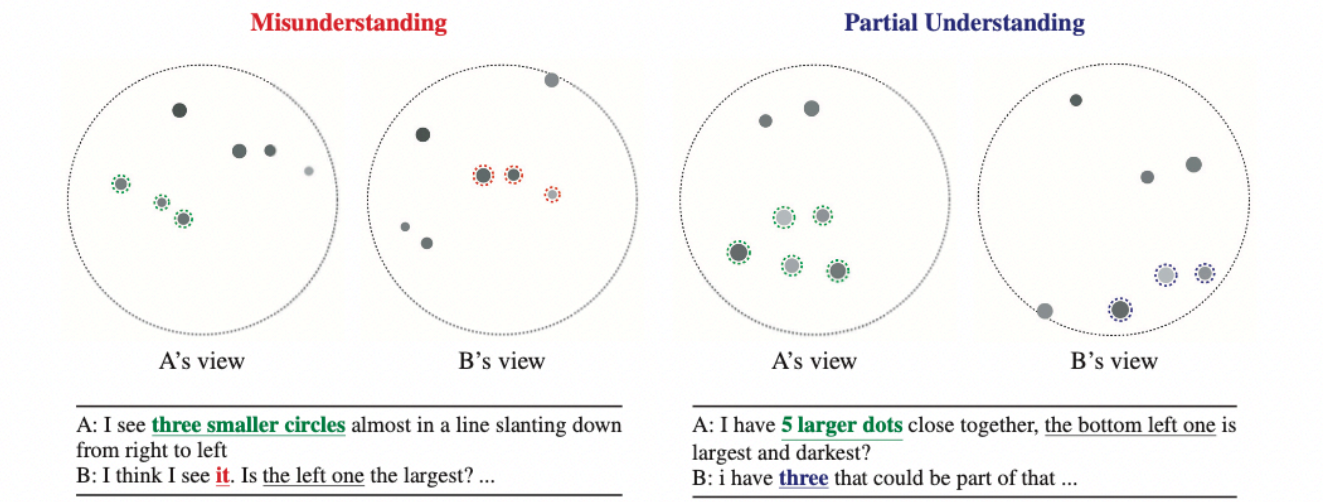
- 这类任务通常涉及两种类型的高级基础能力（advanced grounding）：符号接地（symbol grounding），它连接了符号化的自然语言和连续的视觉感知；以及共同接地（common grounding），它指的是通过连续的对话发展相互理解的过程。
- 视觉背景的连续性引入了涉及细微差别和实用性表达的、有挑战性的符号接地问题。有些进一步引入了部分可观察性（partial observability），即说话人不共享完全相同的背景信息，需要高级的共同接地才能解决这类复杂的误解问题。
- 尽管最近这些任务上取得了进展，但仍然不清楚哪些类型的语言结构可以（或不能）被现有模型正确识别，原因有二。首先，现有的数据集往往包含不良的偏差，这使得在不识别精确的语言结构的情况下做出正确的预测成为可能。其次，现有的数据集严重缺乏复杂的语言学分析，这使得研究者很难理解存在哪些类型的语言学结构，或者它们如何影响模型的性能。
- 研究者专注于OneCommonCorpus语料库，这是在连续和部分可观察语境（continuous and partially-observable context）下的一个简单但具有挑战性的协作指代任务（collaborative referring task）。在这个数据集中，视觉语境保持简单和可控，以消除不希望出现的偏差，同时增强语言的多样性。总共收集了5191段自然对话，并对指代表达（他们称之为markables）及其所指进行了充分的标注，这些可以被用来进行进一步的语言学分析。
- 为了捕捉这些对话中的语言结构，研究者标注了空间表达，这些表达在以视觉为基础的对话中起到了核心作用。研究者从现有的注释框架（Pustejovsky等人，2011a、2011b；Petruck和Ellsworth，2018；Ulinski等人，2019）中得到启发，但也做了一些简化和修改以提高覆盖率、效率和可靠性。



