Tutorial on Spatial Language Understanding:
Representation, Reasoning, and Grounding

Parisa Kordjamshidi, Michigan State University, USA, kordjams@msu.edu

Marie-Francine Moens, KU Leuven, Belgium, sien.moens@cs.kuleuven.be

James Pustejovsky, Brandeis University, USA, jamesp@cs.brandeis.edu

The 29th Conference on Computational Linguistics
COLING-2022

Oct 12th-17th

Gyeongju, Republic of Korea

Learning | Reasoning

Heterogeneous

Introduction

Spatial Language Challenges

"Hi! You are just on time! Please get me a piece of cake.

It's in the kitchen. Go out to the hall; you will see a door with a table on it. It's on the kitchen's table. A plate is under the counter, in the drawer. Utensils are next to it. There are also tissue papers above the table.

Be careful! there will be a vase on the ground on your left … .

…

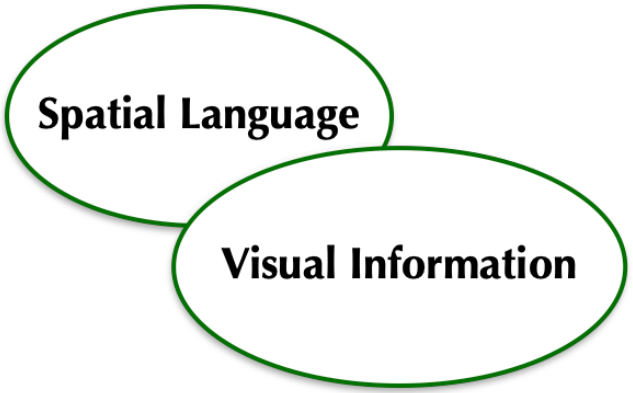
Great! You are on top of it!"

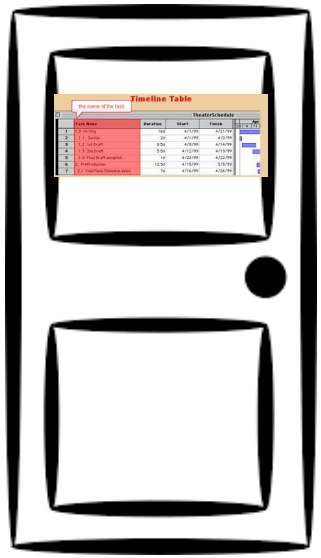
- Lexical variability
- Structural variability
- Polysemy
- Ambiguity
- Discourse and Common Sense

Spatial Language Challenges

"Hi! You are just on time! Please get me a piece of cake. It's in the kitchen. Go out to the hall; you will see a door with a table on it. It's on the kitchen's table.

A plate is under the counter, in the drawer. Utensils are next to it. There are also tissue papers above the table. Be careful! there will be a vase on the ground on your left … . Great! You are on top of it!"



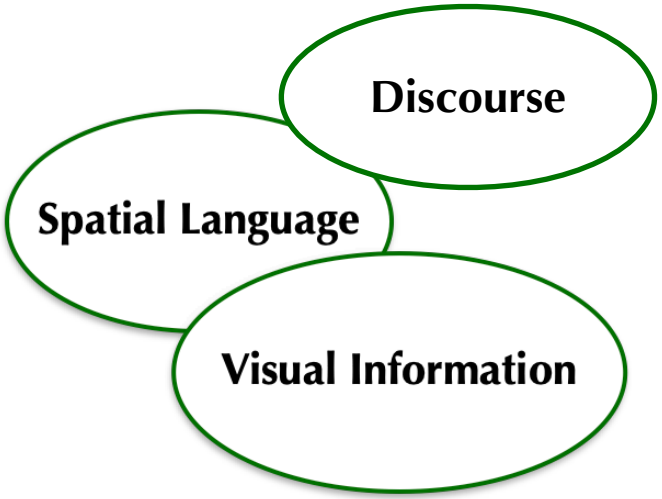


Spatial Language Challenges

"Hi! You are just on time! Please get me a piece of cake. It's in the kitchen. Go out to the hall; you will see a door with a table on it. It's on the kitchen's table.

A plate is under the counter, in the drawer. Utensils are next to it.

There are also tissue papers above the table. Be careful! there will be a vase on the ground on your left … . Great! You are on top of it!"



Spatial Language Challenges

Complex Linguistic Utterances

I: Complex locative statements

The vase is in the living room, on the table under the window.

II: Sequential scene descriptions

Behind the shops is a church, to the left of the church is the town hall, in front of the town hall is a fountain.

III: Path and route descriptions

The man came from between the shops, ran along the road and disappeared down the alley by the church.

[Barclay, Michael & Galton, Antony. (2008). A Scene Corpus for Training and Testing Spatial Communication Systems.]

Spatial Language Challenges

Implicit spatial semantics

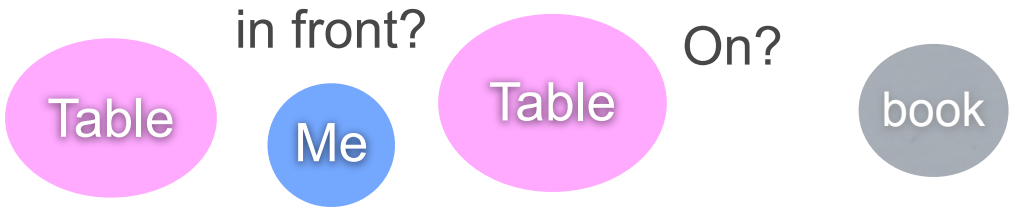
Put the milk in the coffee vs. Put the milk in the refrigerator

Fly a kite vs. Carry a kite

Spatial Language Applications

Navigation Instruction Following

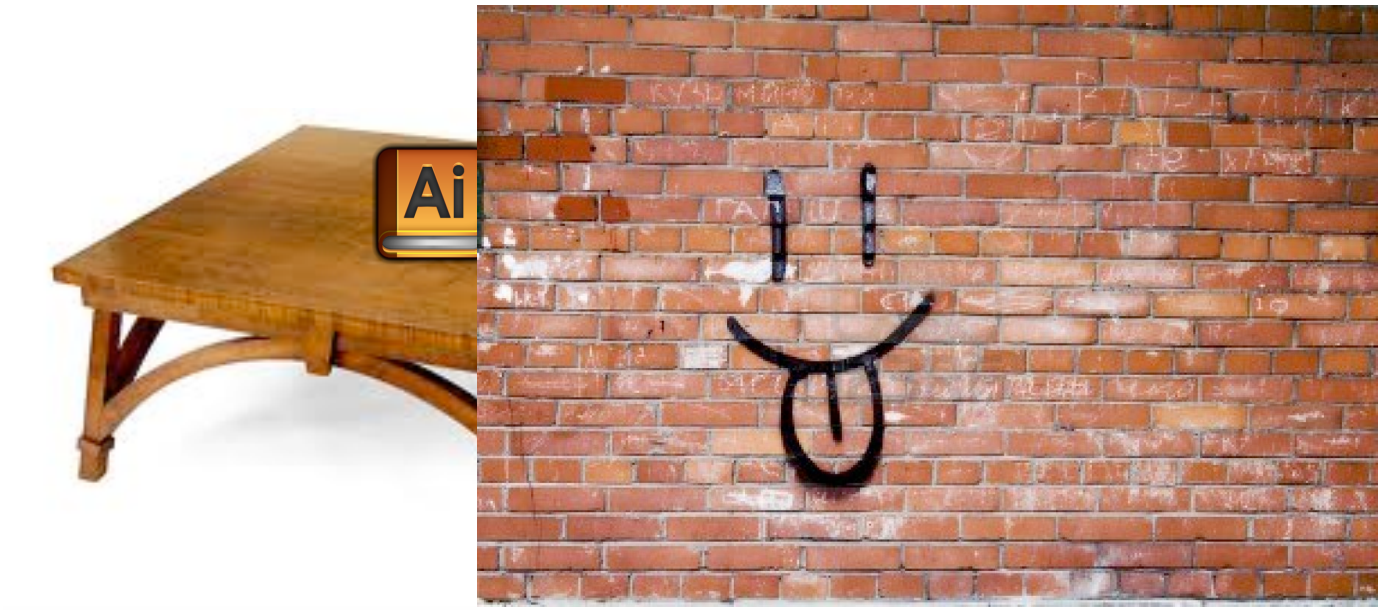
"Give me the book on AI on the big table in front of you!"



Spatial Language Applications

Text to Scene conversion (Visualization)

“The book on AI is on the big table behind the wall.”



Spatial Language Applications

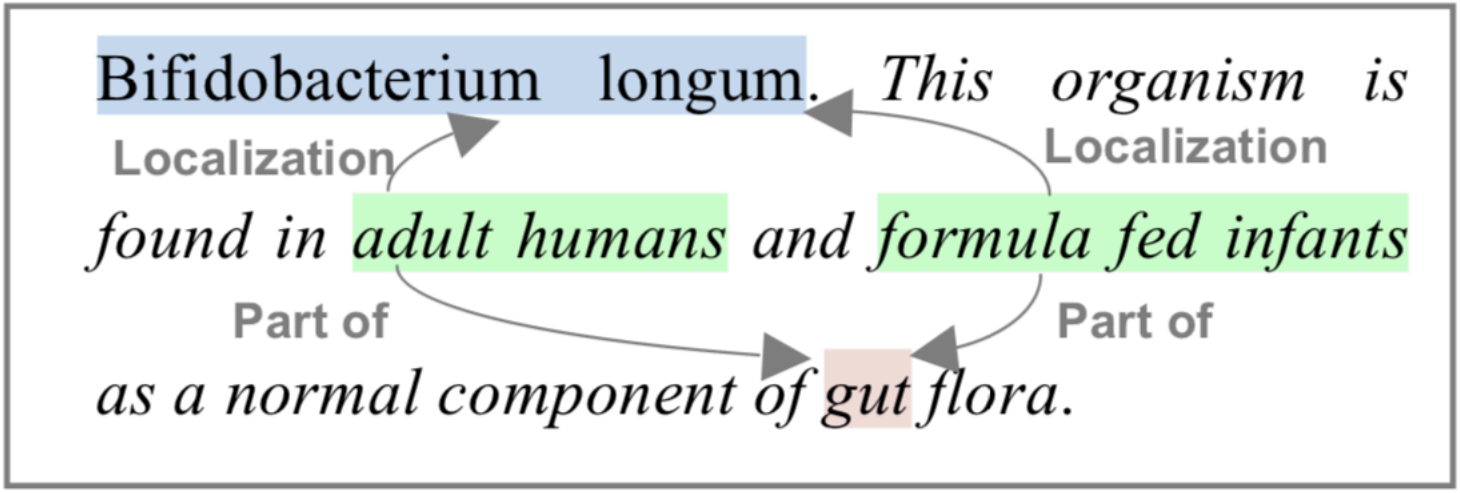
Scene to Text conversion (Image captioning)

ref: Google images, dpreview.com

Spatial Language Applications

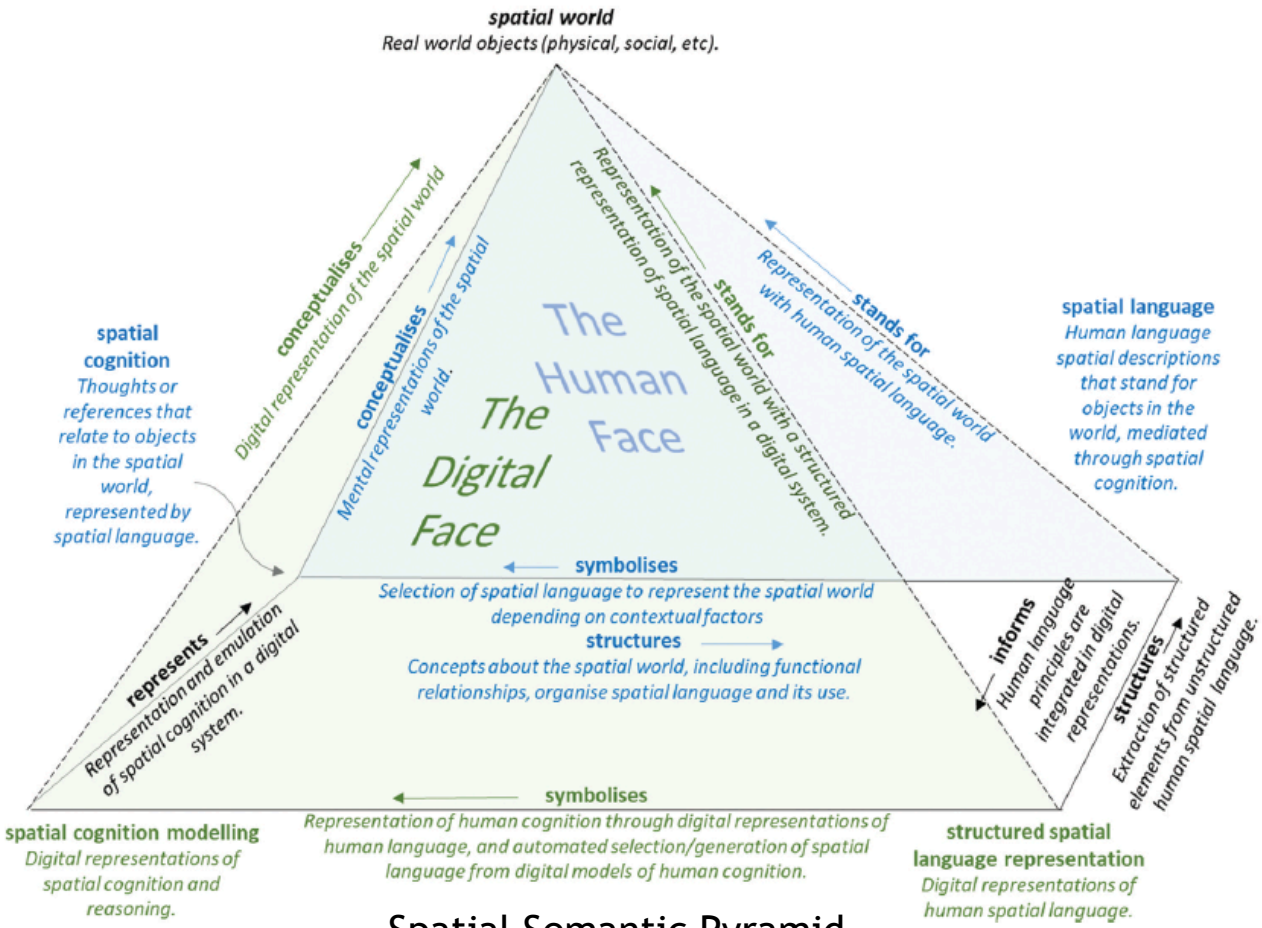
Scientific text: Biomedical Domain

- Whether Bacteria X can live in human body?
- What are the habitats of Bacteria Y?
- What kind of Bacteria can be found in home made Yogurt that do not live in commercial Yogurt?



