Tutorial: Spatial and Temporal 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

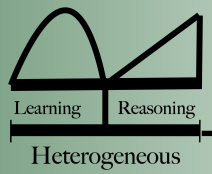
James Pustejovsky, Brandeis University, USA, jamesp@cs.brandeis.edu

Qiang Ning, Amazon, USA, qning@amazon.com

2024 Annual Conference of North American Chapter of the Association for Computational Linguistics (NAACL-2024)

Mexico City, Mexico

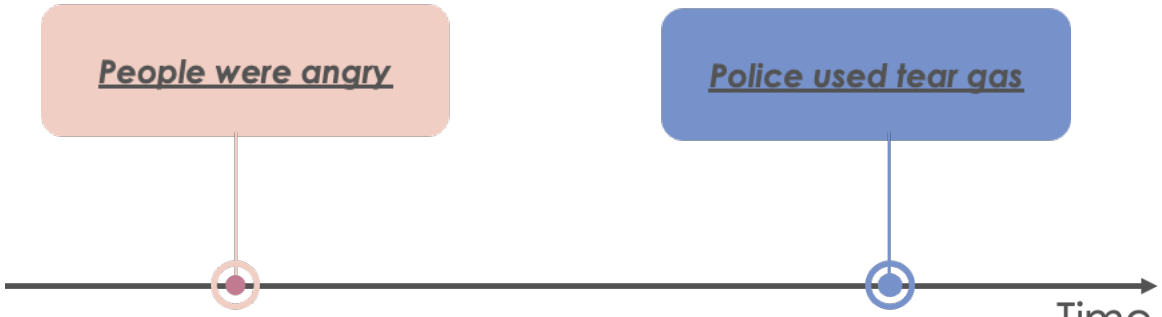
June 16



Introduction

Spatial & Temporal Understanding is important

- Are these two events relevant? (Space)
- Which one causes the other? (Time)



*People **were angry** first (likely causing chaos), and then the police **used tear gas** (to restore order).*

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!"

- Lexical variability

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

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

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

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

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

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

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

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

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

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

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 … .

…

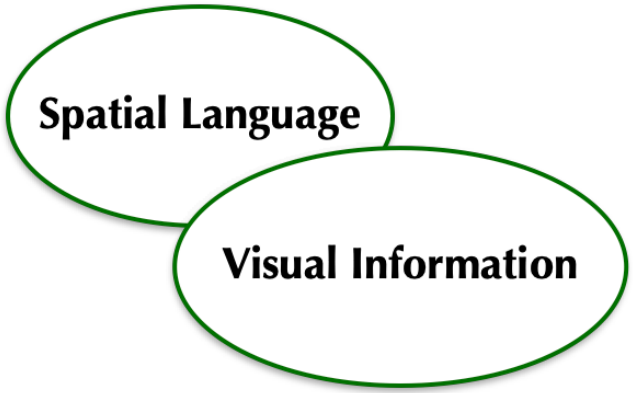
Great! You are on top of it!"

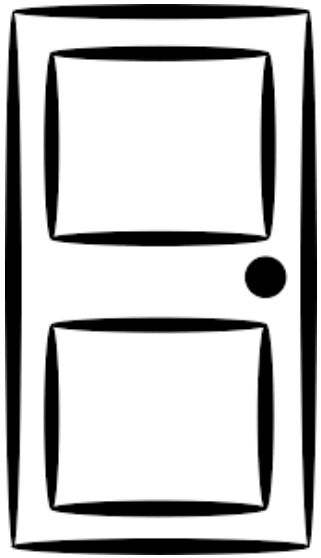
- Lexical variability
- Structural variability
- Polysemy
- Ambiguity
- Discourse and Common Sense

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!"

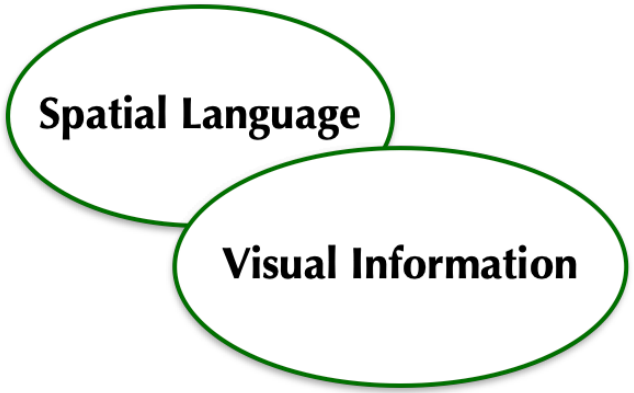


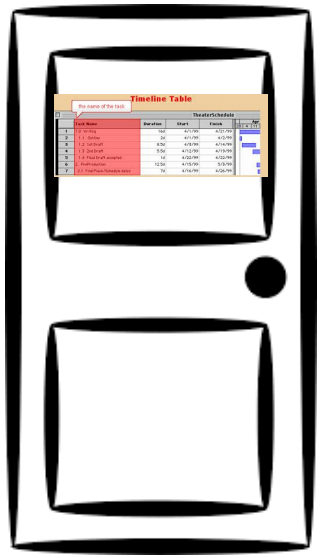


Spatial Language Challenges

"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!"

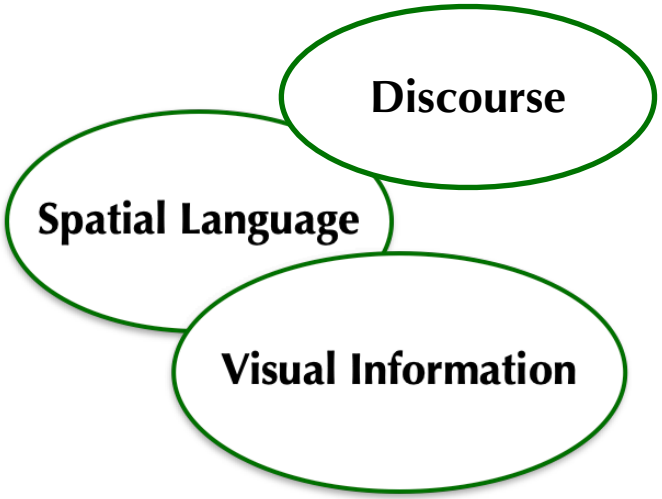




Spatial Language Challenges

"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!"

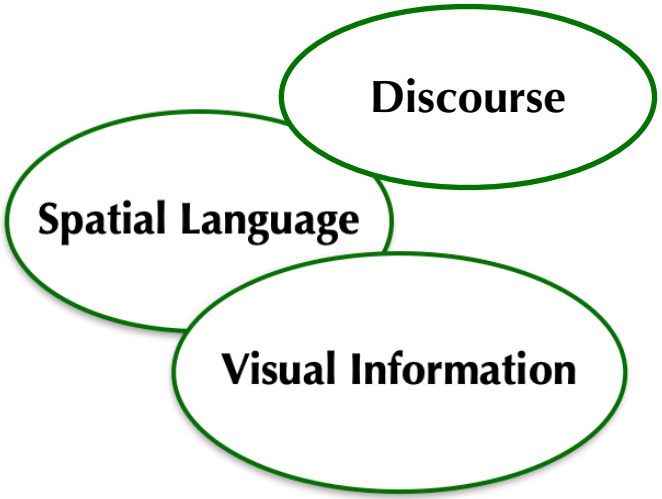


Spatial Language Challenges

"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!"

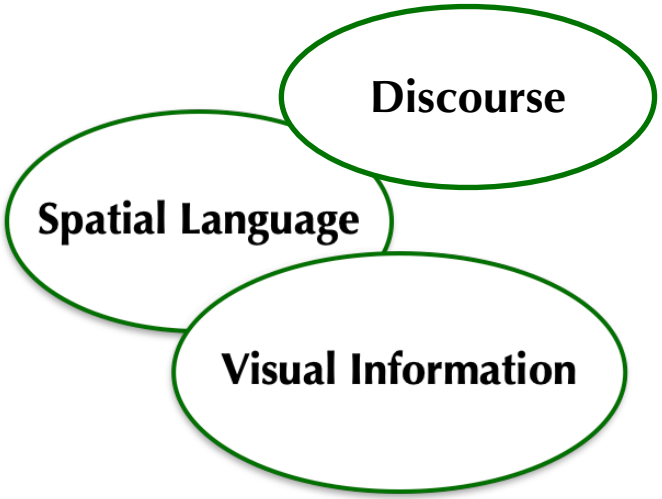


Spatial Language Challenges

"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!"

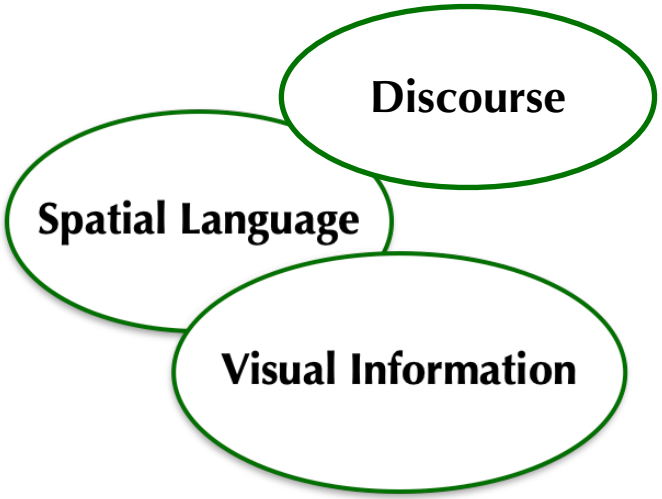


Spatial Language Challenges

"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!"