- ◦ OneCommonCorpus对指代消解的标注（左）及研究者对空间表达的标注（右）

2 OneCommonCorpus

- 研究者的工作扩展了Udagawa和Aizawa（2019）提出的OneCommon Corpus。在原数据集中，对话者A和B分别获得了一个稍微不同的、视角重叠的二维图像，每个人可以看到7个实体。由于其中只有一些（4、5或6个）实体是相同的，这种部分可观察的（partially-observable）设定会带来复杂的误解和部分理解（misunderstandings and partial understandings）。此外，每个实体只有连续（continuous）的属性（X值、Y值、颜色和大小），这就引入了各种细微的（nuanced）和实用的（pragmatic）表达。



- ◦ 误解和部分理解示例
- 所有的实体属性都是随机生成的，以增强语言的多样性，减少数据集的偏差。两个亚马逊众包平台的标注者被要求用自然语言在线自由交谈，以获得对相同实体的共同注意力。

Total dialogues	6,760
Avg. utterances per dialogue	4.76
Avg. tokens per utterance	12.37
Successful dialogues	5,191
Annotated markables	40,172
% markables with 1 referent	71.81
% markables with 2 referents	14.85
% markables with $\geq$3 referents	12.03
% markables with 0 referent	1.31

- OneCommonCorpus数据

Nuance Type	% Utterance	Example Usage
Approximation	3.98	**almost** in the middle
Exactness	2.71	**exactly** horizontal
Subtlety	9.37	**slightly** to the right
Extremity	9.35	**very** light dot
Uncertainty	5.79	**Maybe** it's different

- OneCommonCorpus中精细表达（nuanced expressions）的统计
- Udagawa和Aizawa（2020）进一步对数据集进行了指代消解方面的标注。具体来说，他们要求标注者识别指代表达（referring expressions）即markables，和它们的所指。
- 在本文中，研究者从最新的语料库中随机抽出600段对话，对空间表达进行进一步的标注。

3 标注

3.1 标注过程

研究者的标注过程包括三个步骤：空间表达检测（spatial expression detection）、论元识别（argument identification）和标准化（canonicalization）。

- A 空间表达检测
  - 根据Pustejovsky等人（2011a、2011b）的定义，空间表达是 "明确指代一个物体的空间属性或物体之间的空间关系的语言结构"。研究者大体上遵循这个定义，识别对话中所有的空间属性和空间关系表达。
  - 为了明确区分，研究者将颜色、大小等实体层面的信息视为空间属性，而将位置和明确的属性比较等其他信息视为空间关系。空间属性可以是形容词（"dark"）、介词短语（"of light color"）或名词短语（"a black dot"），空间关系可以是形容词（"lighter"）、介词（"near"）等等。

- B 论元识别
  - 研究者将识别出的空间表达式视为谓词（predicates），而将指代表达式（markables）标注为他们的论元（argument）。
- C 标准化
  - 研究者最后对空间表达式和修饰语进行标准化。研究者只关注核心空间关系的标准化。
  - 根据Landau（2017），空间语言中有2类关系：功能类（functional class），其核心含义为力-动态关系（force-dynamic relationship）（如on，in）；几何类（geometry class），其核心含义为几何（如left，over）。由于功能关系在这个数据集中不太常见，而且由于其模糊性和上下文依赖性而更难定义（Platonov和Schubert，2018），研究者专注于以下5类几何关系和属性比较，包括总共24种可以明确定义的标准关系。
    -

类别	描述	关系子类
Direction	requires the subjects and objects to be placed in certain orientation	left, right, above, below, horizontal, vertical, diagonal
Proximity	is related to distance between subjects, objects or other entities	near, far, alone
Region	restricts the subjects to be in a certain region specified by the objects	interior, exterior
Color comparison	is related to comparison of color between subjects and objects	lighter, lightest, darker, darkest, same color, different color
Size comparison	is related to comparison of size between subjects and objects	smaller, smallest, larger, largest, same size, different size