[Kordjamshidi, Roth, Moens,. BMC-BioInformatics. Structured learning for spatial information extraction from biomedical text: bacteria biotopes, 2015.]

Speaking of Location



Spatial Semantic Pyramid

Kristin Stock, Christopher B. Jones, and Thora Tenbrink. Speaking of location: a review of spatial language research. Spatial Cognition & Computation, 0(0):1–40, 2022.

Table of Content

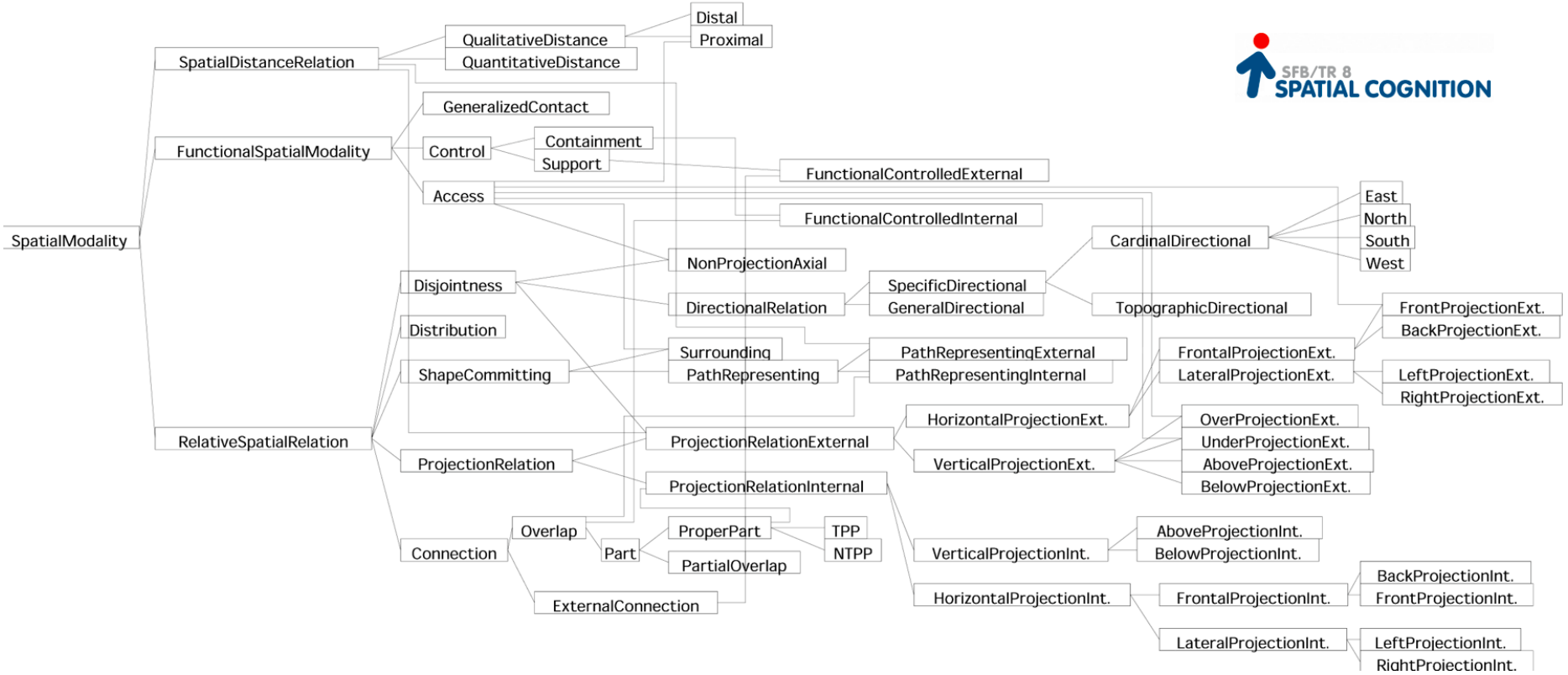
Spatial Information Representation

Spatial Representation

- Symbolic Semantic Representations
  - Cognitive Linguistic Conceptualizations
  - Spatial Knowledge Representation and Reasoning
- Continuous Representations
  - Learning Representations (corpora and sources of supervision)
  - Vision and Language Models

Linguistically motivated representations

General Upper Model (GUM) ontology

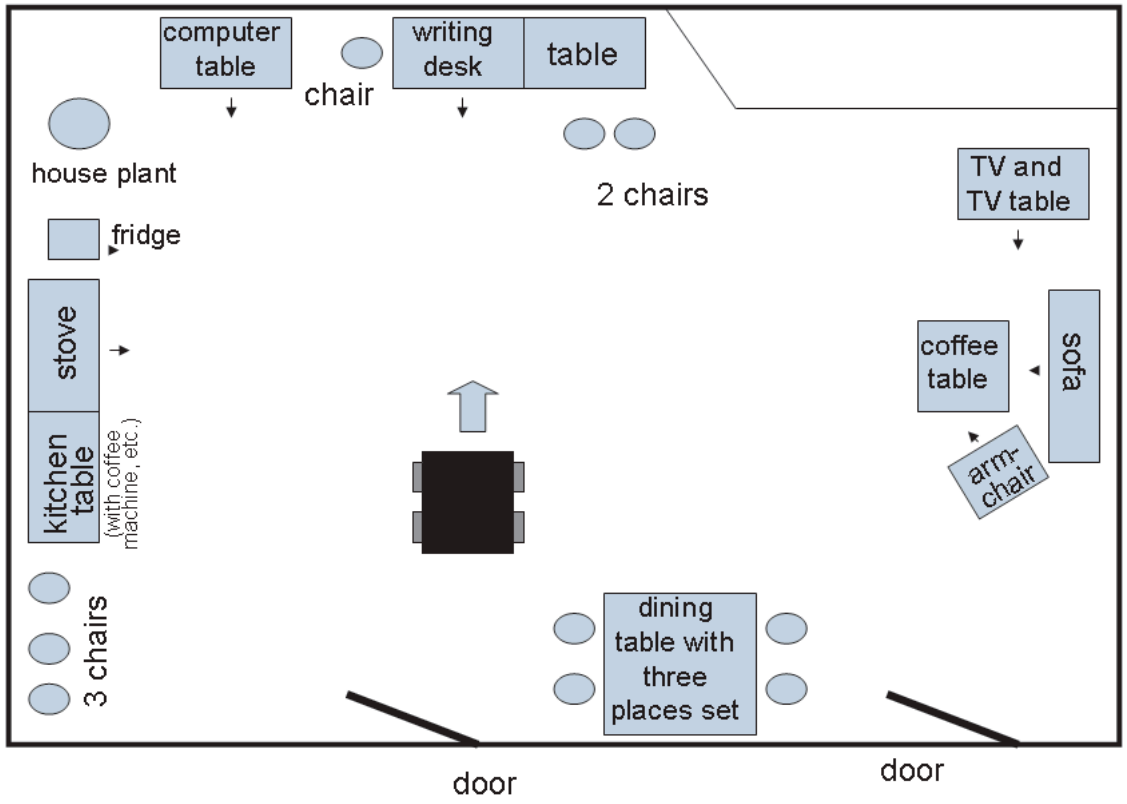


[J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink. A linguistic ontology of space for natural language processing. Artificial Intelligence, 174(14):1027–1071, 2010.]

[L. Talmy, The fundamental system of spatial schemas in language, in: From Perception to Meaning: Image Schemas in Cognitive Linguistics, Mouton de Gruyter, Berlin, 2006, pp. 37–47.]

[M. Bierwisch, How much space gets into language, in: P. Bloom, M.A. Peterson, L. Nadel, M.F. Garrett (Eds.), Language and Space, MIT Press, Cambridge,MA, 1999, pp. 31–76.]

Linguistically motivated representations



1. so from here exactly opposite is my desk.

2. and next to that left of that is my computer, perhaps a meter away.

3. (breathing) ähm.

4. next to that at the wall is my kitchen, first there is my fridge all the way to the right.

[J. Bateman, T. Tenbrink, and S. Farrar. The role of conceptual and linguistic ontologies in discourse. Discourse Processes , 44(3):175–213, 2007.]

Linguistically motivated representations

so from here exactly opposite is my desk…and next to that left of that is my computer, perhaps a meter away…

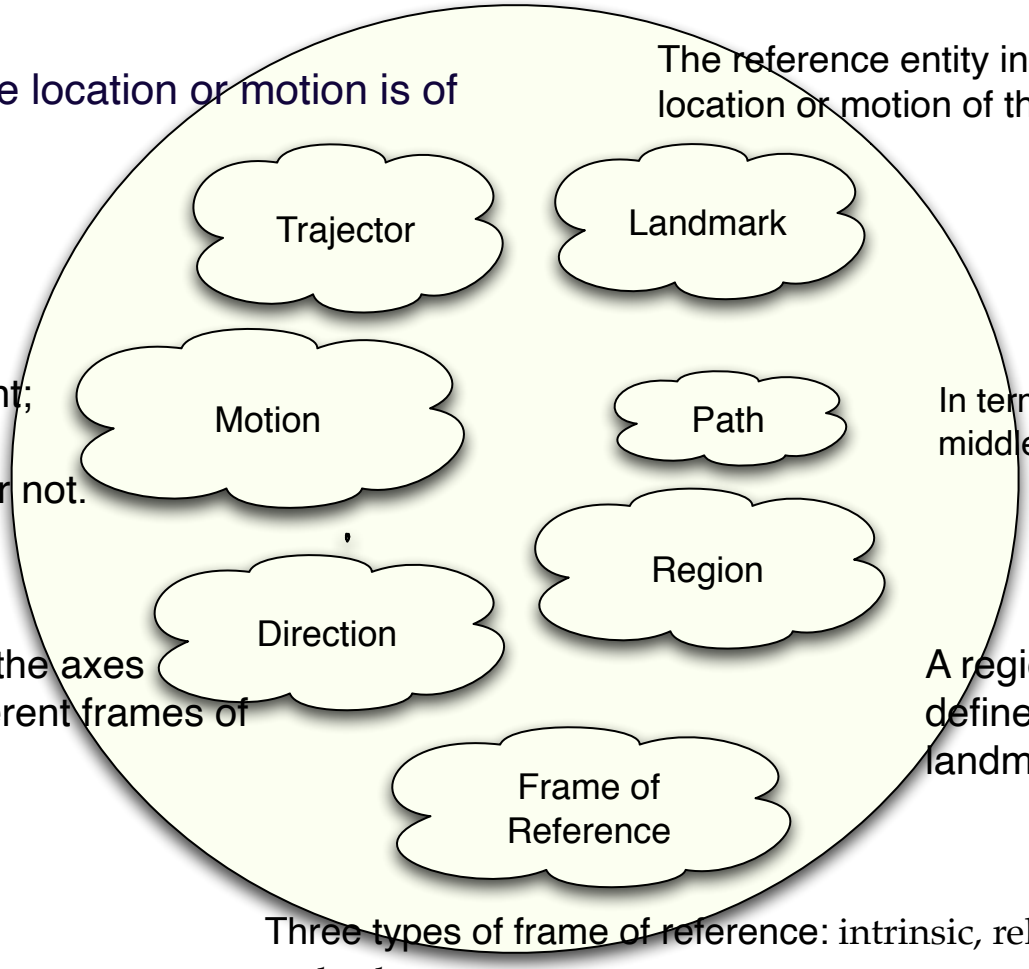
Utterance	*Locatum*	*Relatum*	*GUM Category*
1	Desk	Self	NonprojectionAxial: opposite
2	Computer	Desk	LeftProjectionExt [distance: 1m]
4	Kitchen	Computer	HorizontalProjectionExt: next
4	Kitchen	Wall	ExternalConnection: at
4	Fridge	Kitchen	RightProjectionInt: rightmost
4	Fridge	Corner	Containment: in
5	Houseplant	Corner	Containment: in
6	Stove	"There"	ExternalConnection: at
6	{Stove, kitchen table}	Fridge	HorizontalProjectionExt: side of
6–7	{Stove, kitchen table}	Fridge	LeftProjectionExt
9	Entrance	Self	BackProjectionExt
10	Dining	table	Self RightProjectionExt

SFB/TR 8 SPATIAL COGNITION

Holistic Spatial Semantics

The entity whose location or motion is of relevance.

The reference entity in relation to which the location or motion of the trajector is determined.



A binary component; whether there is perceived motion or not.

In terms of its beginning, middle and end.

The direction along the axes provided by the different frames of reference.

A region of space which is defined in relation to a landmark.

Three types of frame of reference: intrinsic, relative or absolute.

[J. Zlatev. Spatial semantics. In D. Geeraerts and H. Cuyckens, editors, The Oxford Handbook of Cognitive Linguistics , pages 318–350. Oxford Univ. Press, 2007.]

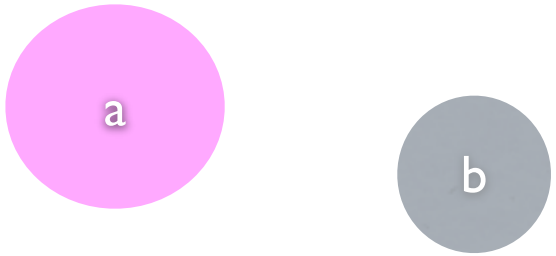
Spatial Logic

Question

What representation is needed for spatial reasoning?

Spatial Knowledge Representation

Formal representation of the meaning.



Disconnected?

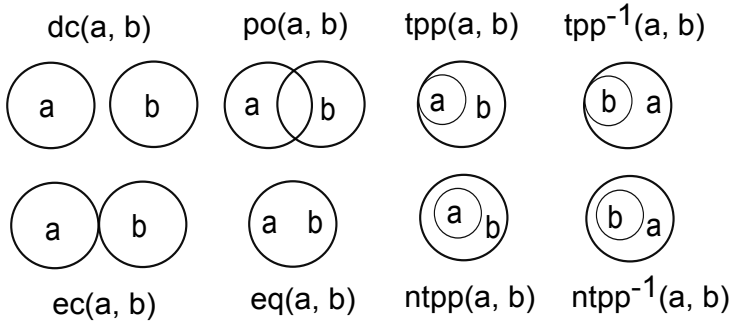
Touch?