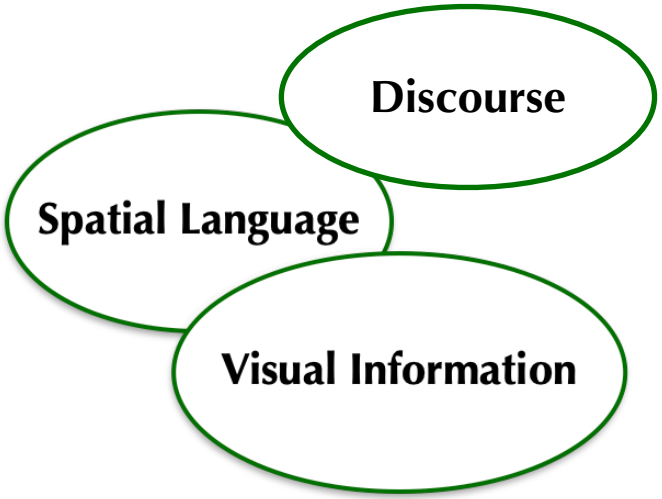


Spatial Language Challenges

"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!"

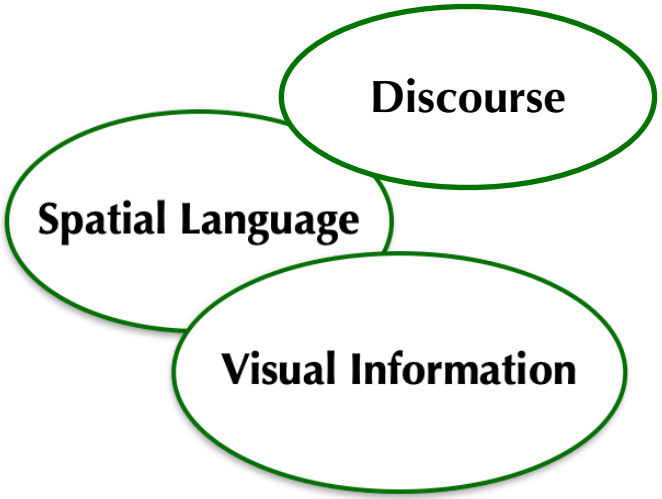


Spatial Language Challenges

"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

Navigation Instruction Following

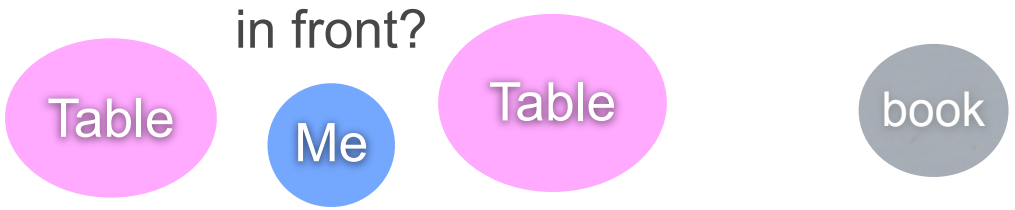
"Give me the book on AI on the big table in front of you!"

Huh?!

Spatial Language Applications

Navigation Instruction Following

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

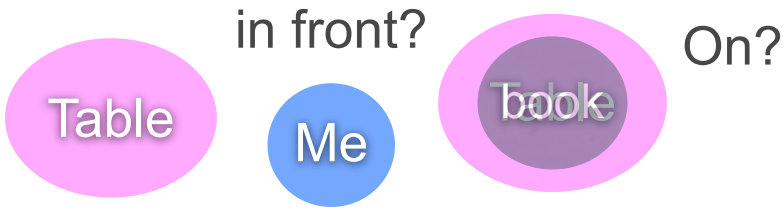


Huh?!

Spatial Language Applications

Navigation Instruction Following

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



Aha!

Spatial Language Applications

Spatial Language Applications

Text to Scene conversion (Visualization)

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

Spatial Language Applications

Text to Scene conversion (Visualization)

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

Spatial Language Applications

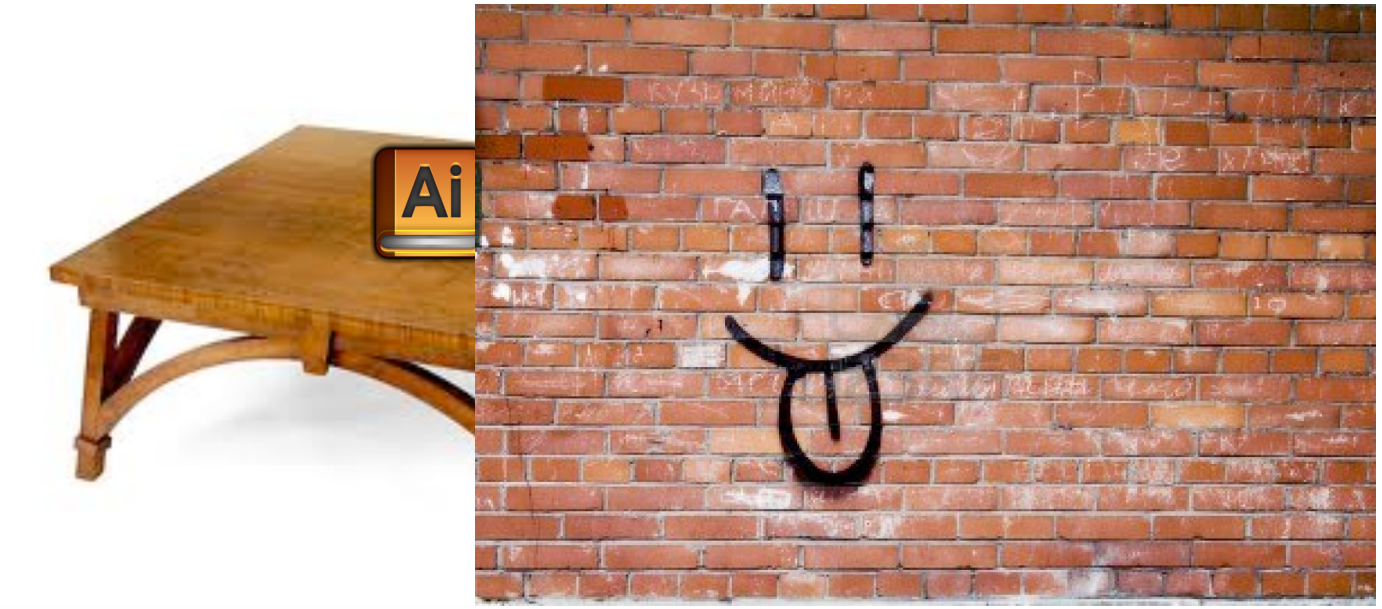
Text to Scene conversion (Visualization)

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

Spatial Language Applications

Text to Scene conversion (Visualization)

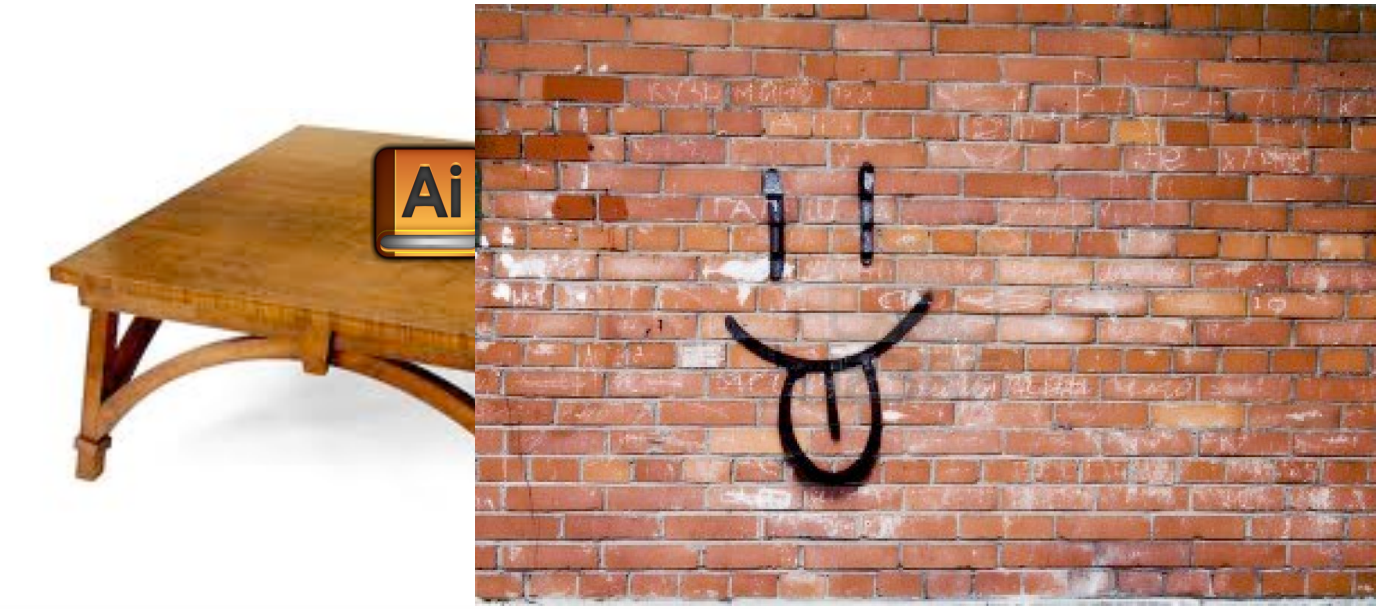
“The book on AI is on the big table behind the wall.”



Spatial Language Applications

Text to Scene conversion (Visualization)

"The book on AI is on the big table behind the wall."