  - 研究者对识别到的每个空间关系是否蕴含（imply）24种标准关系中的任何一种进行标注。每个空间关系都可以蕴含多个标准关系（例如 "on the upper right "蕴含right和above），也可以不蕴含（例如 "triangle "不蕴含上述任何关系）。
  - 此外，研究者定义了6种修饰类型（subtlety, extremity, uncertainty, certainty, neutrality and negation），并将每个修饰语标准化为一种类型。例如，"very slightly "被认为是类型subtlely。

3.2 结果

- 标注一致性

Annotation	% Agreement	Cohen's κ
Attribute Span	98.5	0.88
Relation Span	95.1	0.87
Modifier Span	99.2	0.86
Subject Ident.	98.8	0.96
Object Ident.	95.9	0.79
Modificand Ident.	99.6	0.98
Relation Canon.	99.7	0.96
Modifier Canon.	87.5	0.83

    - 标注一致性

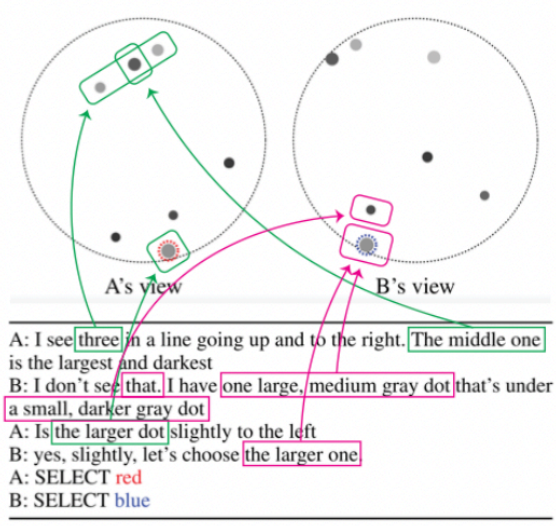
- 标注数据统计

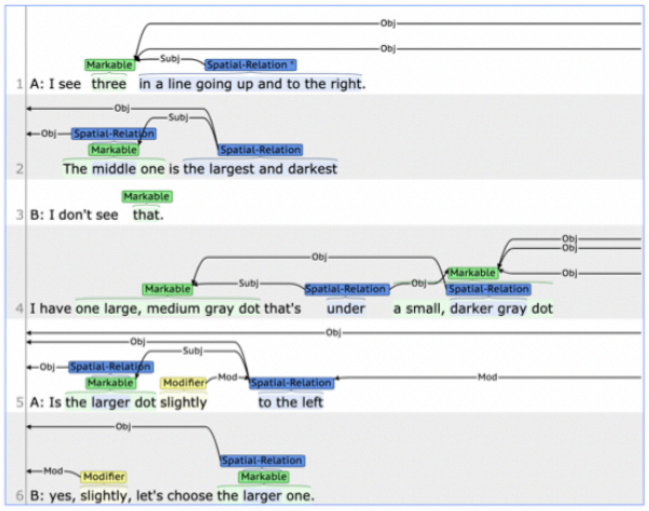
	Attribute	Relation
Total	378	4,300
Unique	121	1,139
Avg. per dialogue	0.63	7.17
% inter-utterance subject	1.59	1.37
% inter-utterance object	-	14.65
% no object	-	30.84
% modified	36.51	16.86
% unannotatable	0.79	0.79