Overlap?

Within?

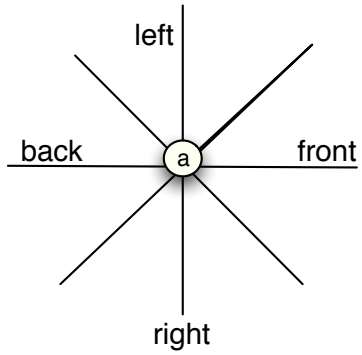
Qualitative Representation and Reasoning

- Topology (Region Connection Calculus)
- Orientation/Directions
- Distances, Sizes and Shapes



The RCC-8 relations.

Answering GIS queries: Retrieve all toxic waste dumps which are within 10 miles of an elementary school and located in Penobscot County and its adjacent counties.



[Cohn, Anthony & Hazarika, Shyamanta. (2001). Qualitative Spatial Representation and Reasoning: An Overview. Fundamenta Informaticae, 46. 1-29]

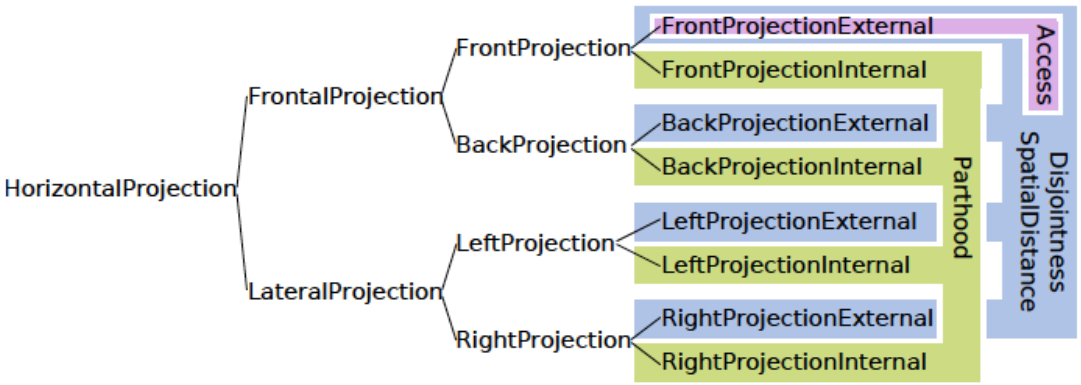
[Andrew U. Frank, Qualitative Spatial Reasoning: Cardinal Directions as an Example, Geographical Information Systems 10(3):269-290, 1996]

[Max J. Egenhofer and Robert D. Franzosa, Point-set topological spatial relations, International Journal of Geographical Information Systems, 1991]

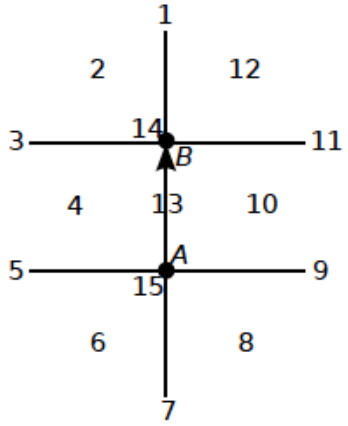
Connecting Linguistic Representation to Formal Calculi Representations

Connections between linguistic representations and logical theories of space

- Connecting linguistically motivated ontologies like GUM to a projective spatial relations formalism, double-cross calculi.



Projective horizontal relations in GUM



DCC's 15 qualitative orientation relations

[J. Hois and O. Kutz. Natural language meets spatial calculi. In C. Freksa, N. S. Newcombe, P. Gärdenfors, and S. Wölfl, editors, Spatial Cognition VI. Learning, Reasoning, and Talking about Space , volume 5248 of LNCS, Springer, 2008.]

Corpus Annotations

Corpus annotations based on both cognitive-linguistic and formal logic representations

Corpus Annotations

SpatialML:

Focused on geographical locations, annotating directional and topological relations.

[Inderjeet Mani, et, al. (2009) SpatialML: Annotation Scheme, Resources, and Evaluation, MITRE Corporation.]

Spatial Role Labeling (SpRL):

Based on holistic spatial semantics also trying to connect to multiple spatial calculi models

[Kordjamshidi, P., van Otterlo, M., Moens, M. F. (2010). Spatial role labeling: Task definition and annotation scheme. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).]

ISO-Space:

More comprehensive by considering dynamics of motion verbs and detailed properties of spatial entities.

[J. Pustejovsky and J. L. Moszkowicz. Integrating motion predicate classes with spatial and temporal annotations. In Donia Scott and Hans Uszkoreit, editors, COLING 2008: Companion volume D, Posters and Demonstrations , pages 95–98, 2008.]
[J. Pustejovsky and J. L. Moszkowicz. The role of model testing in standards development: The case of iso-space. In Proceedings of LREC'12 , pages 3060–3063. European Language Resources Association (ELRA), 2012.]
[Handbook of linguistic annotation, N Ide, J Pustejovsky, Springer, 2017.]

And,

Watch for this Book: Annotation-based Semantics for Space and Time in Language, Kiyong Lee,
Cambridge University Press

Corpus Annotations

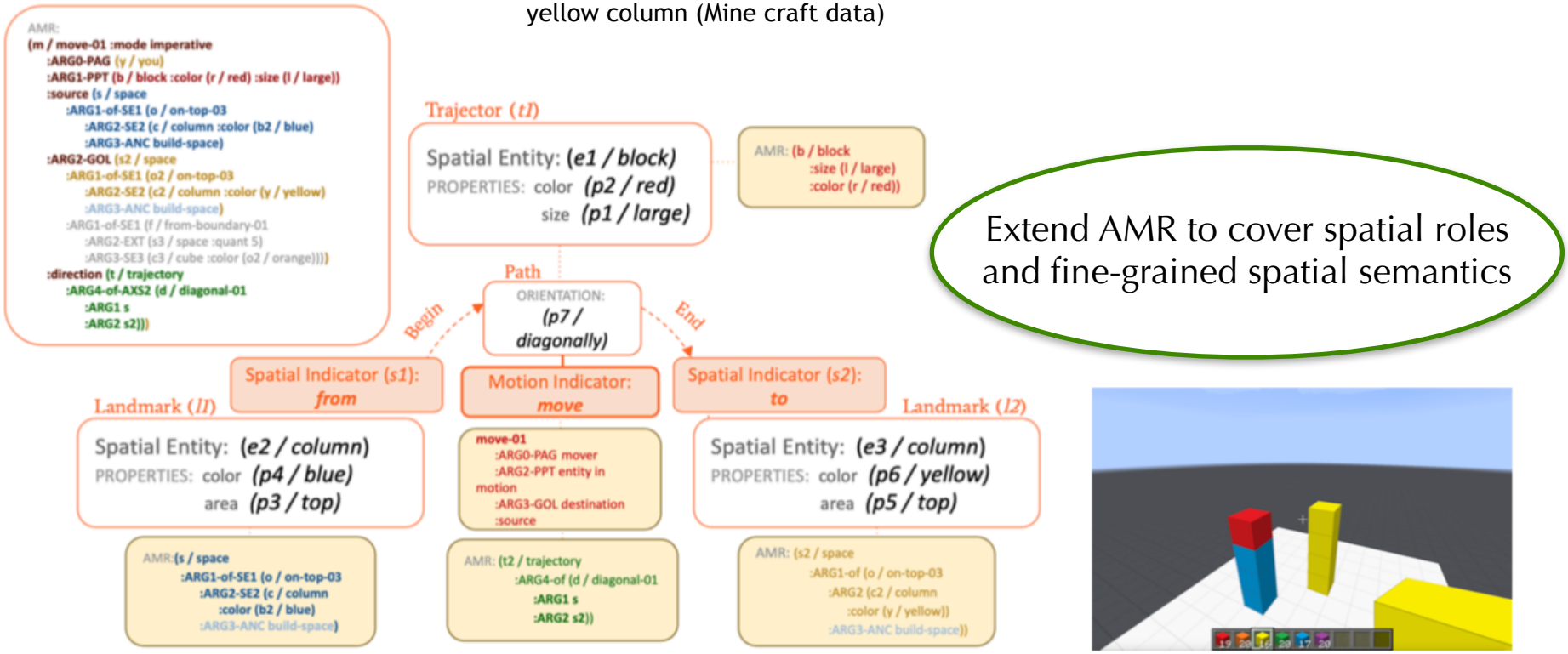
Annotations are applied on various datasets

- Degree Confluence Project (DCP)
- Cross Language Evaluation Forum (CLEF)
- Ride for Climate (RFC)
- Generalized Upper Model (GUM) Maptask corpus

AMR and Spatial Roles

Abstract Meaning Representation with Spatial Roles

- Move the large red block diagonally from the top of the blue column to the top of the yellow column (Mine craft data)



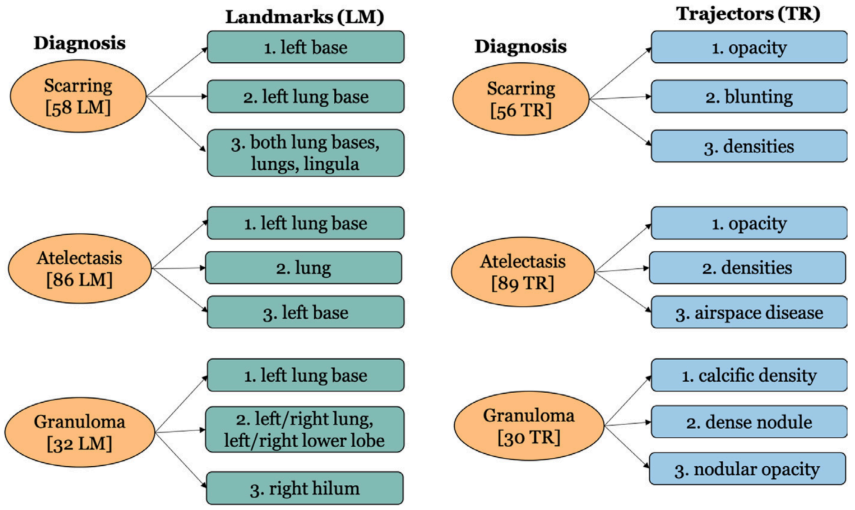
[Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, Dan Roth, Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, Dan Roth. LREC 2020.]
[Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus, Julia Bonn, Martha Palmer, Jon Cai, Kristin Wright-Bettner, LREC-2020.]

Spatial Annotations in Medical Domain

Spatial roles under RadSpRLRelation

TRAJECTOR	Radiological entity (usually a radiographic finding whose position is described
LANDMARK	Anatomical location of a TRAJECTOR
DIAGNOSIS	Potential diagnosis associated with a spatial relation
HEDGE	Any uncertainty phrase used to describe a finding or diagnosis

- 2000 chest X-ray reports from a pool of 3996 de-identified reports collected from the Indiana Network for Patient Care —released by the National Library of Medicine.
- Annotations further extended and connected to spatial configurations in Rad-SpatialNet resource.



[A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations, Surabhi Datta, Kirk Roberts, In J Biomed Inform, 2020]

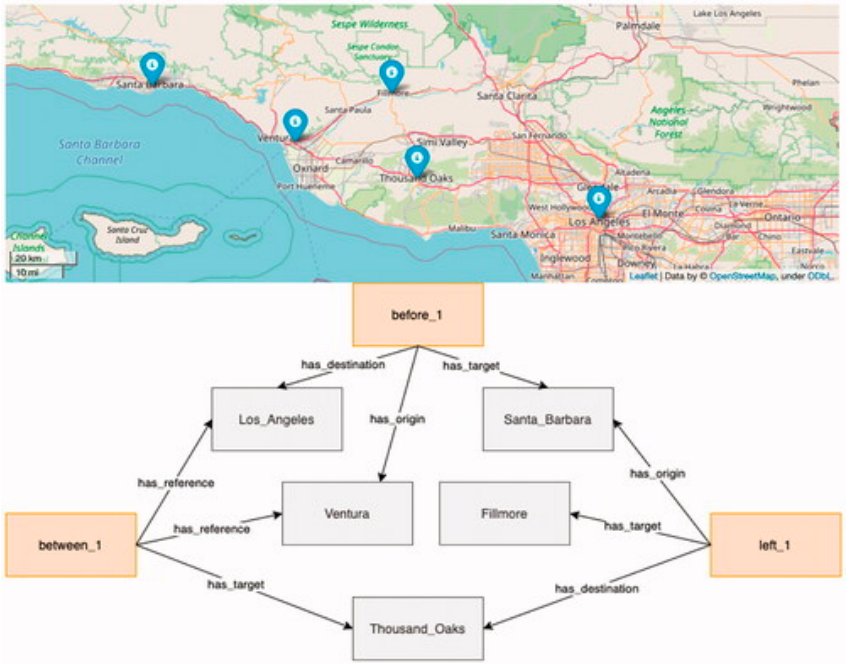
[Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports, Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F. Riascos-Castaneda, Kirk Roberts, LREC 2020.]

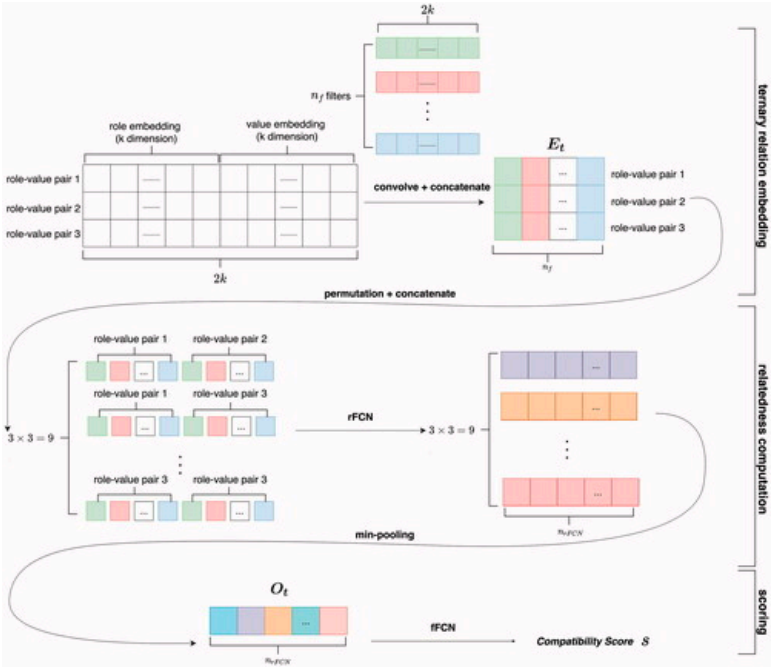
[SpatialNet: A Declarative Resource for Spatial Relations, Morgan Ulinski, Bob Coyne, Julia Hirschberg, SpLU-RoboNLP-2019]

[Bob Coyne, Daniel Bauer, and Owen Rambow. 2011. VigNet: Grounding Language in Graphics using Frame Semantics. In Proceedings of the ACL 2011 Workshop on Relational Models of Semantics, pages 28–36, Portland, Oregon, USA.]