..and nowadays video generation…

Spatial Language Applications

ref: Google images, dpreview.com

Spatial Language Applications

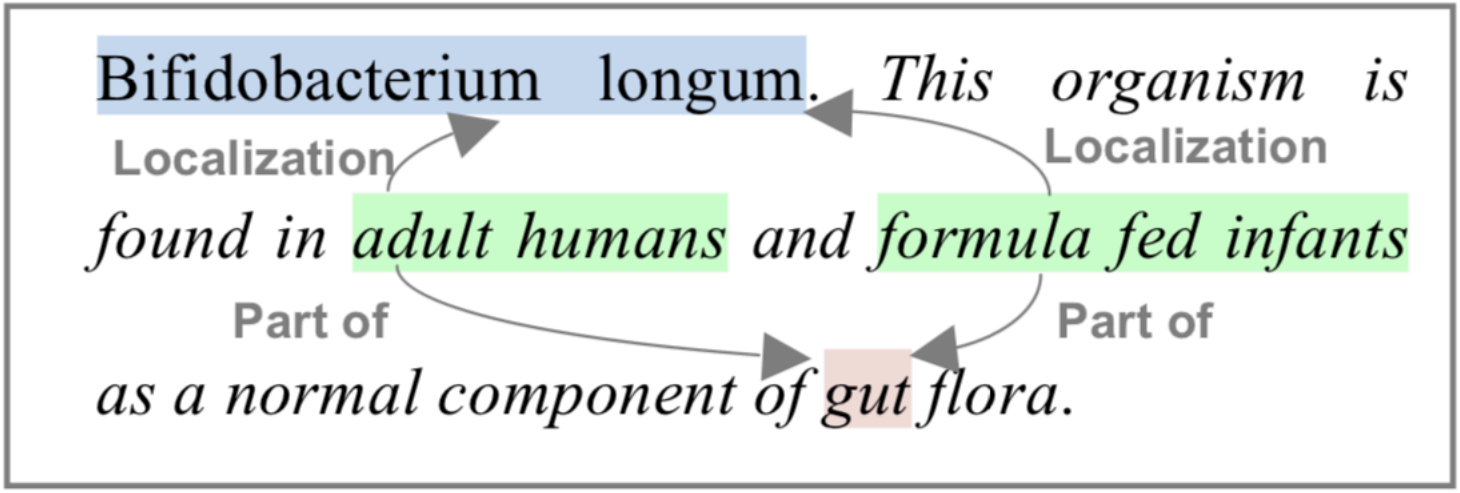
Scene to Text conversion (Image captioning)

ref: Google images, dpreview.com

Spatial Language Applications

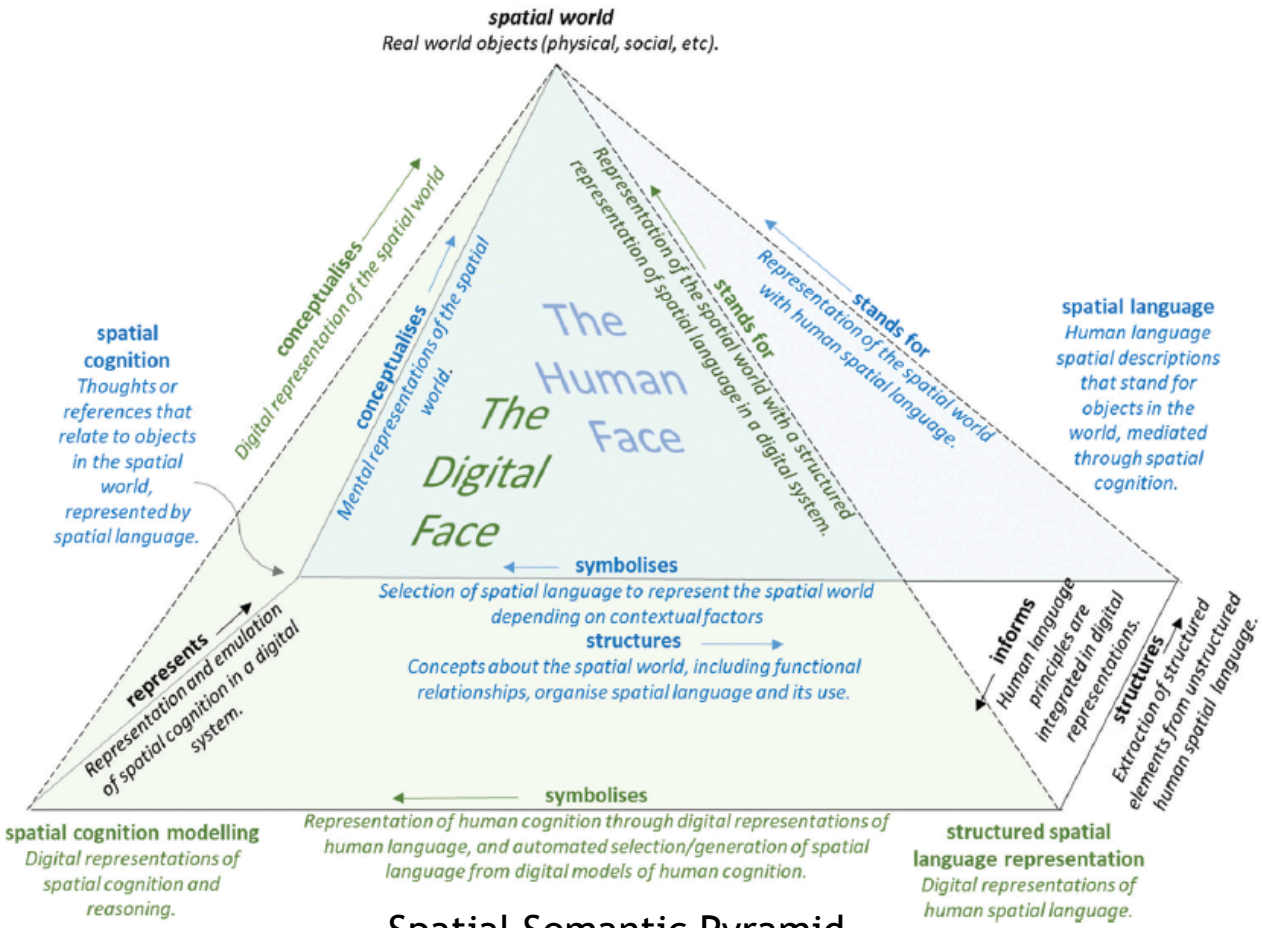
Scientific text: Biomedical Domain

- 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.

Outline

- Introduction 10 mins (Parisa):
  - Introduction to Spatiotemporal Challenges
- Parisa 45 mins:
  - Spatial Annotation Schemes
  - Spatial Information Extraction
  - Spatial Reasoning over Text: Spatial QA
  - Spatial Reasoning over Vision and Language
  - Downstream Application: Navigation
- Qiang 40 mins
  - Temporal Annotation Schemes
  - Temporal Information Extraction
  - Temporal Reasoning: Temporal QA
  - Downstream Application: Robot Path Planning
- James 40 mins
  - Spatial reasoning with AMR and GLAMR;
  - Situated grounding: Human-Object Interactions and more
  - Dense paraphrasing - Data augmentation, contextualization, entailment,
  - Vision Language Action Models
- Sien 40 mins
  - Spatial Commonsense with LLMs
  - Projection of Language into Physical Coordinates (Spatial)
  - Projection of Events into 1D timelines (Temporal)
- Conclusion 5 mins (Parisa)

Outline

- Introduction 10 mins (Parisa):
  - Introduction to Spatiotemporal Challenges
- Parisa 45 mins:
  - **Spatial Annotation Schemes**
  - **Spatial Information Extraction**
  - **Spatial Reasoning over Text: Spatial QA**
  - **Spatial Reasoning over Vision and Language**
  - **Downstream Application: Navigation**
- Qiang 40 mins
  - Temporal Annotation Schemes
  - Temporal Information Extraction
  - Temporal Reasoning: Temporal QA
  - Downstream Application: Robot Path Planning
- James 40 mins
  - Spatial reasoning with AMR and GLAMR;
  - Situated grounding: Human-Object Interactions and more
  - Dense paraphrasing - Data augmentation, contextualization, entailment,
  - Vision Language Action Models
- Sien 40 mins
  - Spatial Commonsense with LLMs
  - Projection of Language into Physical Coordinates (Spatial)
  - Projection of Events into 1D timelines (Temporal)
- Conclusion 5 mins (Parisa)