    - 标注数据统计

4 实验

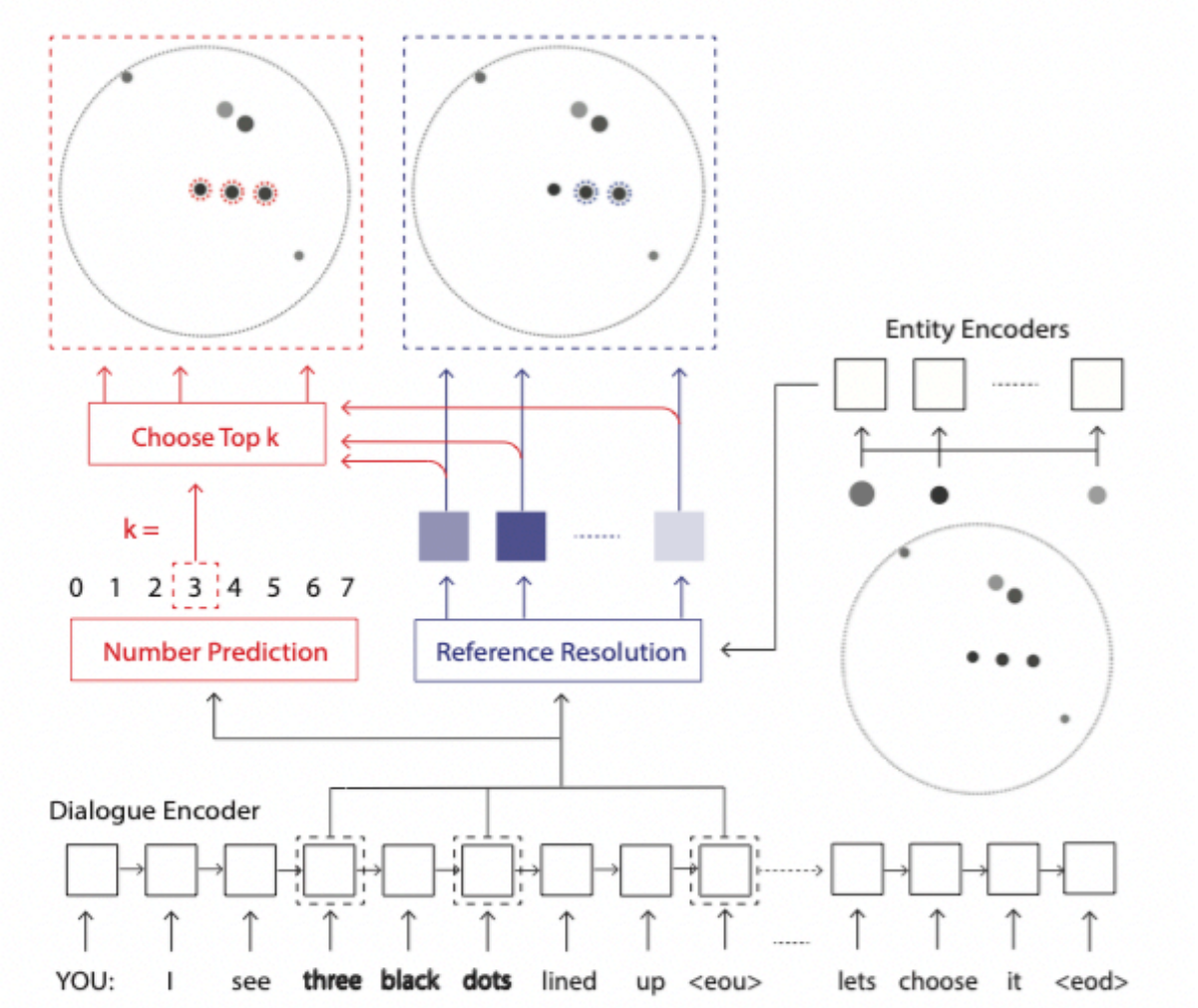
- 任务设计





- 指代消解任务

- 模型设计
  - 基线模型
    -



  - 实验结果（消融实验）
    -

	Entity-Level Accuracy	Markable-Level Exact Match
REF	85.71±0.23	33.15±1.00
REF−location	84.28±0.27	30.53±0.84
REF−color	83.08±0.32	17.09±1.04
REF−size	83.50±0.22	19.41±0.98
NUMREF	**86.03±0.33**	**54.94±0.76**
NUMREF−location	83.35±0.26	49.77±0.64
NUMREF−color	81.19±0.41	39.74±1.31
NUMREF−size	82.39±0.20	43.40±0.67
Human	96.26	86.90