Spatial KG & Distributed Representations

- Forming KG of geographical places to learning representation of the places and their spatial relations





Reasoning over higher-order qualitative spatial relations via spatially explicit neural networks, Rui Zhu, Krzysztof Janowicz, Ling Cai & Gengchen Mai, International Journal of Geographical Information Science, July 2022.

Information Extraction

Spatial Information Extraction

Spatial IE can be seen as:

A formal symbolic spatial meaning representation that can be used in down stream tasks.

Spatial Semantics Shared Tasks

SemEval series of workshops: 2012, 2013 and 2015

- Spatial Role Labeling (SpRL)
- ISO-space

Multimodal SpRL (mSpRL) workshop: CLEF-2017

See more info here: https://www.cse.msu.edu/~kordjams/SpRL.htm

Information Extraction

Two Layers of Semantics:

Based on cognitive linguistic elements and multiple calculi.

1. **SpRL**: Spatial role labeling

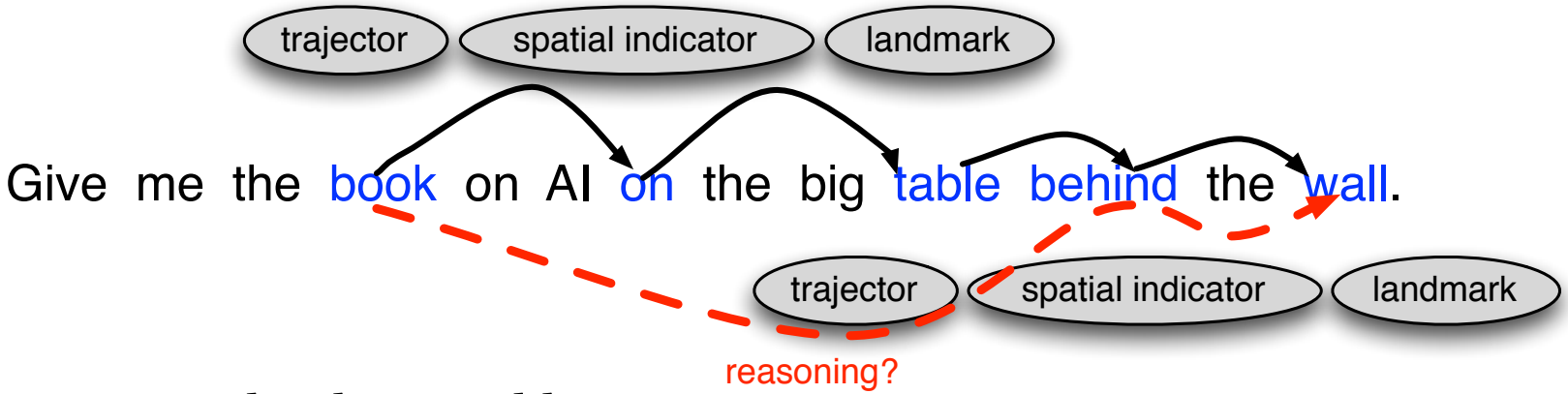
   -Identifying objects, roles and relations

2. **SpQL**: Spatial qualitative labeling

   -Identifying types of relations based on spatial calculi models

[Kordjamshidi et.al, 2012, Learning to interpret spatial natural language in terms of qualitative spatial relations Series Explorations in Language and Space.]

Spatial Role Labeling (SpRL)



$<on_{SP}\ book_{TR}\ table_{LM}>$ $<behind\ book\ wall>$

$<behind_{SP}\ table_{TR}\ wall_{LM}>$

Come over here! Implicit roles?

$<over_{SP}\ undefined_{TR}\ here_{LM}>$

[P Kordjamshidi, M Van Otterlo, MF Moens, Spatial role labeling: Towards extraction of spatial relations from natural language ACM-Transactions in speech and language processing, 2011]

[P Kordjamshidi, P Frasconi, M Van Otterlo, MF Moens, L De Raedt, Relational learning for spatial relation extraction from natural language; International Conference on Inductive Logic Programming, ILP proceedings, LNCS, 2012]

Spatial Qualitative Labeling (SpQL)

Based on multiple calculi models

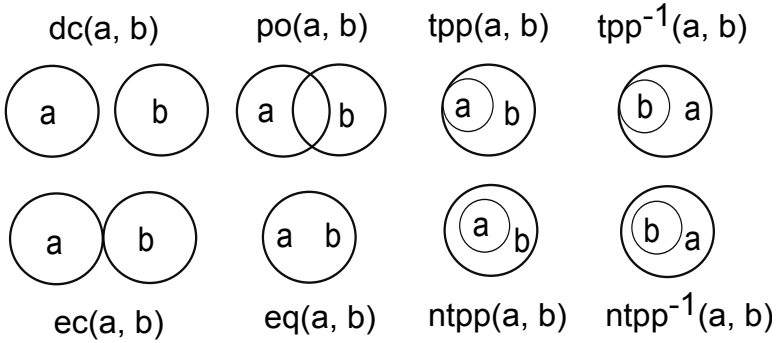
- **Topological**

{EQ, DC, EC, PO, PP}

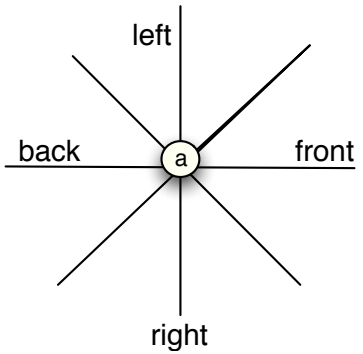
- **Directional**

{Right, Left, Above, Below, Front, Back}

- **Distal**



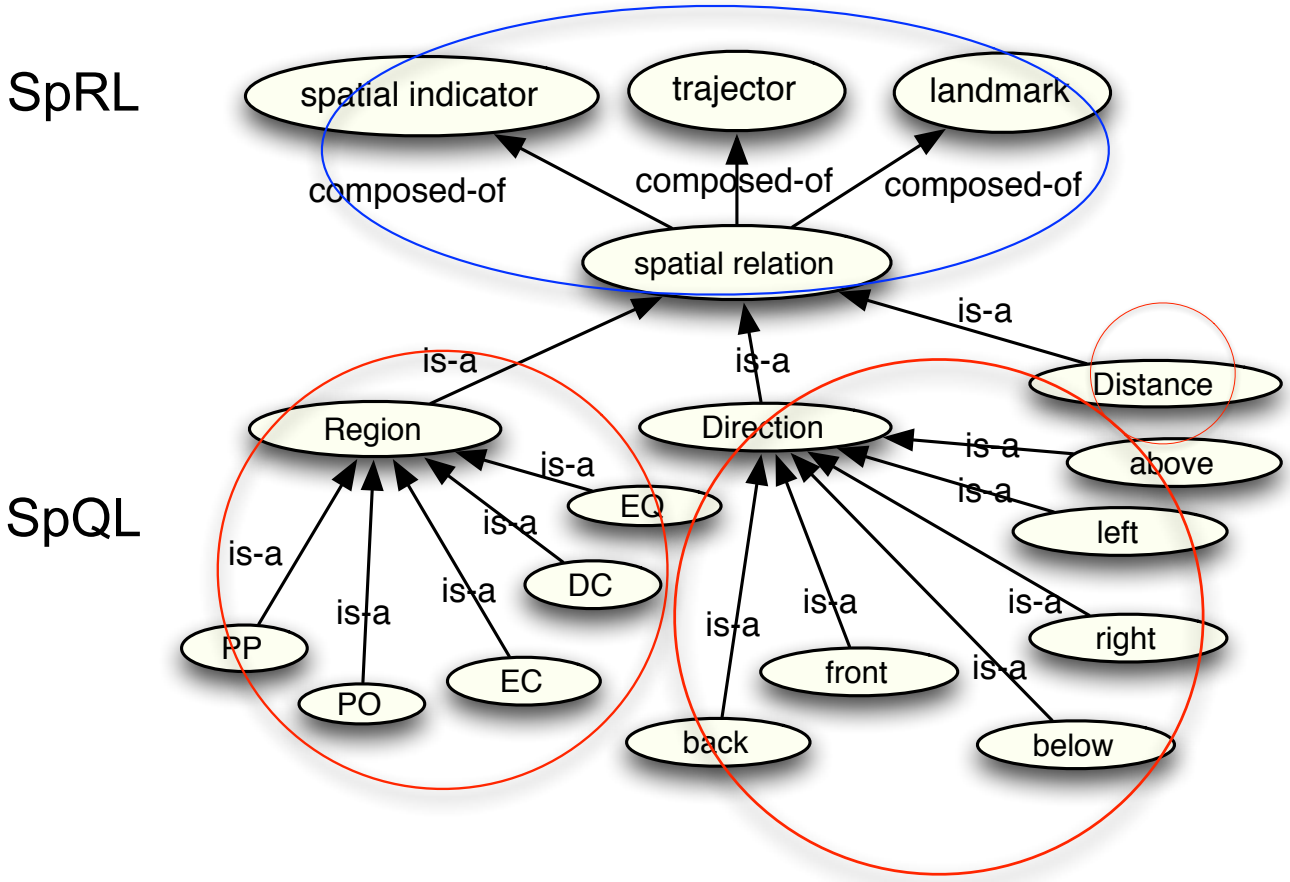
The RCC-8 relations.



[Kordjamshidi, P., van Otterlo, M., Moens, M.F.. From language towards formal spatial calculi. Computational Models of Spatial Language Interpretation Workshop (COSLI-2010) at COSIT.]

Spatial Ontology

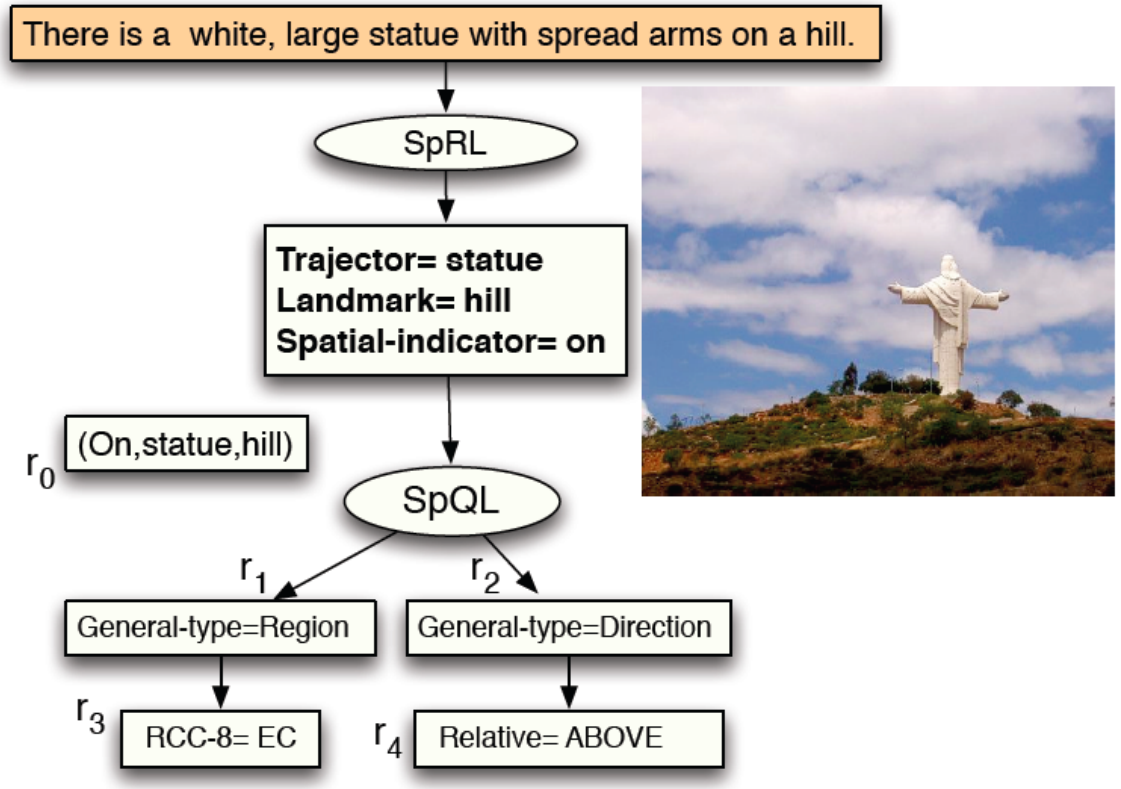
Based on cognitive linguistic elements and multiple calculi.



[Kordjamshidi, P., van Otterlo, M., Moens, M. F., Spatial role labeling: Task definition and annotation scheme. LREC-2010).]

[Kordjamshidi, P., Hois, J., van Otterlo, M., Moens, M. F., Learning to interpret spatial natural language in terms of qualitative spatial relations. Series Explorations in Language and Space. 2011.]

SpRL data



SemEval-2012/2013/2015 and CLEF/mSpRL-2017 benchmarks.