Spatial Semantic Representations

Spatial Representation

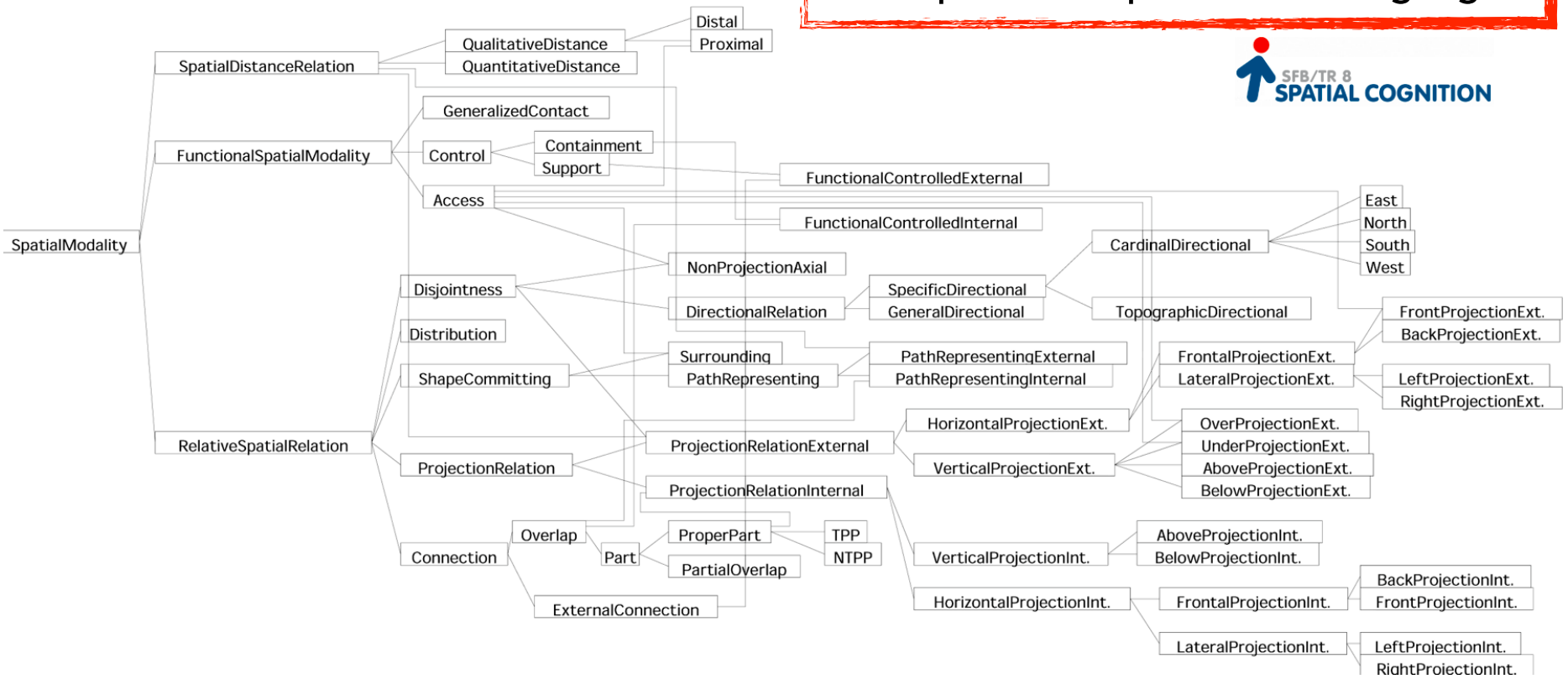
How space is expressed in language?

How can we automatically reason over spatial language? Qualitative/Quantitative

Linguistically Motivated Representations

General Upper Model (GUM) ontology

How space is expressed in language?

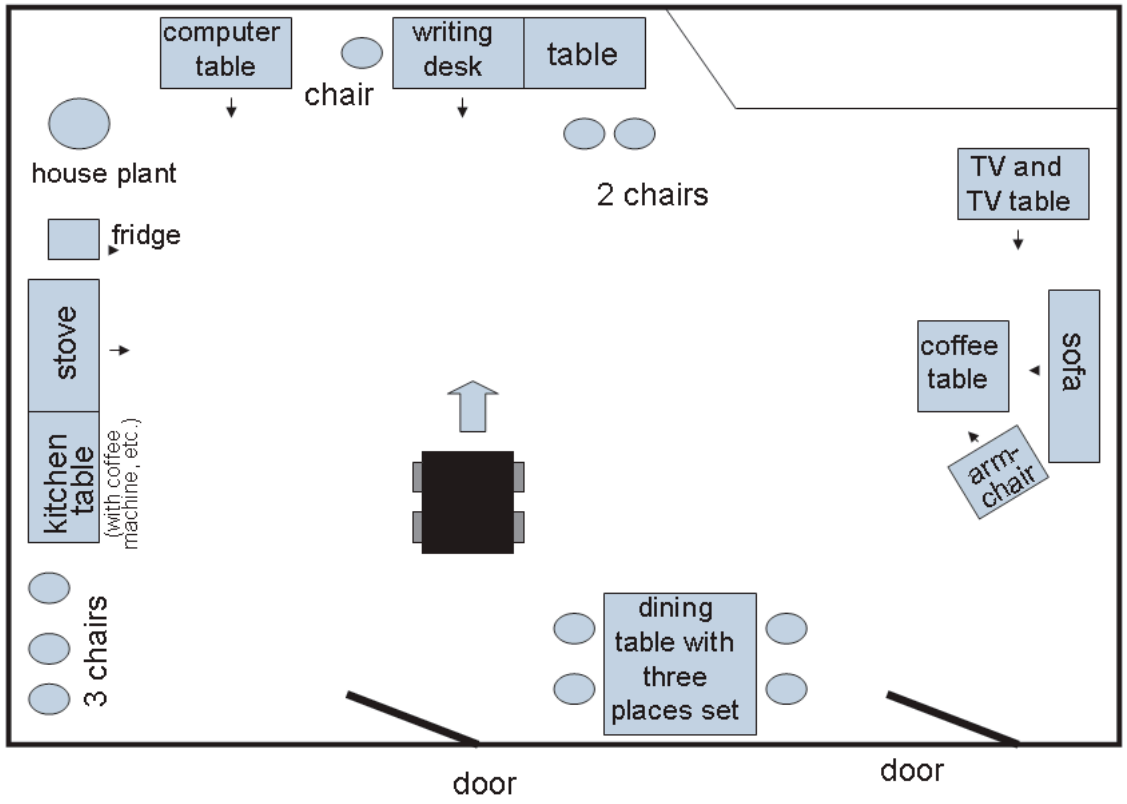


[J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink. A linguistic ontology of space for natural language processing. Artificial Intelligence, 174(14):1027–1071, 2010.]

[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

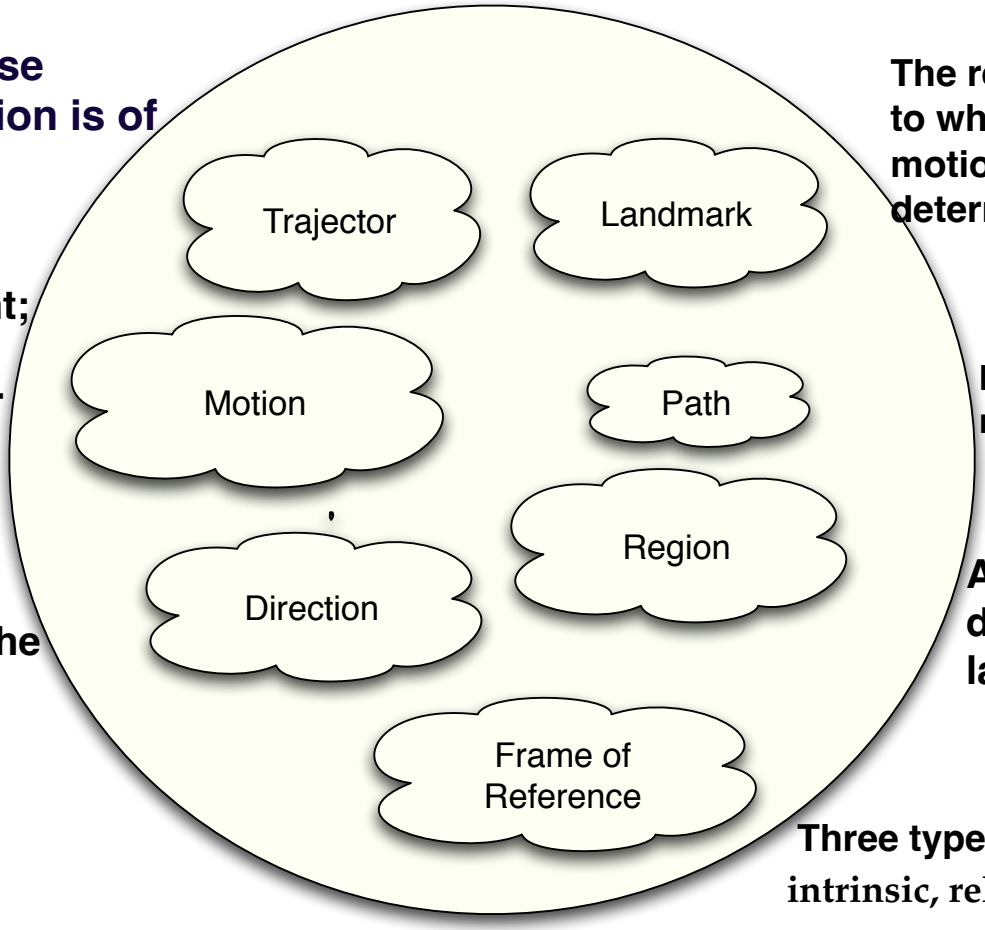
SFB/TR 8 SPATIAL COGNITION

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.

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

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

Three types of FoR:
intrinsic, relative or absolute.

How space is expressed in language?

[J. Zlatev. Spatial semantics. The Oxford Handbook of Cognitive Linguistics , pages 318–350. Oxford Univ. Press, 2007.]

Spatial Logic

How can we automatically reason over spatial language? Qualitative vs. Quantitative

What formal representation is appropriate for spatial reasoning?

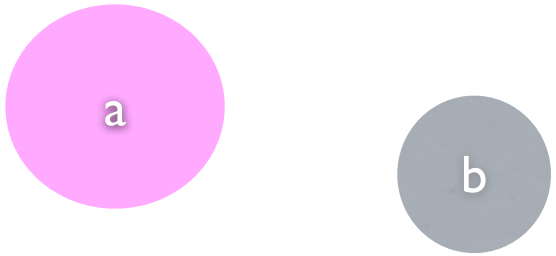
Spatial Knowledge Representation

Spatial Knowledge Representation

Formal representation of the meaning.

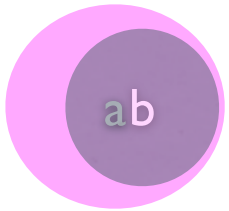
Spatial Knowledge Representation

Formal representation of the meaning.



Spatial Knowledge Representation

Formal representation of the meaning.



Disconnected?