- 模型分析
  - 为了详细研究模型的长处和短处，研究者分析了模型预测结果与核心空间表达的一致性。
  - 空间属性
    -

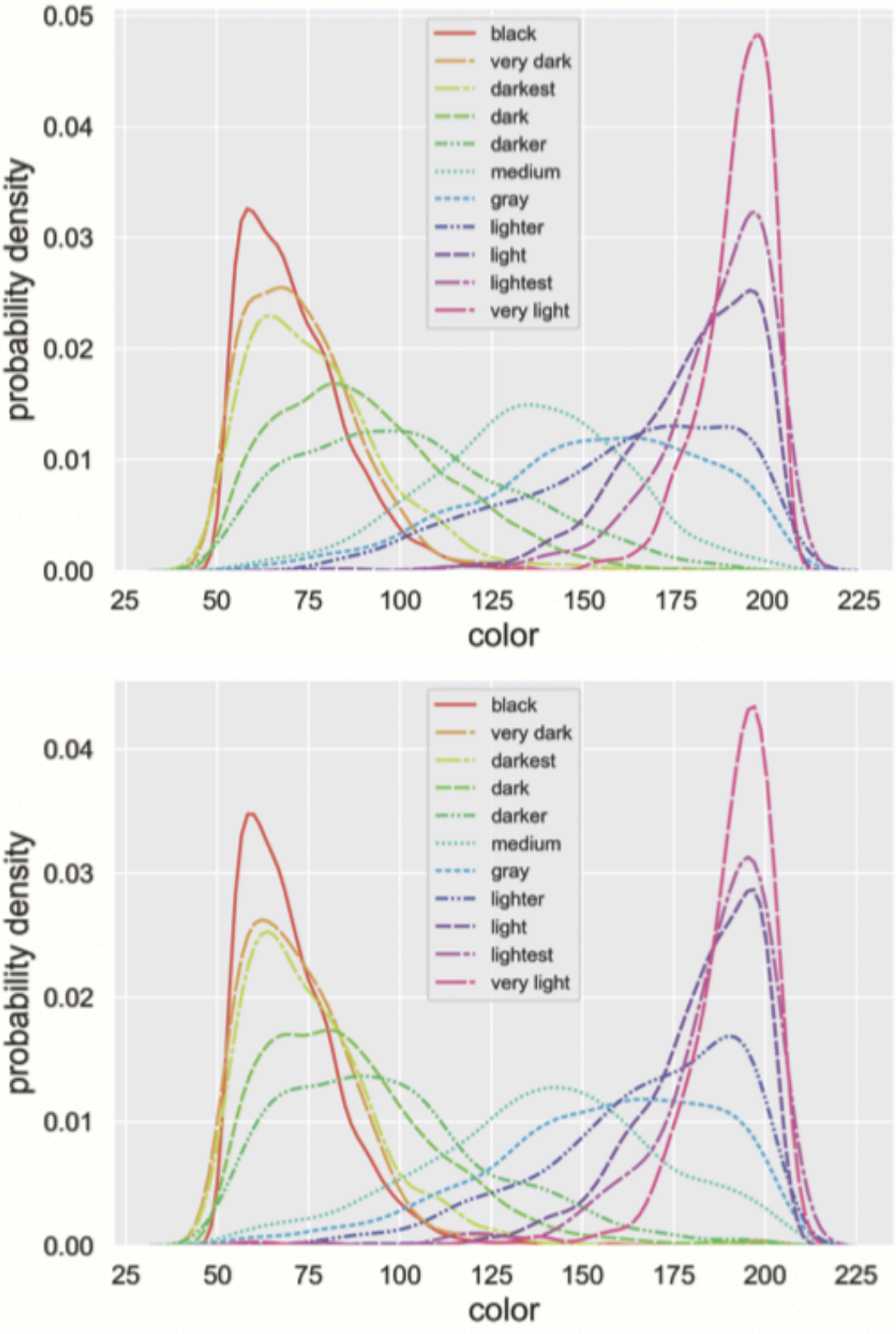


Figure 3: Referent color distributions. Top is human, bottom is NUMREF (smaller is darker in color axis).