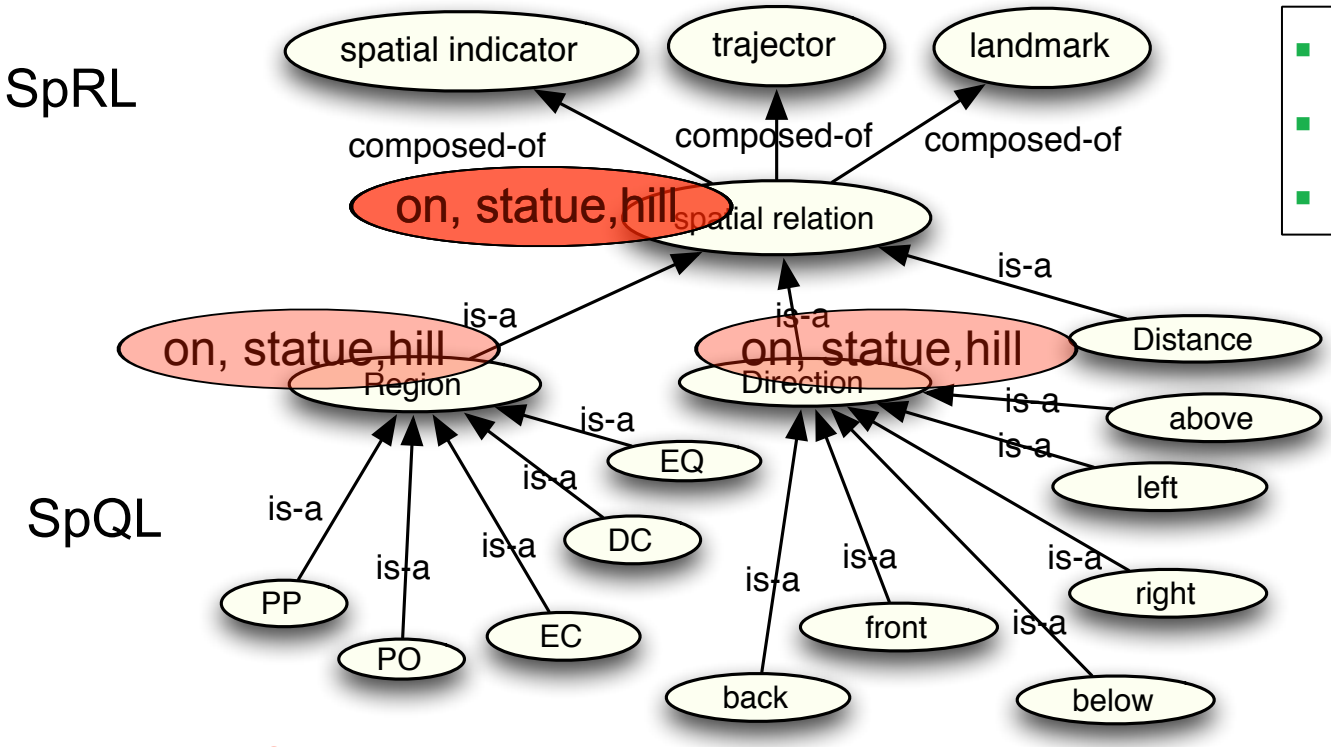
[Kordjamshidi et al. SemEval2012] [Kolomiyets, et.al. SemEval2013] [Pustejovsky et.al, SemEval2015]

[Kordjamshidi et.al. CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview.]

Exploit ontological information and structure

Semantic representation via Ontology population

There is a white, large **statue** with spread arms **on** a **hill**.



- Mutual exclusivity
- Is-a
- Composed-of

Structured (Deep) machine learning!

[Kordjamshidi, Moens. Global machine learning for spatial ontology population; Journal of Web Semantics, 2015]

Using ISO-space for Extraction

ISO Standard for Annotation of Spatial Information as expressed in Language

a. PLACES AND SPATIAL ENTITIES: natural or artificial locations in the world, as well as objects participating in spatial relations.
b. EVENTS AND MOTION EVENTS: Eventualities involving movement from one location to another.
c. SPATIAL SIGNALS AND SPATIAL MEASURES: linguistic markers that establish relations between places and spatial entities.
d. SPATIAL RELATIONSHIPS: The specific qualitative configurational, orientational, and metric relations between objects.

Spatial Relations in ISO space

Spatial Relations in ISO-Space

1. QSLINK – qualitative spatial links;
2. OLINK – orientation information;
3. MOVELINK – movement links;

DC	*the [grill] outside of the [house]*
EC	*the [cup] on the [table]*
PO	*[Russia] and [Asia]*
EQ	*[boston] and the [capital] of Massachusetts*
TPP	*the [shore] of [Delaware]*
TPPi	
NTPP	*[Austin], [Texas]*
NTPPi	
IN	*the [bookcase] in the [room]*

a. [**Boston**$_{pl1}$] is [**north of**$_{s1}$] [**New York City**$_{pl2}$].
olink(ol1, figure=pl1, ground=pl2, trigger=s1, relType="NORTH",
frame_type=ABSOLUTE, referencePt=NORTH, projective=TRUE)
b. [**The dog**$_{sne1}$] is [**in front of**$_{s2}$] [**the couch**$_{sne2}$].
olink(ol2, figure=sne1, ground=sne2, trigger=s2, relType="FRONT",
frame_type=INTRINSIC, referencePt=sne2, projective=FALSE)
c. [**The dog**$_{sne3}$] is [**next to**$_{s3}$] [**the tree**$_{sne4}$].
olink(ol3, figure=sne3, ground=sne4, trigger=s3, relType="NEXT TO",
frame_type=RELATIVE, referencePt=VIEWER, projective=FALSE)

a. [**Boston**$_{pl1}$] is [**north of**$_{s1}$] [**New York City**$_{pl2}$].
olink(ol1, figure=pl1, ground=pl2, trigger=s1, relType="NORTH",
frame_type=ABSOLUTE, referencePt=NORTH, projective=TRUE)
b. [**The dog**$_{sne1}$] is [**in front of**$_{s2}$] [**the couch**$_{sne2}$].
olink(ol2, figure=sne1, ground=sne2, trigger=s2, relType="FRONT",
frame_type=INTRINSIC, referencePt=sne2, projective=FALSE)
c. [**The dog**$_{sne3}$] is [**next to**$_{s3}$] [**the tree**$_{sne4}$].
olink(ol3, figure=sne3, ground=sne4, trigger=s3, relType="NEXT TO",
frame_type=RELATIVE, referencePt=VIEWER, projective=FALSE)

SpaceEval 2015 Tasks

Enriches SpRL (SemEval 2012)

- **SE**: Spatial Element Identification.
- **SS**: Spatial Signal Identification.
- **MS**: Motion Signal Identification.
- **MoveLink**: Motion Relation Identification.
- **QSLink**: Spatial Configuration Identification.
- **OLink**: Spatial Orientation Identification.

[James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, Zachary Yocum. SemEval-2015 Task 8: SpaceEval; SemEval2015 workshop.]

Papers Overviewing Spatial Shared Tasks

SemEval-2012 task 3: Spatial role labeling. Kordjamshidi, P., Bethard, S., Moens, M.F. (2012). *{*SEM 2012}: The First Joint Conference on Lexical and Computational Semantics -- Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation {(SemEval 2012)}: Vol. 2. SemEval-2012. Montreal- Canada, 7-8 June (pp. 365-373) ACL, 2012.*

SemEval-2013 task 3: Spatial role labeling. Kolomiyets, O., Kordjamshidi, P.,Bethard, S., Moens, M.F. (2013). *Second joint conference on lexical and computational semantics (*SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013). Second joint conference on lexical and computational semantics. Atlanta, USA, 14-15 June 2013 (pp. 255-266). East Stroudsburg, PA: ACL, 2013*

SemEval-2015 Task 8: SpaceEval. Pustejovsky J., Kordjamshidi P., Moens M.F., Levine A., Dworman S., Yocum Z., (2015). SemEval2015.

CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview. P. Kordjamshidi, T. Rahgooy, M-F. Moens, J. Pustejovsky, U. Manzoor and K. Roberts. *LNCS volume 10456 on Experimental IR Meets Multilinguality, Multimodality, and Interaction; Proceedings of 8th International Conference of the CLEF Association*

Combining Vision and Language for Spatial Inf. Extraction

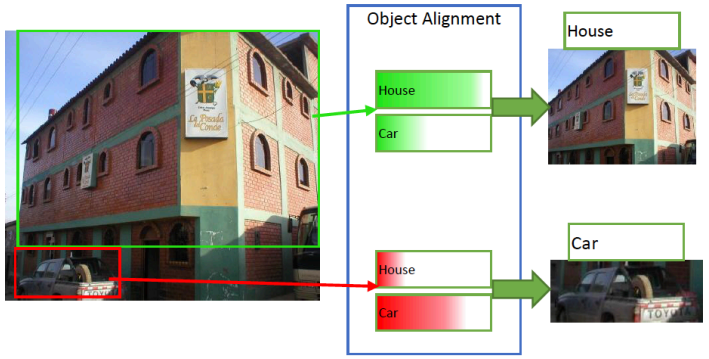
Nesting spatial constructs are a major sources of error in spatial relation extraction. *A car in front of the house on the left* can be interpreted as:
*(**A car** in front of the house) **on the left**.*

VS

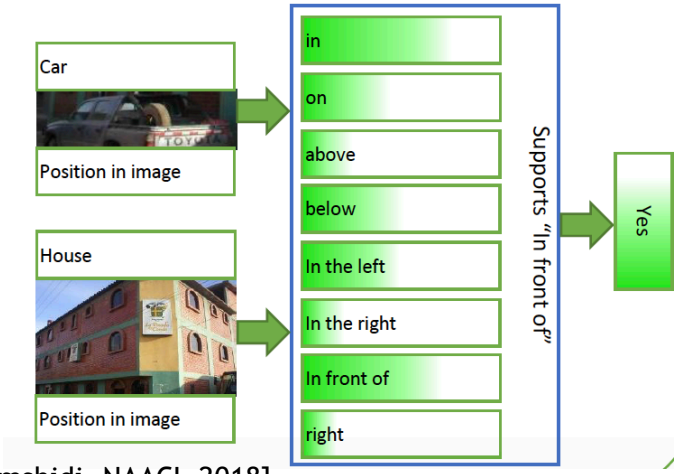
*A car in front of (**the house on the left**.)*

Image helps!

2% boost when using visual constraints!



- Align role candidates with image objects.
- For each candidate triplet check if the image supports the relation.
- Decide jointly based on image and text classifiers.



[D. Schlangen, S. Zarrieß, and C. Kennington. Resolving references to objects in photographs using the words-as-classifiers model, ACL-2016.]

[Visually Guided Spatial Relation Extraction from Text, Rahgooy, Manzoor, Kordjamshidi, NAACL-2018]

External Resources: Coreference Resolution for SpRL

"A narrow, rising street with colourful houses on both sides, among **them** a green house with balconies and a white car parked in front of it, and a blue-and-white church on the right."

Relations with pronoun landmark:

R_1 : a green house$_{tr}$, among$_{sp}$, them$_{lm}$ => "Them" is referring to "colorful houses".

R_2 : a white car$_{tr}$, in front of$_{sp}$, it$_{lm}$ => "it" is referring to "a green house".

Visual Genome Data gave another 2% boost!
Visual information can be see as a source of common sense.

[Manzoor, Kordjamshidi, Anaphora Resolution for Improving Spatial Relation Extraction from Text; NAACL, 2018, SpLU workshop]

Results

	Trajector			Landmark			Spatial indicator			Spatial triplet		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
BM	56.72	69.57	62.49	72.97	86.21	79.05	94.76	97.74	96.22	75.18	45.47	56.67
BM+C	65.56	69.91	67.66	77.74	87.78	82.46	94.83	96.86	95.83	75.21	48.46	58.94
BM+E	55.87	77.35	64.88	71.47	89.18	79.35	94.76	97.74	96.22	66.50	57.30	61.56
BM+E+C	64.40	76.77	70.04	76.99	89.35	82.71	94.85	97.48	96.15	68.34	57.93	**62.71**
BM+E+I	56.53	79.29	66.00	71.78	87.44	78.84	94.76	97.74	**96.22**	64.12	57.08	60.39
BM+E+I+C	64.49	77.92	**70.57**	77.66	89.18	**83.02**	94.87	97.61	96.22	66.46	57.61	61.72
SemEval-2012	78.2	64.6	70.7	89.4	68.0	77.2	94.0	73.2	82.3	61.0	54.0	57.3
SOP2015-10f	-	-	-	-	-	-	90.5	84	86.9	67.3	57.3	61.7

BM: Baseline, BM: Baseline, C: Constraints, E: Text Embeddings, I: Image embeddings

[Kordjamshidi, et.al., EMNLP 2017, Structured NLP workshop]

[Rahgooy, Manzoor, Kordjamshidi, NAACL-2018]

[Manzoor, Kordjamshidi, NAACL-2018, SpLU workshop]

Spatial Comprehension by Language Models

Spatial Question Answering

Do we have relevant corpora to evaluate spatial meaning representations and their impact on downstream tasks?

- SQuAD, Hotpot QA, WiQA
- bAbi (task 17 on spatial reasoning), BoolQA

Checking samples of these datasets, we realized:

- No complex spatial descriptions included
- Spatial reasoning is not a key issue for solving these tasks

Spatial Question Answering

A new Benchmark: SpartQA

Formal Representations:

- Topological relations. (contains, part-of, overlap,…)
- Relative directions. (Left, Right, under, above)
- Qualitative distance. (near to, close to, far from)

The girl is on the left of the bookcase. She holds a box with a cat in it.
What is to the right of the cat? The girl or the bookcase?

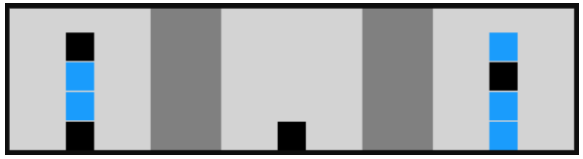
Rules of Reasoning:

- Symmetry : near to (girl, cat) -> near to (cat, girl)
- Transitivity : left (girl, bookcase) & left (cat, girl) -> left (cat, bookcase)
- Reverse : left (girl, bookcase) -> right (bookcase, girl)

Roshanak Mirzaee, et. al,. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. *2021 Conference of the North American Chapter of the Association for Computational Linguistics(NAACL): Human Language Technologies*, pages 4582–4598

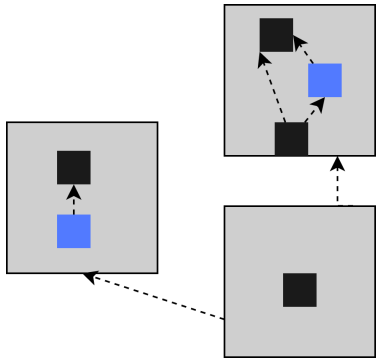
Spatial Reasoning QA dataset

Generate Dataset (SPARTQA) use Visual info and Rules of reasoning as a distant source of supervision



NLVR1 image

Random Sampling




```
[[{"y_loc":80,"size":20,"type":"square","x_loc":40,"color":"Black"},
  {"y_loc":59,"size":20,"type":"square","x_loc":40,"color":"#0099ff"},
  {"y_loc":38,"size":20,"type":"square","x_loc":40,"color":"#0099ff"},
  {"y_loc":17,"size":20,"type":"square","x_loc":40,"color":"Black"}],
```


NLVR1 scene graph (image data)

Story

We have three blocks, A, B and C. Block B is to the right of block C and it is below block A. Block A has two black medium squares. Medium black square number one is below medium black square number two and a medium blue square. It is touching the bottom edge of this block. The medium blue square is below medium black square number two. Block B contains one medium black square. Block C contains one medium blue square and one medium black square. The medium blue square is below the medium black square.