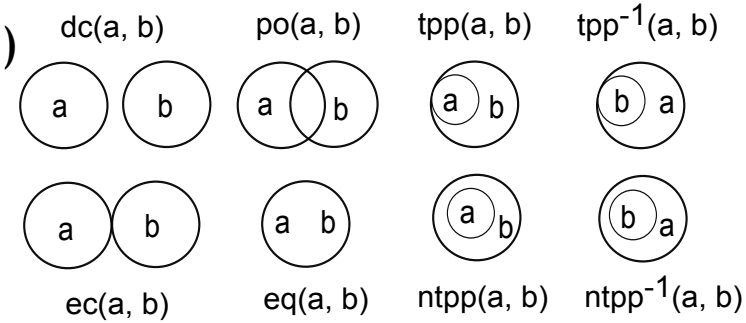
Touch?

Overlap?

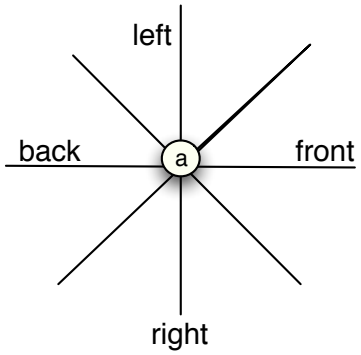
Within?

Qualitative Representation and Reasoning

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



The RCC-8 relations.



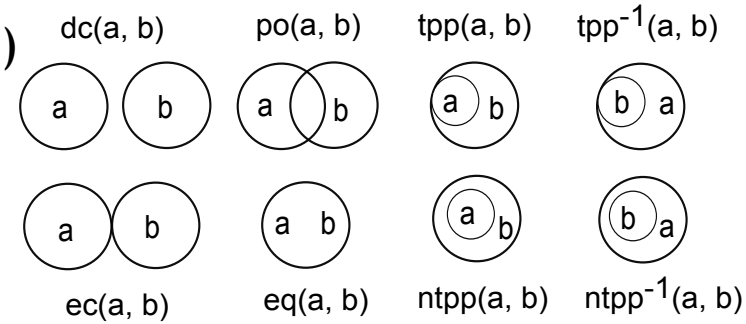
[Cohn, Anthony & Hazarika, Shyamanta. (2001). Qualitative Spatial Representation and Reasoning: An Overview. Fundamenta Informaticae, 46. 1-29]

[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]

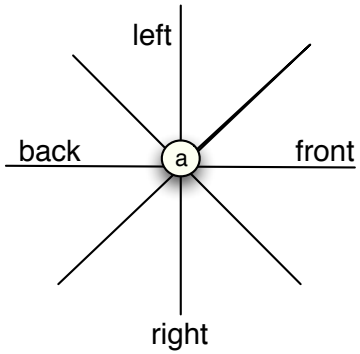
Qualitative Representation and Reasoning

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



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.

The RCC-8 relations.



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

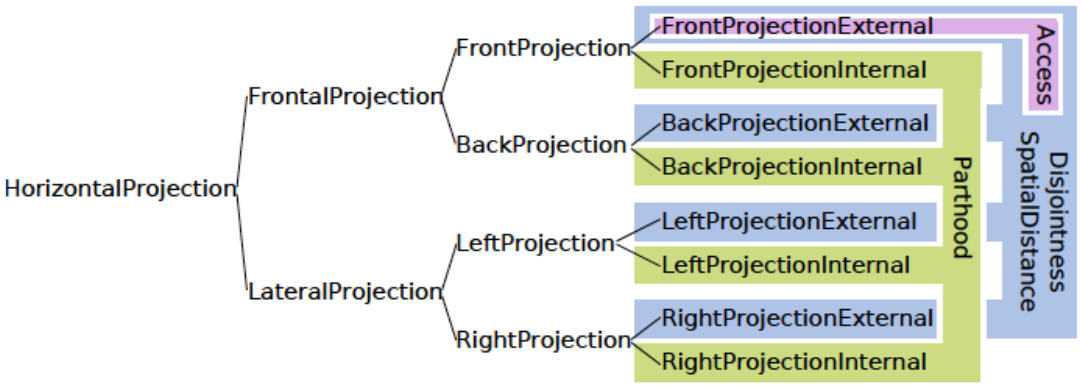
[Andrew U. Frank, Qualitative Spatial Reasoning: Cardinal Directions as an Example, Geographical Information Systems 10(3):269-290, 1996]

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

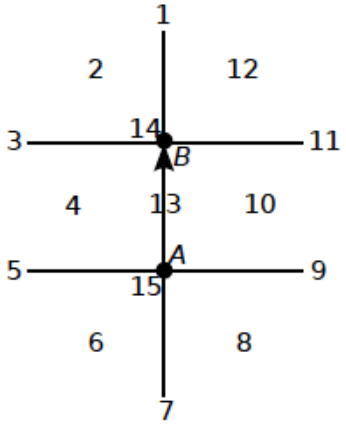
Connecting Linguistic Representation to Formal Calculi Representations

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,

Book: Annotation-based Semantics for Space and Time in Language, Kiyong Lee, Cambridge University Press, 2023

Corpus Annotations

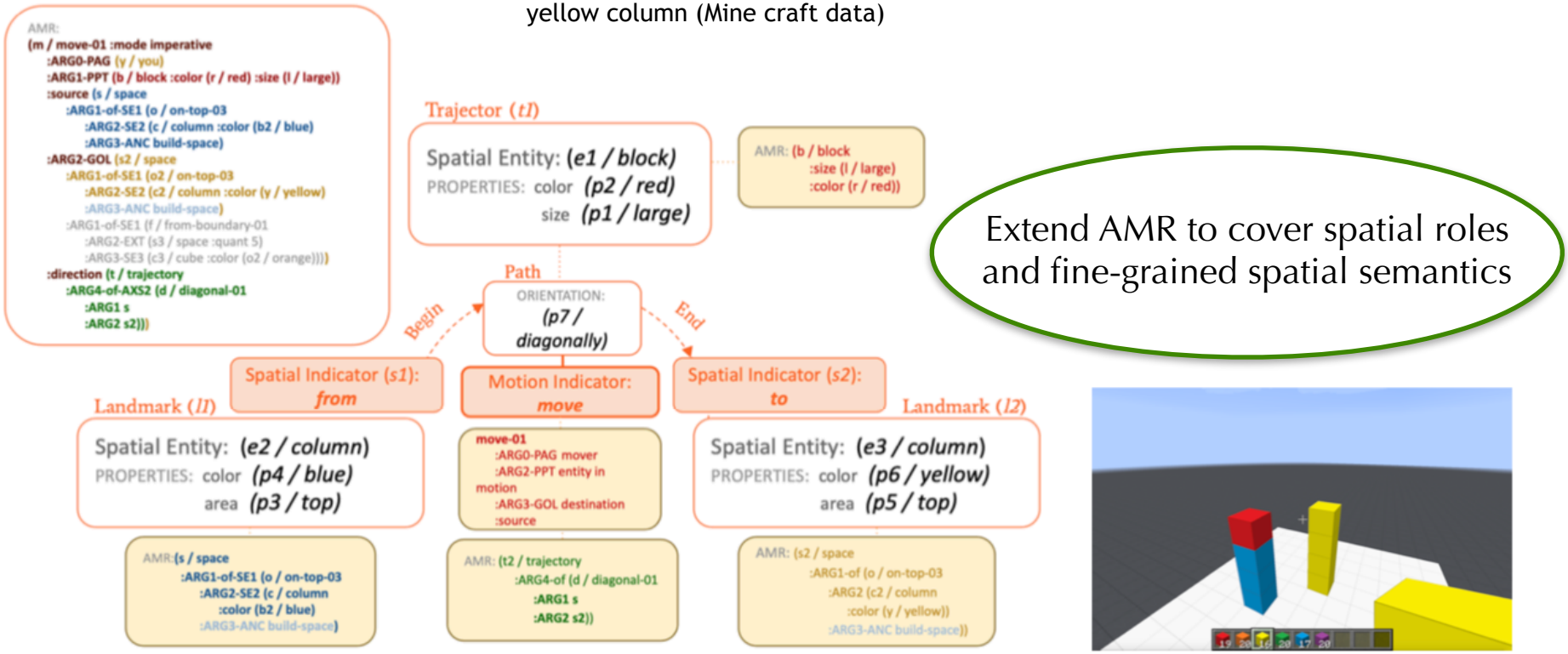
Annotations were applied on various datasets of different nature

- Degree Confluence Project (DCP)
- Cross Language Evaluation Forum (CLEF, images with textual descriptions)
- 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)



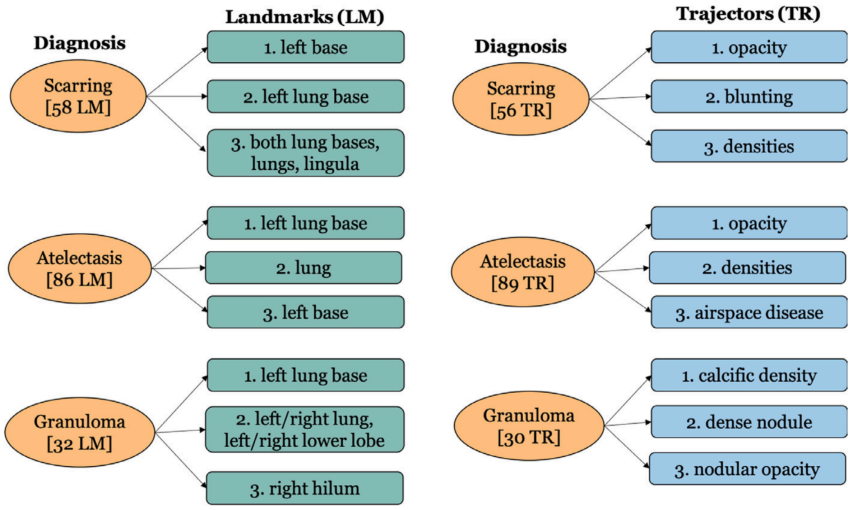
[Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, Dan Roth, Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, Dan Roth. LREC 2020.]
[Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus, Julia Bonn,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