- 空间关系
  -

Algorithm 1: Test for *left* relation

Input: subject referents $\mathcal{S}$, object referents $\mathcal{O}$, boolean no_object

Output: boolean $satisfy$, boolean $valid$

if no_object **then**

$valid \leftarrow |\mathcal{S}| > 0$

$satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < 0$

else

$valid \leftarrow |\mathcal{S}| > 0 \wedge |\mathcal{O}| > 0$

$satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < mean(\mathcal{O}.x)$

return $satisfy, valid$

-

Models			REF		REF-abl		NUMREF		NUMREF-abl		Human	
Category	Relation	# Cases	satisfy	valid	satisfy	valid	satisfy	valid	satisfy	valid	satisfy	valid
Direction	*left*	412	23.5	32.3	21.1	28.9	**67.0**	**99.5**	62.4	**99.5**	95.9	97.6
	right	468	28.0	35.5	24.6	30.8	67.3	**98.7**	**68.2**	**98.7**	95.3	96.4
	above	514	28.6	37.4	24.7	33.1	65.2	99.2	**66.5**	**99.4**	96.7	98.6
	below	444	25.2	34.5	21.6	27.9	**66.0**	**99.1**	62.2	**99.1**	96.4	96.8
	horizontal	37	54.1	70.3	27.0	59.5	**59.5**	**100.0**	51.4	97.3	91.9	100.0
	vertical	46	37.0	73.9	23.9	54.3	43.5	**95.7**	**45.7**	**95.7**	82.6	100.0
	diagonal	50	48.0	74.0	30.0	50.0	**60.0**	**98.0**	**60.0**	**98.0**	90.0	100.0
	All	1,971	27.8	37.6	23.4	31.9	**65.5**	**99.0**	64.1	**99.0**	95.5	97.6
Proximity	*near*	271	49.4	61.3	29.9	49.1	**77.1**	94.5	56.1	**95.2**	95.2	96.7
	far	27	29.6	40.7	33.3	40.7	77.8	**100.0**	**92.6**	**100.0**	96.3	96.3
	alone	111	36.9	44.1	45.0	54.1	**68.5**	**94.6**	67.6	**94.6**	91.9	94.6
	All	409	44.7	55.3	34.2	49.9	**74.8**	94.9	61.6	**95.4**	94.4	96.1
Region	*interior*	135	38.5	52.6	27.4	39.3	**62.2**	93.3	58.5	**94.1**	96.3	100.0
	exterior	62	40.3	48.4	40.3	53.2	80.6	**98.4**	**87.1**	**98.4**	98.4	98.4
	All	197	39.1	51.3	31.5	43.7	**68.0**	94.9	67.5	**95.4**	97.0	99.5
Color	*lighter*	147	23.1	25.9	6.8	8.2	**84.4**	**100.0**	57.1	99.3	97.3	98.0
	lightest	42	45.2	66.7	14.3	33.3	**61.9**	**100.0**	31.0	**100.0**	83.3	100.0
	darker	171	24.0	26.3	7.0	10.5	**83.0**	**99.4**	53.2	**99.4**	95.9	98.8
	darkest	48	56.2	64.6	14.6	33.3	**66.7**	**100.0**	35.4	**100.0**	89.6	97.9
	same	50	12.0	30.0	8.0	30.0	**40.0**	**88.0**	32.0	86.0	92.0	96.0
	different	14	64.3	71.4	71.4	71.4	64.3	**100.0**	**78.6**	92.9	92.9	100.0
	All	472	28.8	35.4	10.4	18.0	**74.8**	**98.5**	49.2	97.9	94.1	98.3
Size	*smaller*	213	27.7	31.5	7.5	9.9	**80.8**	**100.0**	59.6	**100.0**	98.6	99.5
	smallest	52	71.2	73.1	21.2	34.6	**86.5**	**98.1**	48.1	**98.1**	92.3	98.1
	larger	238	23.1	28.6	9.7	16.0	**73.5**	**99.6**	48.7	**99.6**	98.3	98.3
	largest	61	52.5	60.7	11.5	24.6	**73.8**	**100.0**	39.3	**100.0**	96.7	100.0
	same	103	34.0	42.7	18.4	27.2	**80.6**	88.3	65.0	**91.3**	98.1	100.0
	different	12	75.0	75.0	66.7	66.7	**91.7**	**91.7**	83.3	83.3	91.7	91.7
	All	679	33.4	38.7	12.4	18.9	**78.2**	97.8	54.3	**98.1**	97.6	99.0