Questions

YN: Is there a square that is below medium square number two above all medium black squares that are touching the bottom edge of a block? Yes

CO: Which object is above a medium black square? the medium black square which is in block C or medium black square number two? medium black square number two

FR: What is the relation between the medium black square which is in block C and the medium square that is below a medium black square that is touching the bottom edge of a block? Left

FB: Which block(s) has a medium thing that is below a black square? A, B, C

FB: Which block(s) doesn't have any blue square that is to the left of a medium square? A, B

Improve Language Models for Spatial Reasoning

Evaluating BERT on spatial Understanding and Reasoning.

Fine-tune BERT on MLM task (using auto-SpartQA stories).

Fine-tune BERT on auto-SpartQA's training set.

#	Model	FB	FR	CO	YN	Avg
1	Majority	28.84	24.52	40.18	53.60	36.64
2	BERT	16.34	20	26.16	45.36	30.17
3	BERT (Stories only; MLM)	21.15	16.19	27.1	**51.54**	32.90
4	BERT (SpartQA-Auto; MLM)	19.23	29.54	**32.71**	47.42	34.88
5	BERT (SpartQA-Auto)	**62.5**	**46.66**	**32.71**	47.42	**47.25**
6	Human	91.66	95.23	91.66	90.69	92.31

Try more LMs and various test sets

#	Models	FB			FR			CO			YN		
		Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*
1	Majority	48.70	48.70	28.84	40.81	40.81	24.52	20.59	20.38	40.18	49.94	49.91	**53.60**
2	BERT	87.13	69.38	62.5	85.68	73.71	46.66	71.44	61.09	32.71	78.29	76.81	47.42
3	ALBERT	97.66	83.53	56.73	91.61	83.70	44.76	95.20	84.55	49.53	79.38	75.05	41.75
4	XLNet	**98.00**	**84.85**	**73.07**	**94.60**	**91.63**	**57.14**	**97.11**	**90.88**	**50.46**	**79.91**	**78.54**	39.69
5	Human	85		91.66	90		95.23	94.44		91.66	90		90.69

find relation (FR), find blocks (FB), choose object (CO), and yes/no/DK (YN)

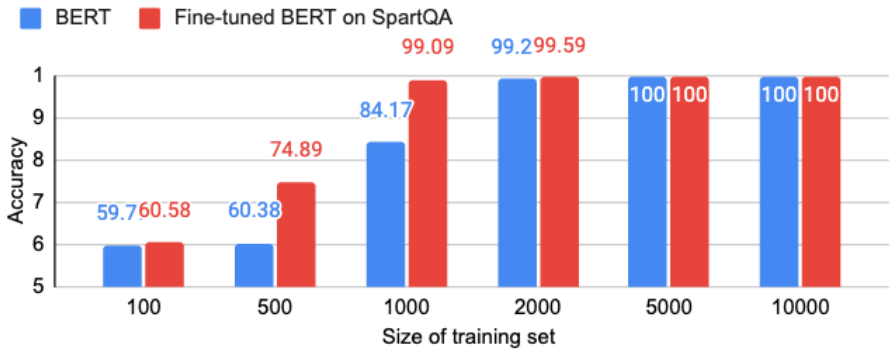
Fine-tuned LM with SpartQA

bAbI dataset (task 17)

The **pink rectangle** is to the *left* of the **red square.**

The **blue square** is to the *right* of the **red square.**

Is the **blue square** to the *left* of the **pink rectangle**? No
Is the **red square** to the *left* of the **blue square**? Yes



Model	Accuracy
Majority baseline	62.2
Recurrent model (ReM)	62.2
ReM fine-tuned on SQuAD	69.8
BERT (our setup)	71.89
ReM fine-tuned on QNLI	71.4
ReM fine-tuned on NQ	72.8
BERT fine-tuned on auto-SPARTQA	**74.18**

boolQ dataset

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]

A paper with similar idea in AAAI-2022 created a dataset called **StepGame**. There is no citation of SpaRTQA of NAACL-2021. So there is a disconnect between the results and comparisons.

Limitations of SpRTQA for Transferability

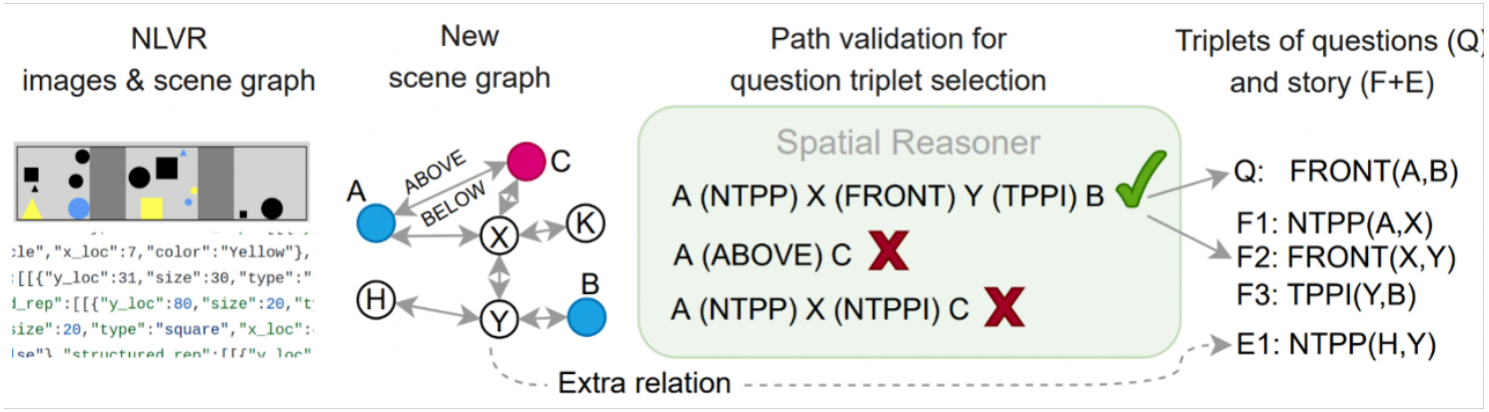
Problems

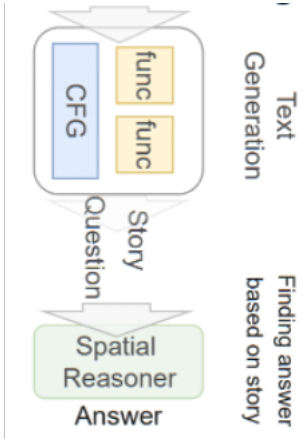
- Few spatial types
  - Compared to real world
- Low generalizability
  - Limited vocabulary
- Complex question text with nesting relations
  - Hard to process the question

Solution

- Gather more relation types
- Extend the vocabulary
- Simplified the questions and kept the reasoning difficulty/multi-hop reasoning on the text side

SPURTAN extension for Transferability





SPURTAN & ReSQ

- On two tasks (annotation):
  - Spatial Question Answering
  - Spatial Role Labeling
    - Spatial Entity/indicator extraction
    - Spatial Relation Extraction

Dataset	Train	Dev	Test
SpaRTUN(YN)	20334	3152	3193
SpaRTUN(FR)	18400	2818	2830

R. Mirzaee, P. Kordjamshidi, 2022. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. The 2022 Conference on Empirical Methods in Natural Language Processing

Three boxes called one, two and three exist in an image. Box one contains a big yellow melon and a small orange watermelon. **Box two has a small yellow apple**. A small orange apple is inside and touching this box. Box one is in box three. **Box two** is to the **south** of, **far from** and to the **west** of **box three**. A **small yellow watermelon** is **inside box three**.

Q: Is **the small yellow apple** to the **west** of the **small yellow watermelon**? Yes

Q: Where is **box two** relative to the **small orange watermelon**? Left, Below, Far

(a) SpaRTUN - A synthetic large dataset provided as source of supervision

A **grey car** is parking in **front** of a **grey house with brown window frames** and **plants on** the **balcony**.

Q: Are the **plants in front of** the **car**? No

Q: Are the **plants in** the **house**? Yes

(b) ReSQ - A human-generated dataset for probing the models on realistic spatial problems

Experimental Results on QA

Model	DS	17^{900}	19^{500}
MB	-	51.9	10.6
BERT	-	87.39	34.53
BERT	SPARTQA-A	90.42	**100**
BERT	StepGame	87.39	99.89
BERT	SPARTUN-S	**92.43**	98.99
BERT	SPARTUN	90.02	99.89

Impact of using synthetic supervision on the bAbI tasks. All the models are further fine-tuned on the training set of task 17 (size = 1k) and 19 (size = 500), and test on bAbI test sets.

Model	DS	YN	FR
MB	-	53.60	24.52
BERT	-	**49.65**	18.18
BERT	SPARTQA-A	39.86	48.05
BERT	StepGame	44.05	11.68
BERT	SPARTUN-S	44.75	37.66
BERT	SPARTUN	48.25	**50.64**
Human	-	90.69	95.23

Impact of transfer learning on SPARTQA-HUMAN. SPARTQA-A stands for SPARTQA-AUTO.

		k steps of reasoning									
Model	DS	1	2	3	4	5	6	7	8	9	10
TP-MANN	-	85.77	60.31	50.18	37.45	31.25	28.53	26.45	23.67	22.52	21.46
BERT	-	98.44	94.77	91.78	71.7	57.56	50.34	45.17	39.69	35.41	33.62
BERT	SPARTQA-A	98.63	94.95	91.94	77.74	68.37	61.67	57.95	50.82	46.86	44.03
BERT	SPARTUN-S	**98.70**	**95.21**	**92.46**	77.93	69.53	62.14	57.37	48.79	44.67	42.72
BERT	SPARTUN	98.55	95.02	92.04	**79.1**	**70.34**	**63.39**	**58.74**	**52.09**	**48.36**	**45.68**

Result of models with and without extra supervision on StepGame.

Experimental Results SpRL

- Spatial Roles and Relations

Model	DS	mSpRL	SpartQA-H
R-Inf	-	80.92	-
SAE	-	**88.59**	55.8
SAE	SpartQA-A	88.41	57.28
SAE	SpaRTUN	88.03	**72.43**

Evaluate spatial argument extraction(SAE) on two mSpRL and SpartQA-Human(SpartQA-H) datasets with and without synthetic supervision.

Model	DS	mSpRL	SpartQA-H
R-Inf	-	68.78	-
SRE	-	69.12	S: 48.58 Q: 49.46
SRE	SpartQA-A	68.38	S: 58.32 Q: 55.17
SRE	SpaRTUN	**74.74**	S: **61:53** Q: **63.22**

Spatial relation extraction on mSpRL and SpartQA-Human(SpartQA-H) with and without synthetic supervision. Since the questions(Q) and stories(S) in SpartQA-Human have different annotations (questions have empty spatial_indicators), we separately train and test this model on each.

ReSQ (Real World Spatial Questions)

A **grey car** is parking in **front** of a **grey house** **with brown window frames** and **plants on** the **balcony**.

Q: Are the **plants in front of** the **car**? No

Q: Are the **plants in** the **house**? Yes

Model	DS	Accu
MB	-	50.21
BERT	-	57.37
BERT	SPARTQA-AUTO	55.08
BERT	StepGame	60.14
BERT	SPARTUN-S	58.03
BERT	SPARTUN	**63.60**
Human	-	90.38

Results of models with and without extra supervision on the ReSQ. The Human accuracy is the performance of human on answering a part of test set.

Finding

- Synthetic data with limited vocabulary still can help transfer learning for both information extraction and question answering for realistic data domain.

Table of Content

Downstream Tasks

Some tasks that involve language and vision modalities and grounding language in physical world.

- Natural Language Visual Reasoning (NLVR)
- (Visual) Question Answering (VQA)
- Navigation and Instruction Following
- Text to scene and Scene to text
- Situated Dialogue and Interactive systems

Following Navigation Instructions

- Spoken dialogue describing a path through a map
- No linguistic annotations
- No alignment between text and route
- Using reinforcement learning
- State space combines linguistic features and the current location in the map, the reward is computed using the reference path

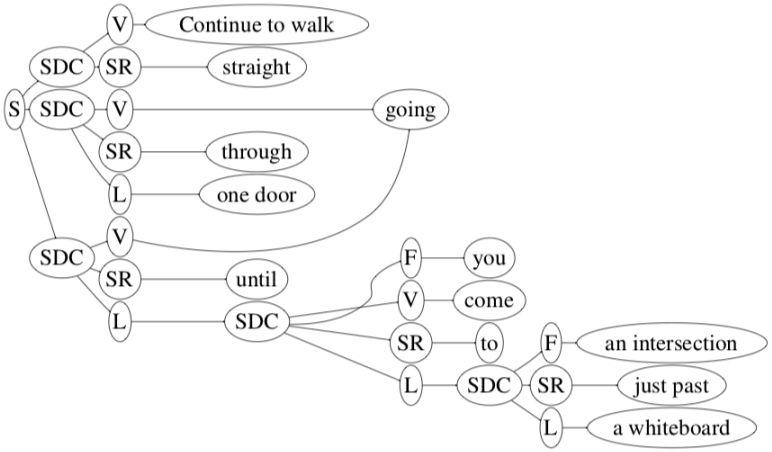
1. go vertically down until you're underneath eh diamond mine
2. then eh go right until you're
3. you're between springbok and highest viewpoint

HCRC Map task corpus

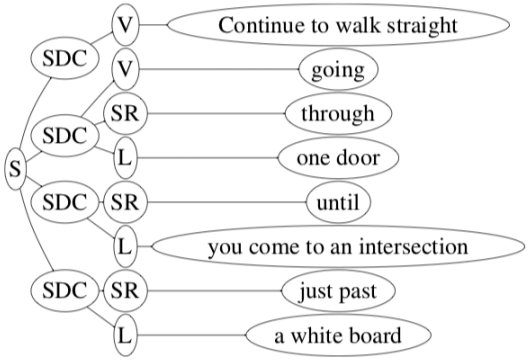
[Learning to Follow Navigational Directions , Adam Vogel and Dan Jurafsky, ACL-2010.]