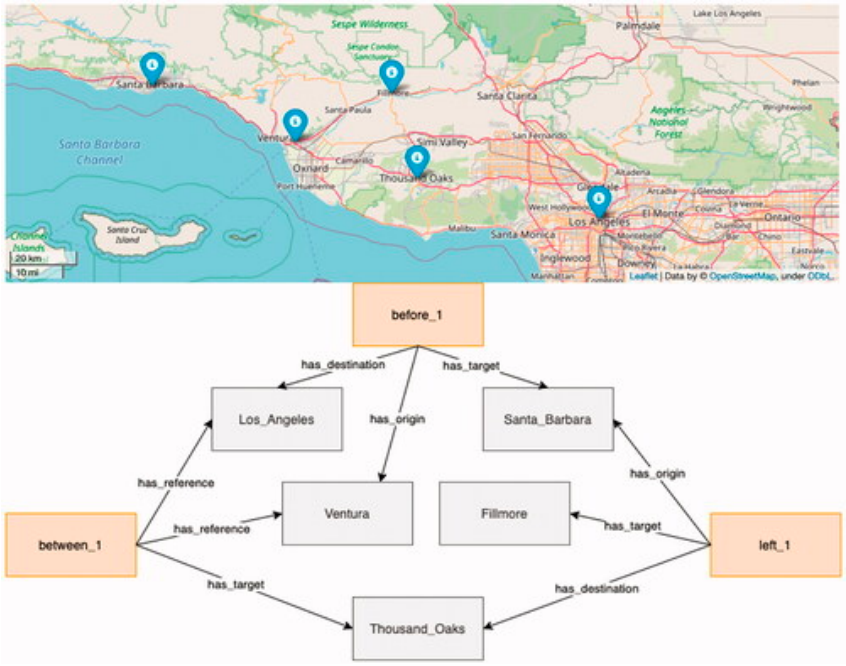
[Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports, Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F. Riascos-Castaneda, Kirk Roberts, LREC 2020.]

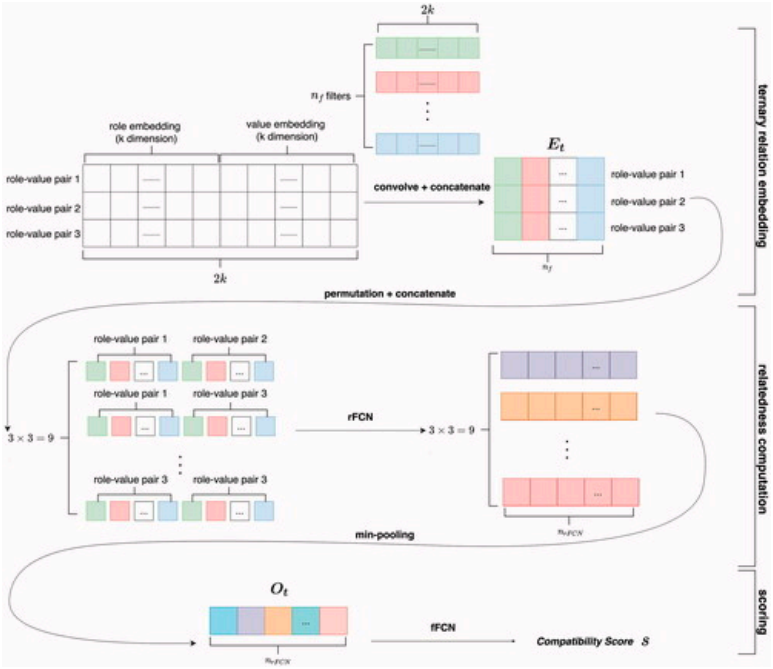
[SpatialNet: A Declarative Resource for Spatial Relations, Morgan Ulinski, Bob Coyne, Julia Hirschberg, SpLU-RoboNLP-2019]

[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.

Spatial 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 Information Extraction

Spatial IE can be seen as:

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

Classically, semantic parsing and its shallow/ domain-specific form, i.e., information extraction, were important steps for reasoning over natural language.

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.

Information Extraction

Two Layers of Semantics:

Based on cognitive linguistic elements and multiple calculi.

1. **SpRL**: Spatial role labeling

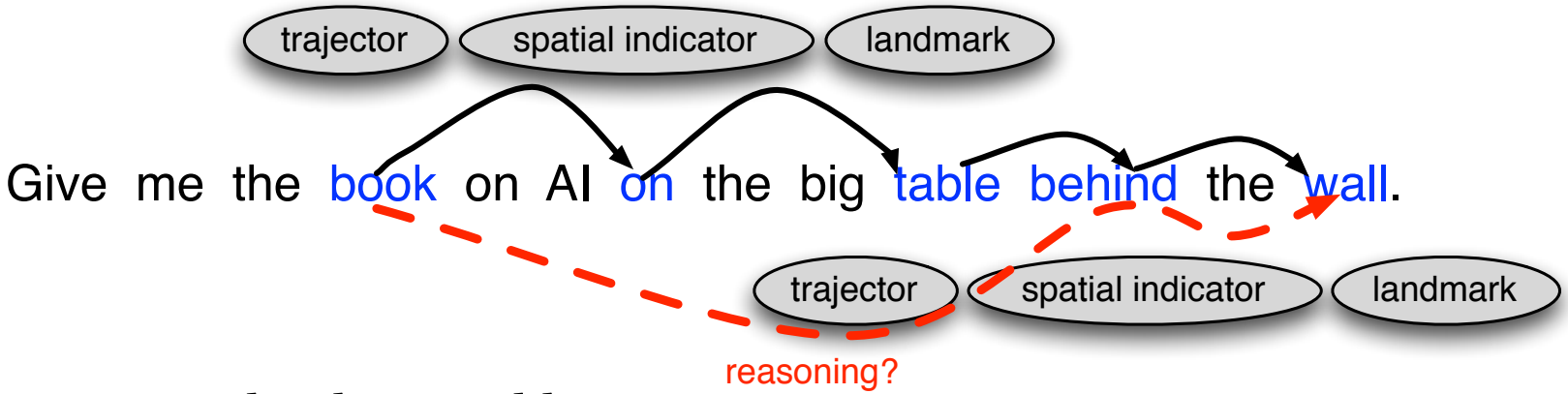
   -Identifying objects, roles and relations

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

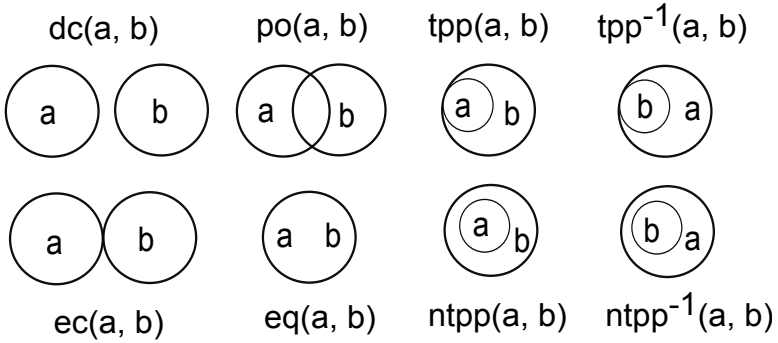
- **Topological**

{EQ, DC, EC, PO, PP}

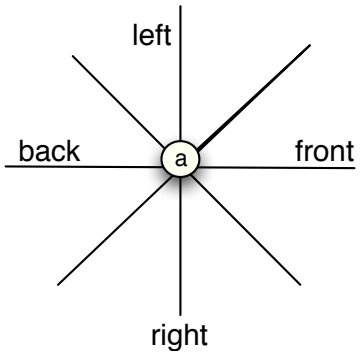
- **Directional**

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

- **Distal**



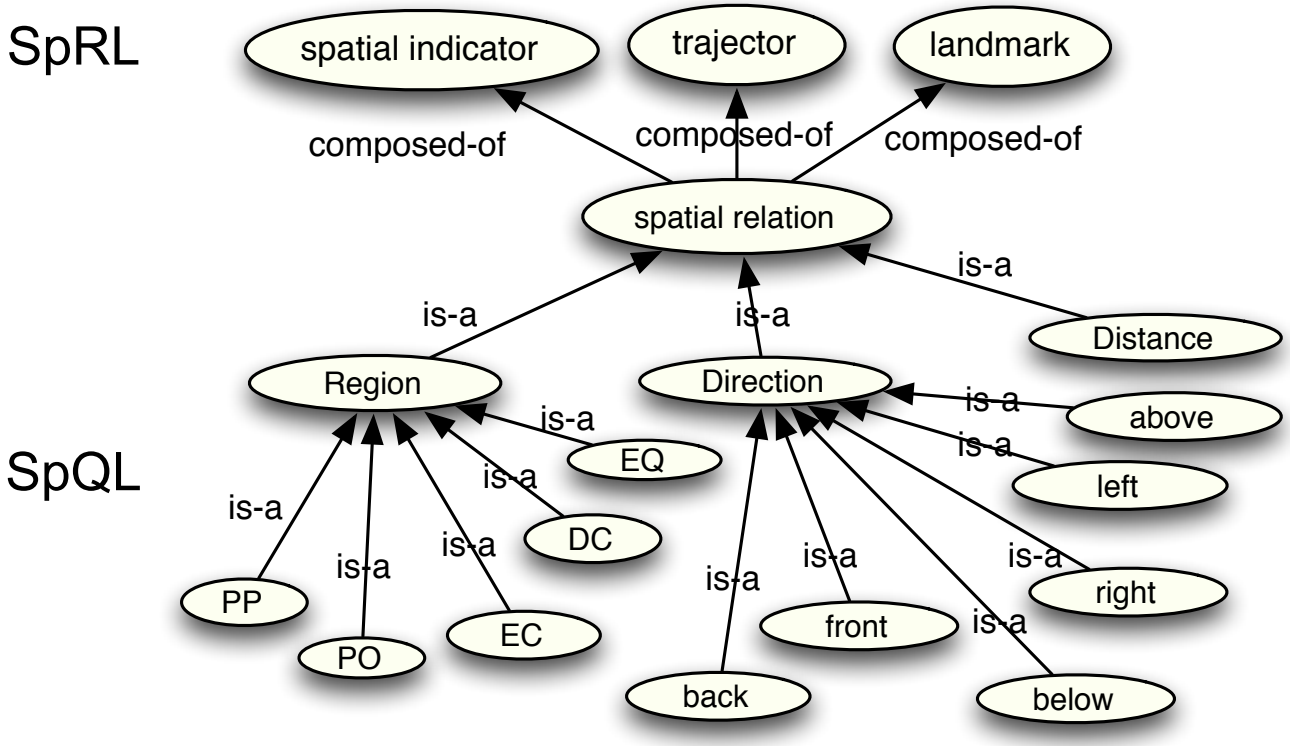
The RCC-8 relations.