Table 6: Canonical relation test results. We compute the *satisfy* and *valid* rate of the predictions for each canonical relation. Best scores of the models are in bold (-abl shows the corresponding feature ablated results).

- 进一步的分析
  -

Linguistic Factors	# Cases	NUMREF	Human
strong modification	149	76.51	95.97
neutral	3,094	70.46	95.77
weak modification	490	66.12	95.10
inter-utterance subject	14	57.14	92.86
inter-utterance object	265	72.08	94.72
no object	1,127	74.45	92.99
ignorable object	1,805	69.64	97.23
unignorable object	796	65.33	96.11
All	3,728	70.17	95.71

Table 7: Satisfy rate classified by linguistic factors.

Models		NUMREF		Human	
value	mod-type	diff.	# valid	diff.	# valid
xy-value	strong	86.06	39	89.15	37
	neutral	80.92	1,586	73.52	1,558
	weak	80.35	200	53.53	198
color	strong	66.23	15	91.80	15
	neutral	56.98	234	60.14	232
	weak	37.73	68	28.55	66
size	strong	3.60	8	4.29	8
	neutral	2.67	337	2.70	320
	weak	1.95	105	1.58	104

Table 8: Absolute difference in comparative relations (number of valid predictions shown in shade).

5 讨论和结论

- 对以视觉为基础的对话的标注
- 可拓展的语言学分析框架
- 项目主页：https://github.com/Alab-NII/onecommon

讨论记录

- 报告了收录在EMNLP 2020 findings的一片论文《针对视觉基础对话任务、基于空间表达的语言学分析（Annotated Corpus of Reference Resolution for Interpreting Common Grounding）》，论文从NLP领域以视觉为基础的对话任务（1）数据中存在偏差（biases），（2）缺少深入的语言学分析，两个问题出发，在OneCommon Corpus数据集的基础上进行空间语义标注。与传统上分析模型在测试集上表现的正确率等统计指标不同，作者将分析重点放在语言学视角上模型的表现（即空间语义标注情况）上。作者分析了模型对标注体系里空间属性和空间关系两类标注对象的识别情况，其中对于空间属性，作者采用比较模型标注结果和人类表现一致性的方法，对于空间关系，作者定义了有效（valid）和满足（satisfy）两步标准，依次计算了模型对24类空间关系识别的有效率和满足率，据此分析得出模型在不同类别空间关系上表现情况的优劣。作者还对修饰语分级考察，得出模型对强烈倾向修饰语（strong modification）的识别效果更好。
- 讨论记录
    - 关于指代消解（reference resolution）任务，一般认为是纯文本的、符号与符号之间的任务，OneCommon Corpus修订后标注的图形与文字对应关系及据此提出的任务不是传统意义上的指代消解任务。
    - 符号接地（symbolic grouding）问题是一个经典问题，主要涉及符号如何落地（grounding）、获得意义的问题。
    - 这篇论文的主要insight有二：（一）提出一个空间语义标注体系，（二）针对如何分析模型表现提出了一个语言学视角的分析框架。