Following Navigation Instructions

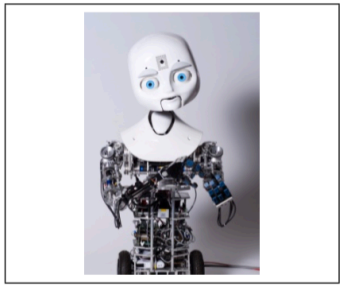
- Spatial language is represented as a hierarchy of spatial description clauses (SDC).
- SDC are hand annotated for a set of instructions.
- A discriminative probabilistic graphical models finds the most probable path by extraction of the SDCs and using the detected visual landmarks.



(a) Ground Truth



(b) Automatic



(a) humanoid

(b) helicopter

With your back to the windows, walk straight through the door near the elevators. Continue to walk straight,…

[Toward Understanding Natural Language Directions, 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Thomas Kollar, Stefanie Tellex, Deb Roy, Nicholas Roy, 2010]

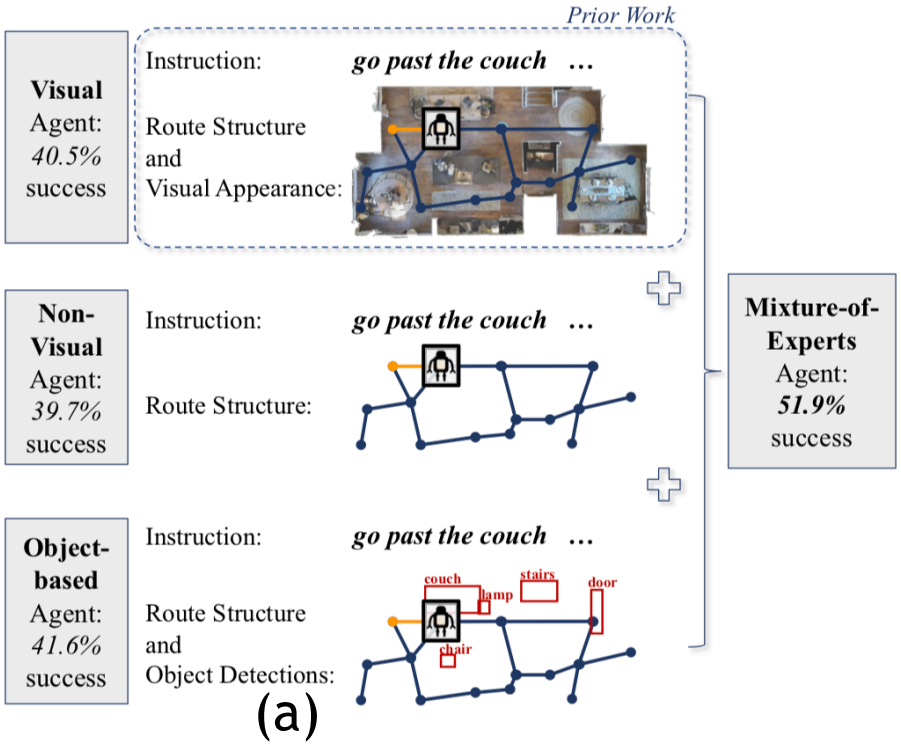
Spatial Semantics in Navigation

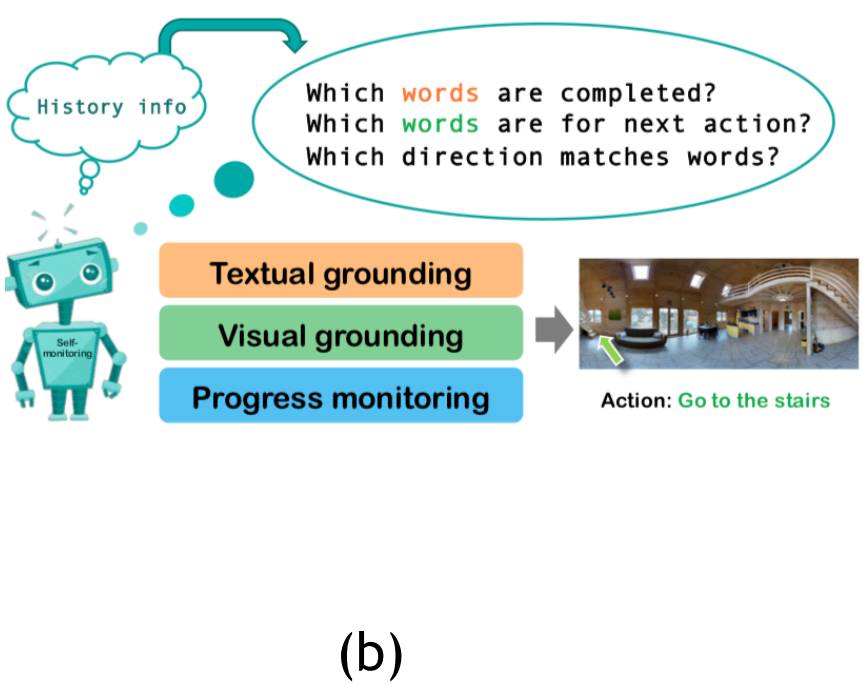
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Room2Room dataset

Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3674-3683.

Following Navigation Instructions





a) Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation, Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, Kate Saenko. ACL-2019.

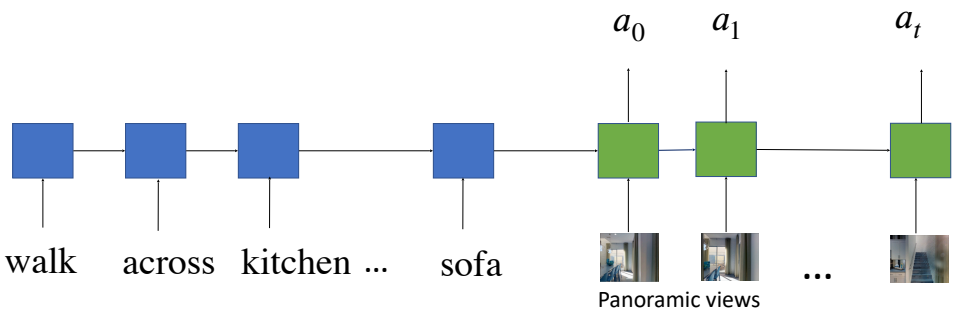
b) Self-Monitoring Navigation Agent via Auxiliary Progress Estimation, Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, Caiming Xiong. ICLR-2019.

c) Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout, Hao Tan, Licheng Yu, Mohit Bansal. NAACL-2019.

Current Neural Models

LSTM-based Model

- Self-Monitor[1]
- Environment Dropout [2]
- Speaker-Follower [3]

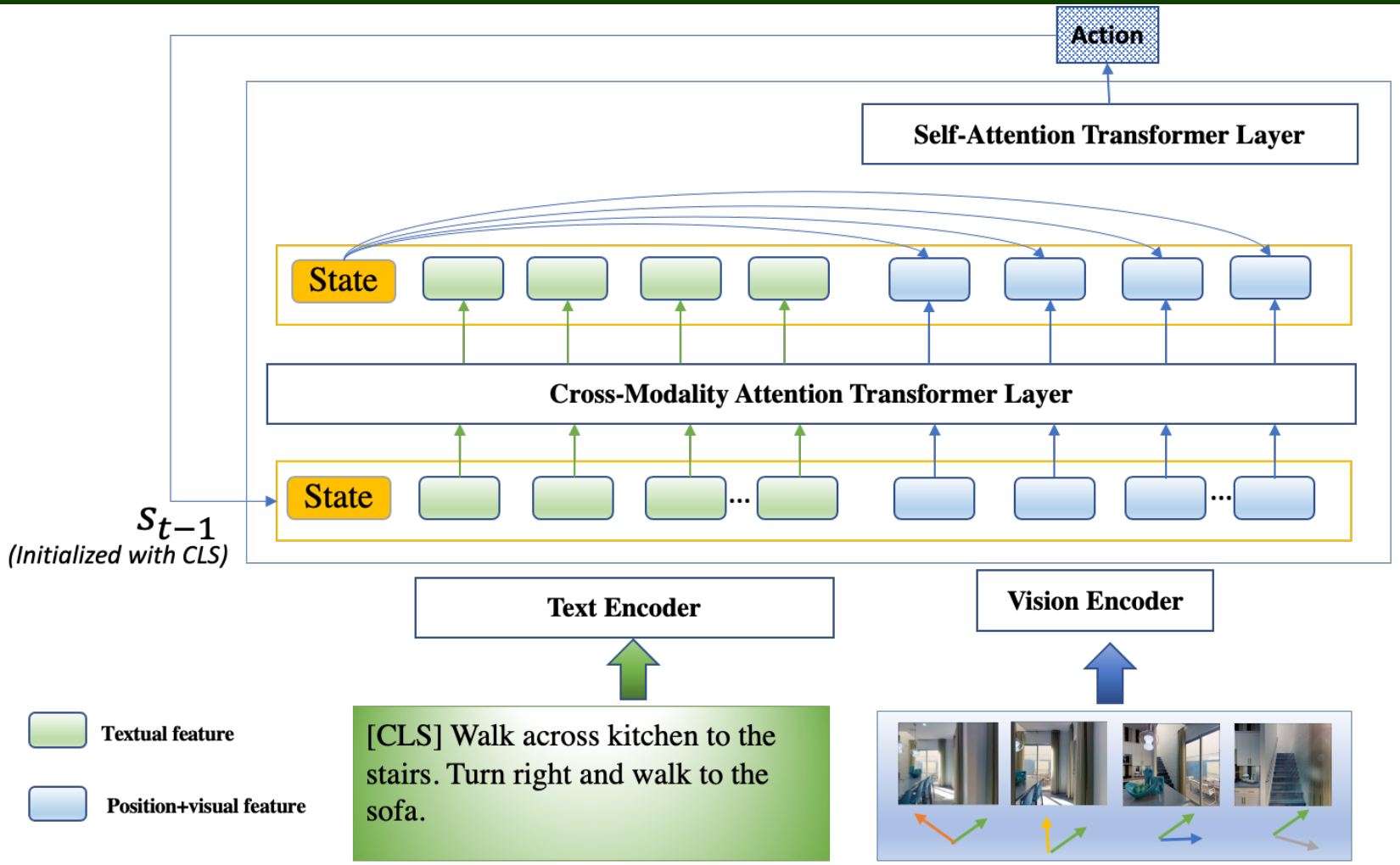


Transformer-based Model

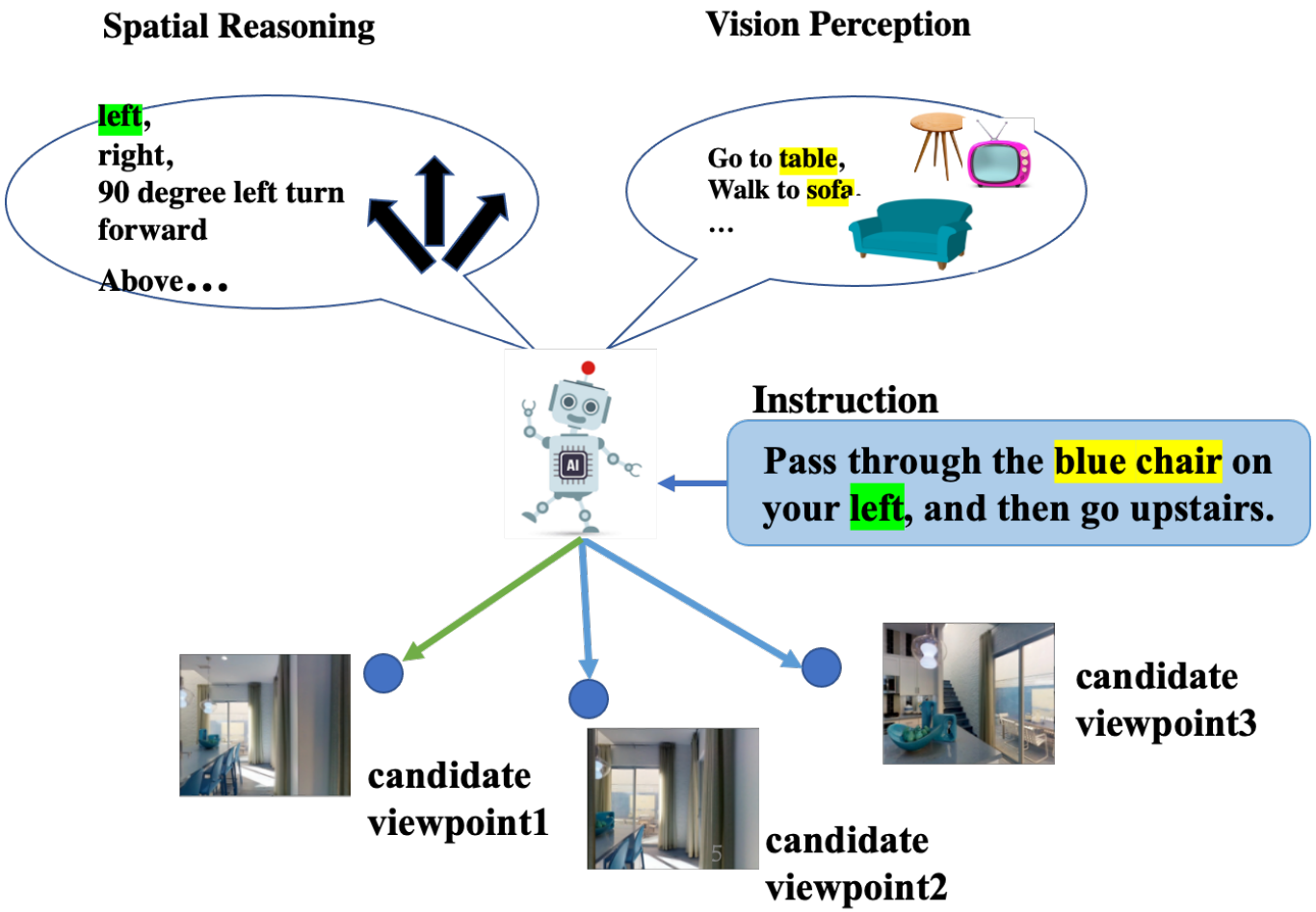
- PREVALENT [4]
- Recurrent VLN [5]
- HAMT[6]

[1] Ma C Y, Lu J, Wu Z, et al. Self-monitoring navigation agent via auxiliary progress estimation[J]. arXiv preprint arXiv:1901.03035, 2019.
[2] Tan H, Yu L, Bansal M. Learning to navigate unseen environments: Back translation with environmental dropout[J]. arXiv preprint arXiv:1904.04195, 2019.
[3] Fried D, Hu R, Cirik V, et al. Speaker-follower models for vision-and-language navigation[J]. Advances in Neural Information Processing Systems, 2018, 31.
[4]Hao W, Li C, Li X, et al. Towards learning a generic agent for vision-and-language navigation via pre-training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13137-13146.
[5] Hong Y, Wu Q, Qi Y, et al. VLN bert: A recurrent vision-and-language bert for navigation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1643-1653.
[6] Chen S, Guhur P L, Schmid C, et al. History aware multimodal transformer for vision-and-language navigation[J]. Advances in Neural Information Processing Systems, 2021, 34.