[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.

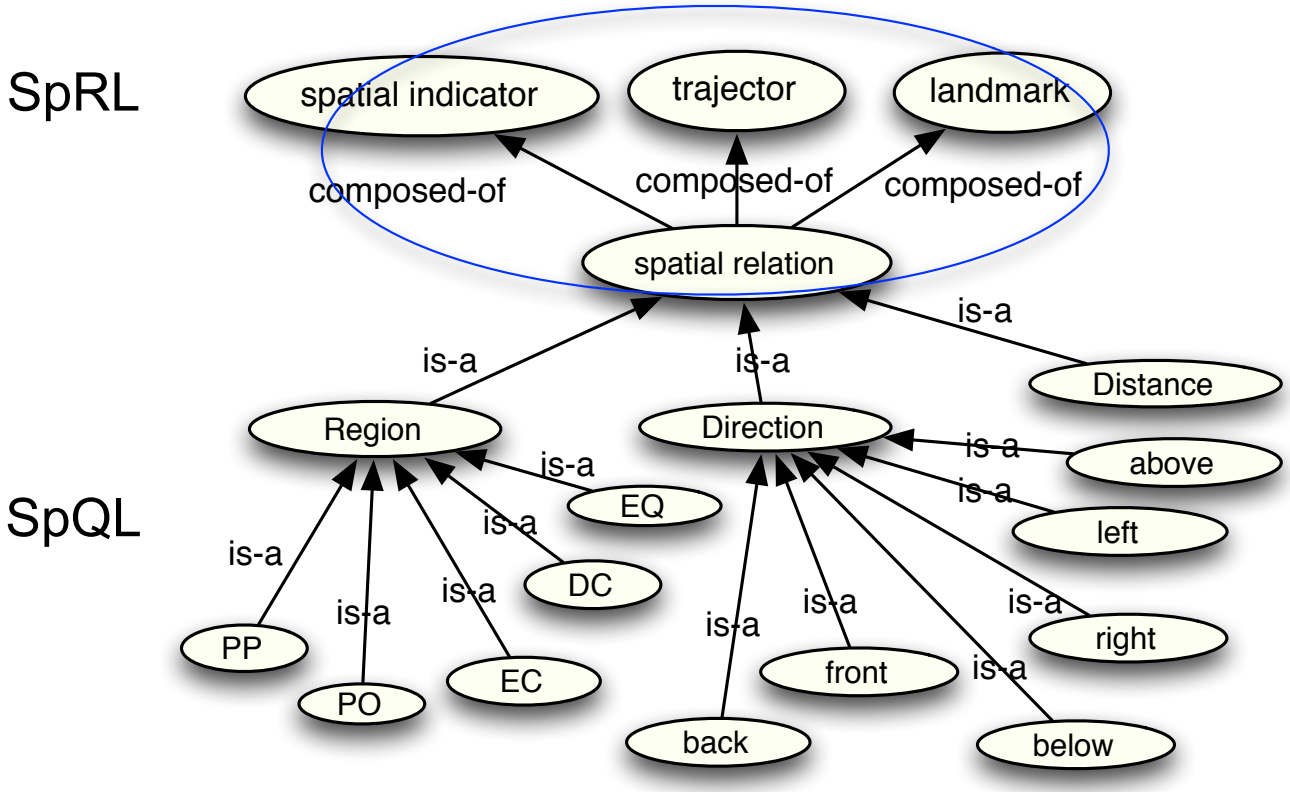


[Kordjamshidi, P., van Otterlo, M., Moens, M. F., Spatial role labeling: Task definition and annotation scheme. LREC-2010).]

[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.]

Spatial Ontology

Based on cognitive linguistic elements and multiple calculi.

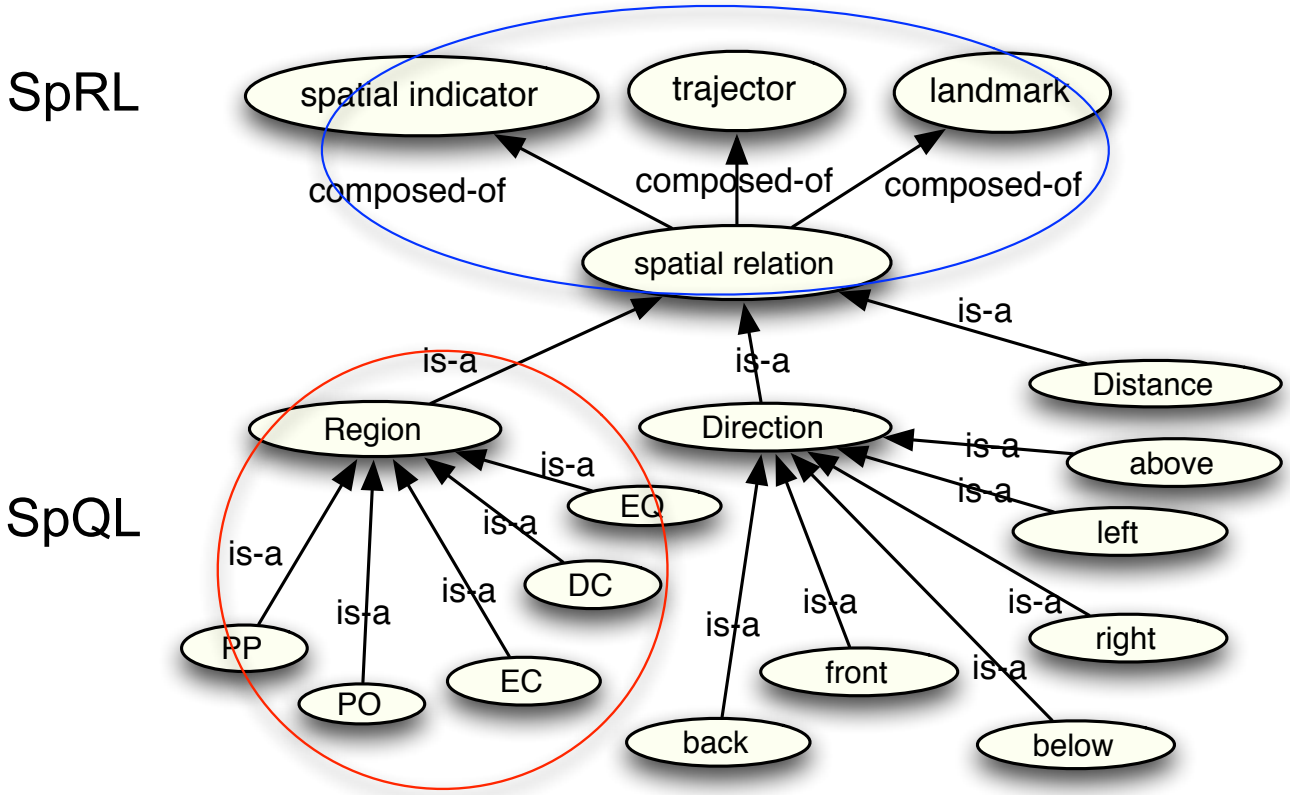


[Kordjamshidi, P., van Otterlo, M., Moens, M. F., Spatial role labeling: Task definition and annotation scheme. LREC-2010).]

[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.]

Spatial Ontology

Based on cognitive linguistic elements and multiple calculi.