Transformer-based Baseline: Recurrent VLN



Reasoning ability of VLN



Issues

- **Spatial reasoning on orientation**

e.g. *"90 degree left turn", "on your right", etc.*

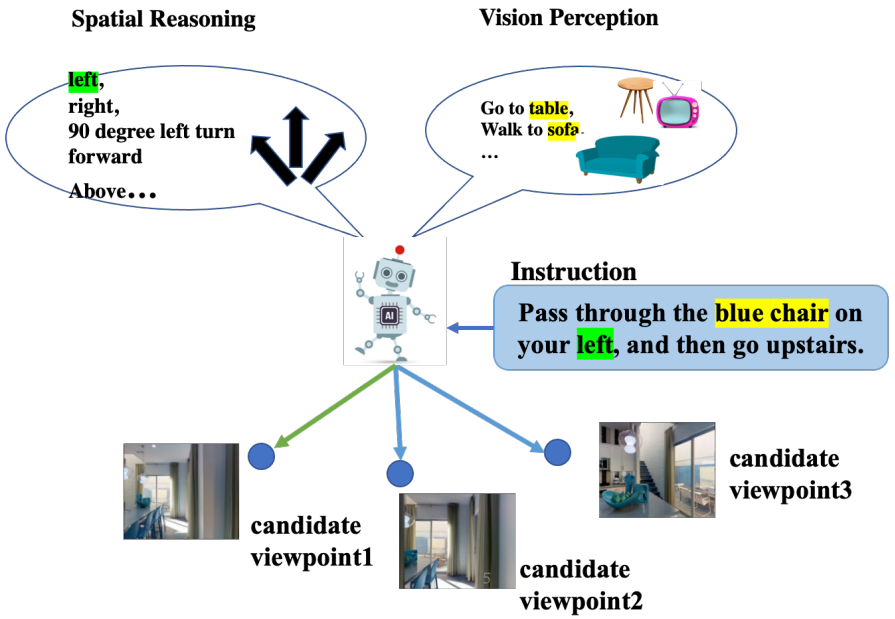
- **Spatial reasoning on visual perception**

e.g. *"walk to the sofa", "pass the table", etc.*

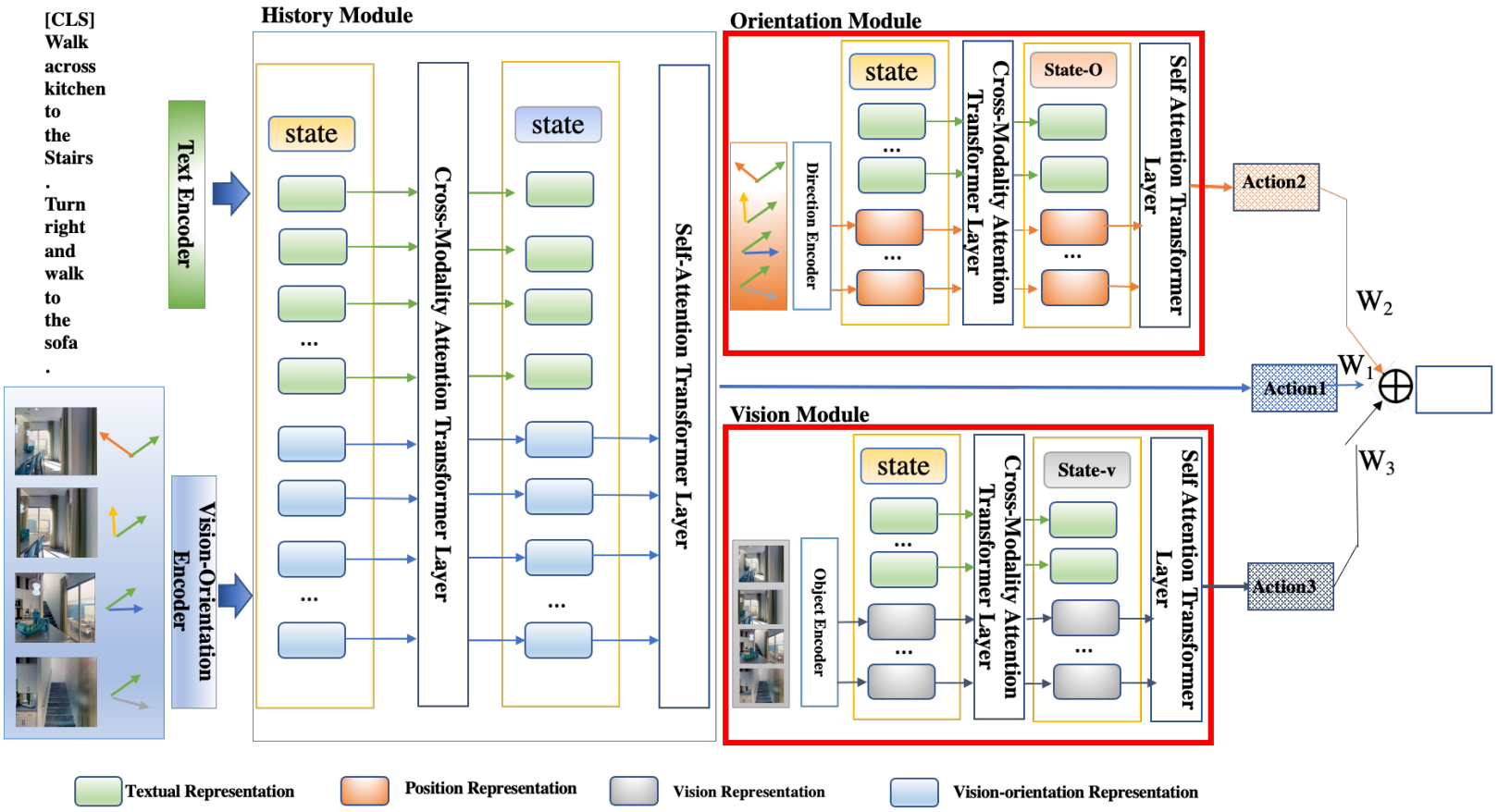
- The entangling of the two abilities maybe **not effective** to the **reasoning ability** and **interpretability** of the model. (Baseline Issue)

LovIS model

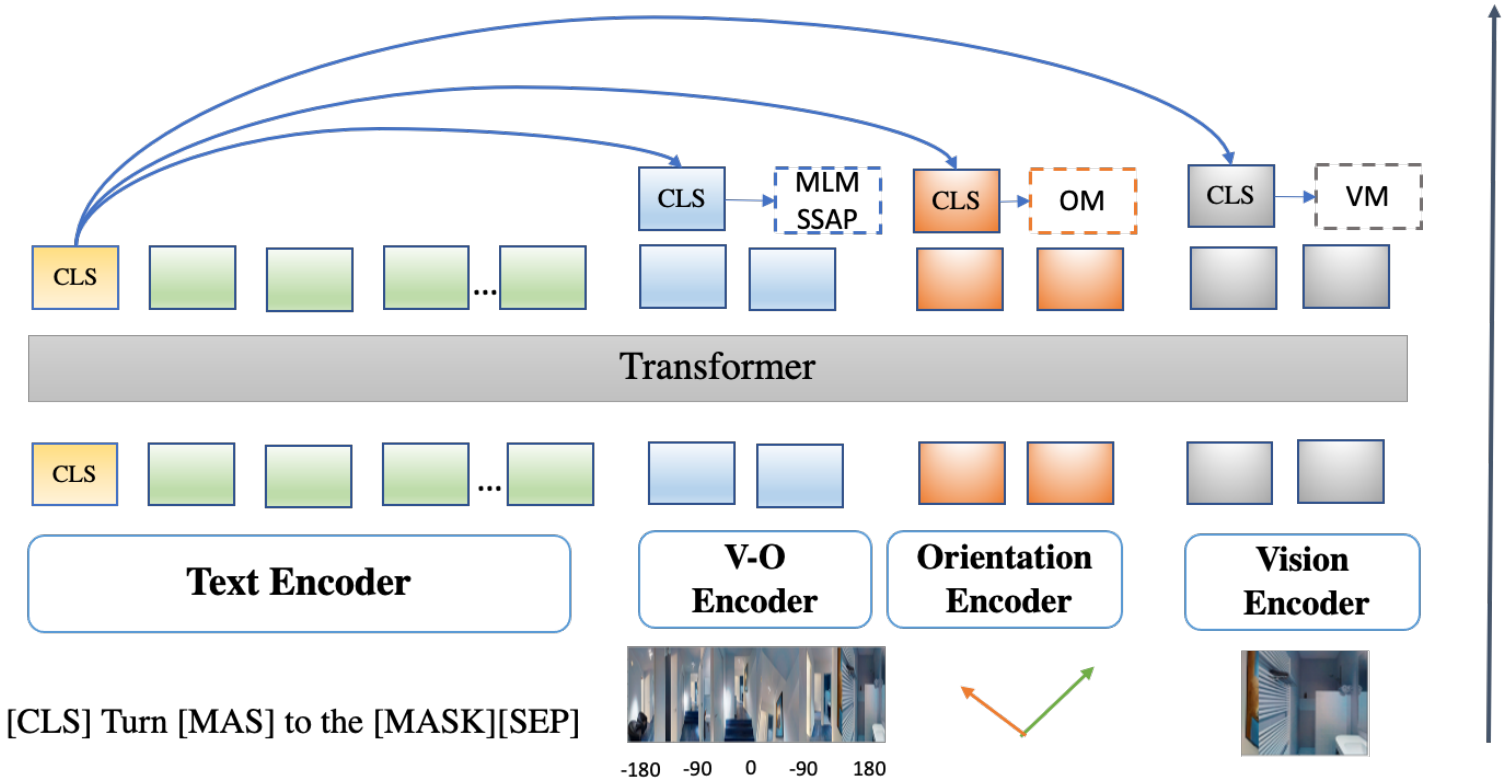
- **Transformer-based navigation agent**
- **Modular design**
  - ❑ Orientation Module
  - ❑ Vision Module
- **Novel pre-training tasks for different modules**
  - ❑ Orientation Matching-> Orientation Module
  - ❑ Vision Matching -> Vision Module



LovIS model: New Modules



LovIS model: New Pre-training tasks



History Module -- > *MLM* (**Masked Language Modeling**); *SSAP* (**Single Step Action Prediction**)
Orientation Module -- > *OM*(**Orientation Matching**)
Vision Module -- > *VM* (**Vision Matching**)

LovIS model: New Pre-training tasks

Masked Language Modeling (MLM): *Predict the landmark and orientation tokens based on CLS representation.*

Single Step Action Prediction (SSAP): *Select an action from discrete actions based on the attention score between CLS representation and vision-orientation representations.*

Orientation Matching (OM): *Predict 4-bits of orientation representation based on the text representation and the initial orientation representation.*

Vision Matching (VM): *Classify whether text and image are matching based on text representation and vision representation.*

LovIS model: Experimental Results

- R2R

	Method	Val seen			Val Unseen			Test(Unseen)		
		NE ↓	SR ↑	SPL↑	NE ↓	SR ↑	SPL↑	NE ↓	SR ↑	SPL ↑
1	Speaker-Follower (Fried et al., 2018)	3.36	0.66	-	6.62	0.35	-	6.62	0.35	0.28
2	Env-Drop (Tan et al., 2019)	3.99	0.62	0.59	5.22	0.47	0.43	5.23	0.51	0.47
3	OAAM (Qi et al., 2020)	-	0.65	0.62	-	0.54	0.50	5.30	0.53	0.50
4	RelGraph (Hong et al., 2020a)	3.47	0.67	0.65	4.73	0.57	0.53	4.75	0.55	0.52
5	NvEM (An et al., 2021)	3.44	0.69	0.65	4.27	0.60	0.55	4.37	0.58	0.54
6	PRESS (Li et al., 2019)	4.39	0.58	0.55	5.28	0.49	0.45	5.49	0.49	0.45
7	PREVALENT (Hao et al., 2020)	3.67	0.69	0.65	4.71	0.58	0.53	5.30	0.54	0.51
8	AirBERT (Guhur et al., 2021)	2.68	0.75	0.70	4.01	0.62	0.56	4.13	0.62	0.57
9	RecBERT (Hong et al., 2021)	2.90	0.72	0.68	3.93	0.63	0.57	4.09	**0.63**	0.57
10	HAMT (Chen et al., 2021)	-	0.69	0.65	-	0.64	0.58	-	-	-
11	RecBERT*	2.99	0.71	0.66	4.03	0.61	0.56	4.35	0.61	0.57
12	Our pretrain + RecBERT	2.90	0.74	0.69	3.75	0.63	0.58	4.20	**0.63**	0.57
13	Our pretrain + LOViS (our model)	**2.40**	**0.77**	**0.72**	**3.71**	**0.65**	**0.59**	**4.07**	**0.63**	**0.58**

- R4R

Method	Val Seen						Val Unseen					
	NE↑	SR↑	SPL↑	CLS↑	nDTW↑	sDTW↑	NE↓	SR↑	SPL↑	CLS↑	nDTW↑	sDTW↑
EnvDrop* (Tan et al., 2019)	-	0.52	0.41	0.53	-	0.27	-	0.29	0.18	0.34	-	0.09
OAAM (Qi et al., 2020)	-	0.56	0.49	0.54	-	0.32	-	0.29	0.18	0.34	-	0.11
NvEM (An et al., 2021)	5.38	0.54	0.47	0.51	0.48	0.35	6.80	0.38	0.28	0.41	0.36	0.20
RecBERT* (Hong et al., 2021)	4.82	0.56	0.46	0.50	0.56	0.38	6.48	0.43	0.32	0.41	0.42	0.21
LOViS (our model)	**4.16**	**0.67**	**0.58**	**0.56**	**0.58**	**0.43**	**6.07**	**0.45**	**0.35**	**0.45**	**0.43**	**0.23**

Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3674-3683.

Related Venues

Spatial Language Understanding (SpLU) workshop at EMNLP-2020:

* https://spatial-language.github.io/

A combined version of SpLU with RoboNLP at ACL-2021:
https://splu-robonlp2021.github.io/

There are plans for a next one for 2023.

Table of Content