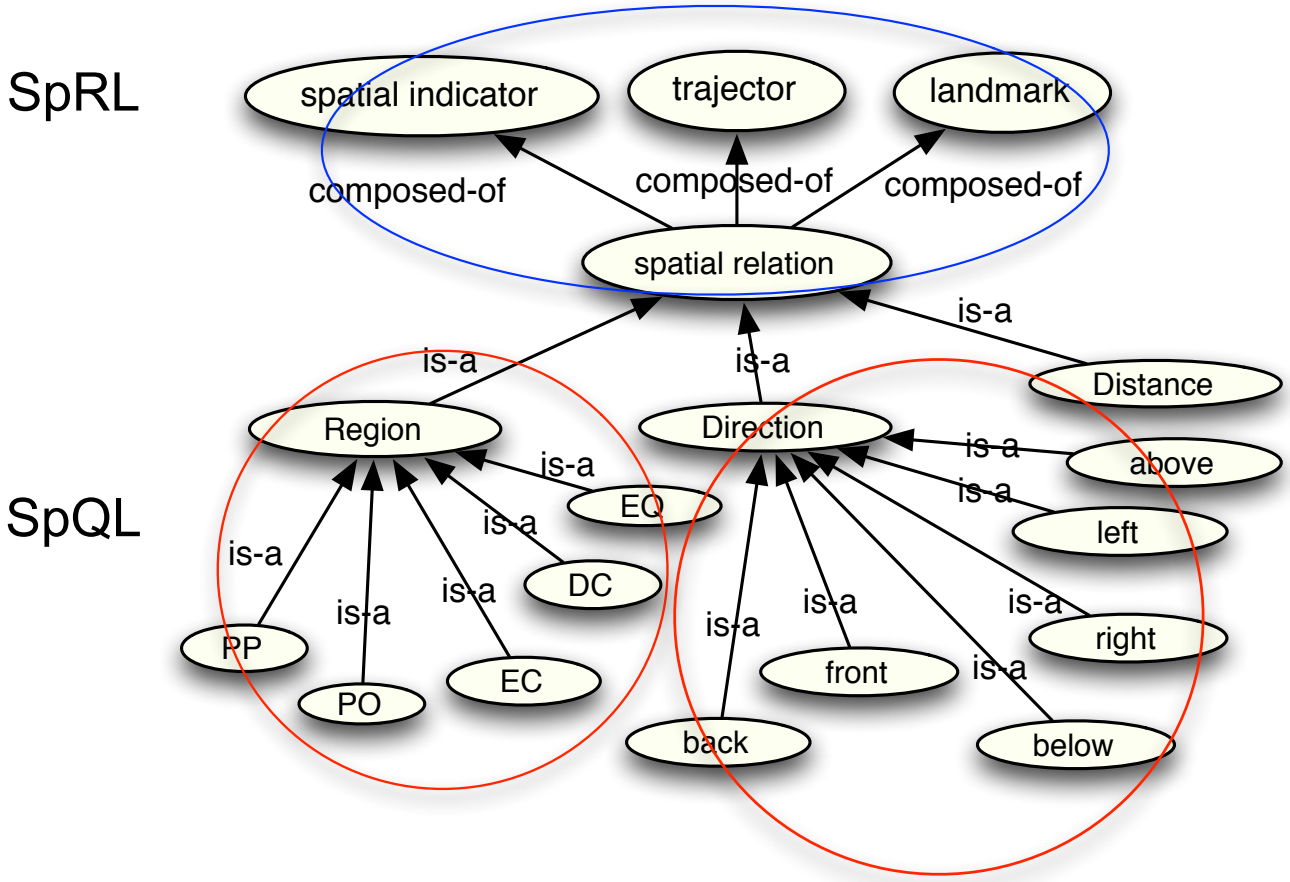


[Kordjamshidi, P., van Otterlo, M., Moens, M. F., Spatial role labeling: Task definition and annotation scheme. LREC-2010).]

[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.]

Spatial Ontology

Based on cognitive linguistic elements and multiple calculi.

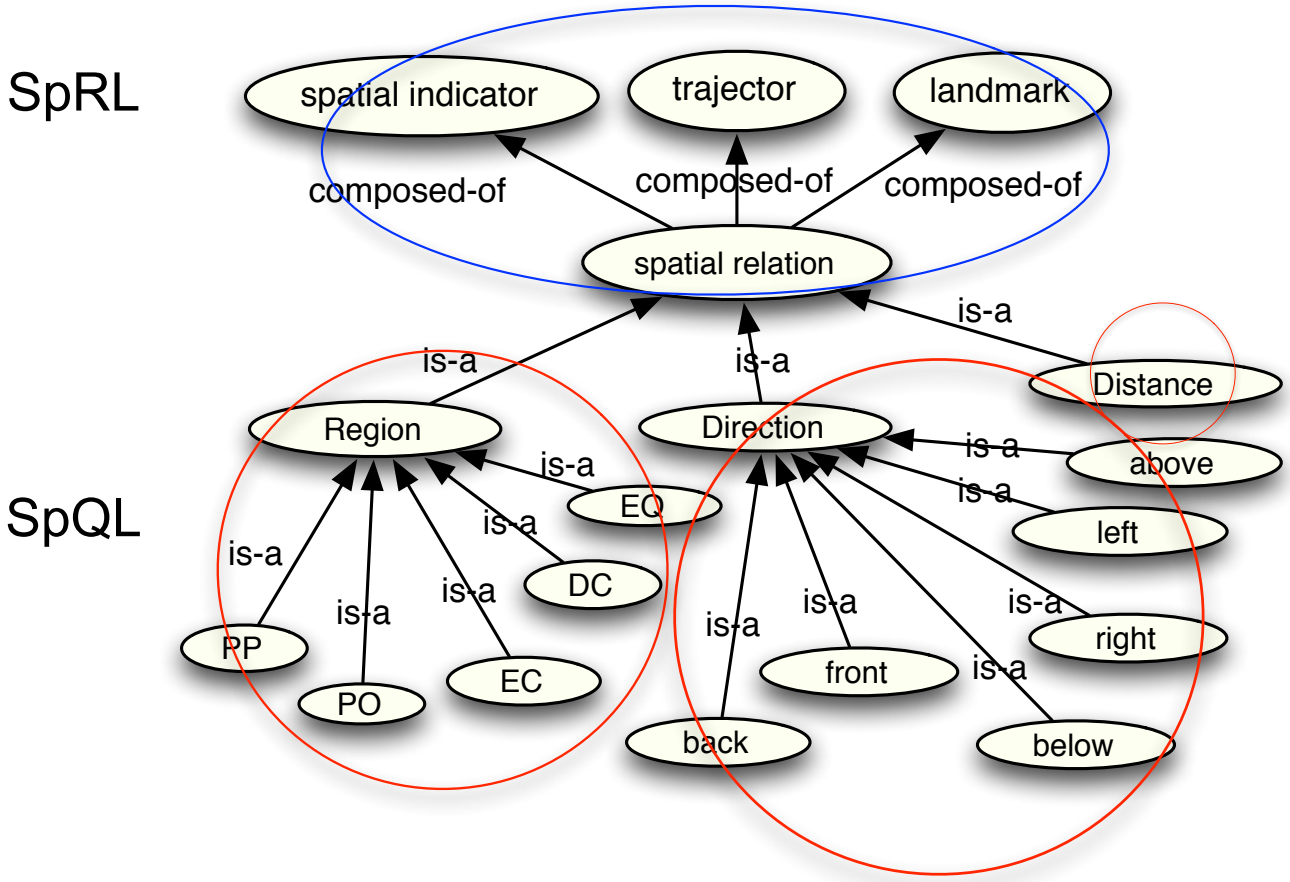


[Kordjamshidi, P., van Otterlo, M., Moens, M. F., Spatial role labeling: Task definition and annotation scheme. LREC-2010).]

[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.]

Spatial Ontology

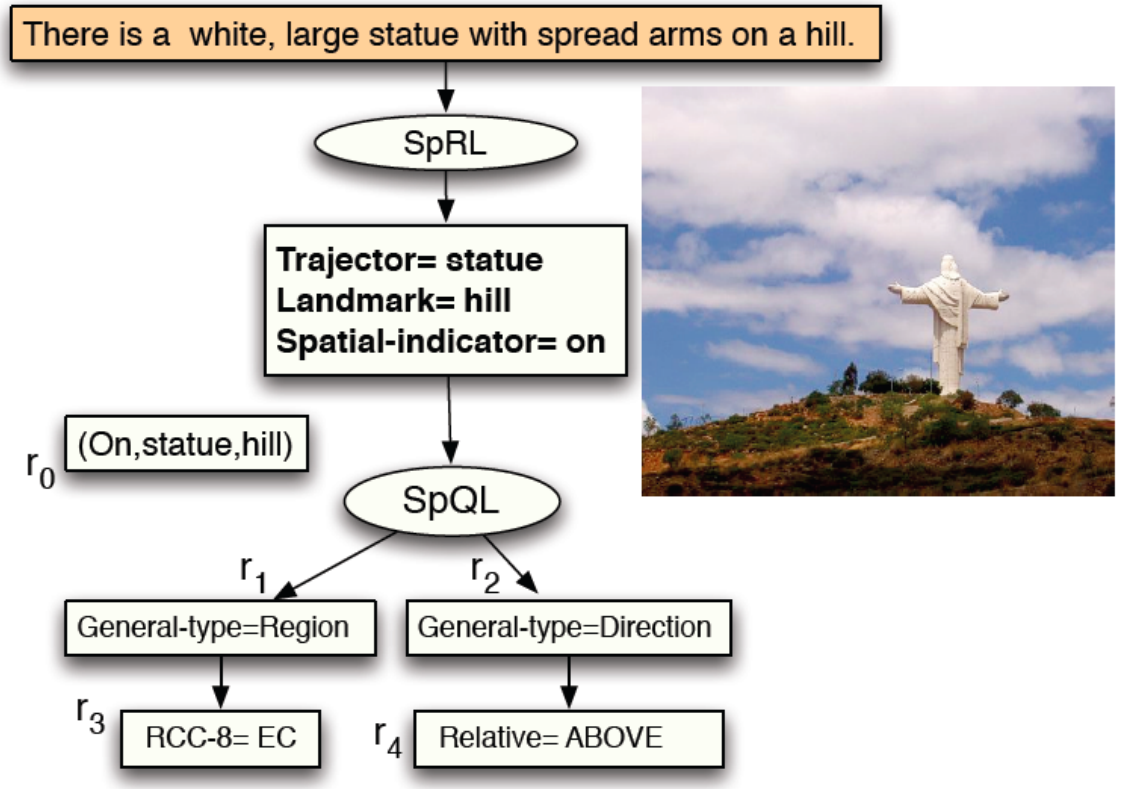
Based on cognitive linguistic elements and multiple calculi.



[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



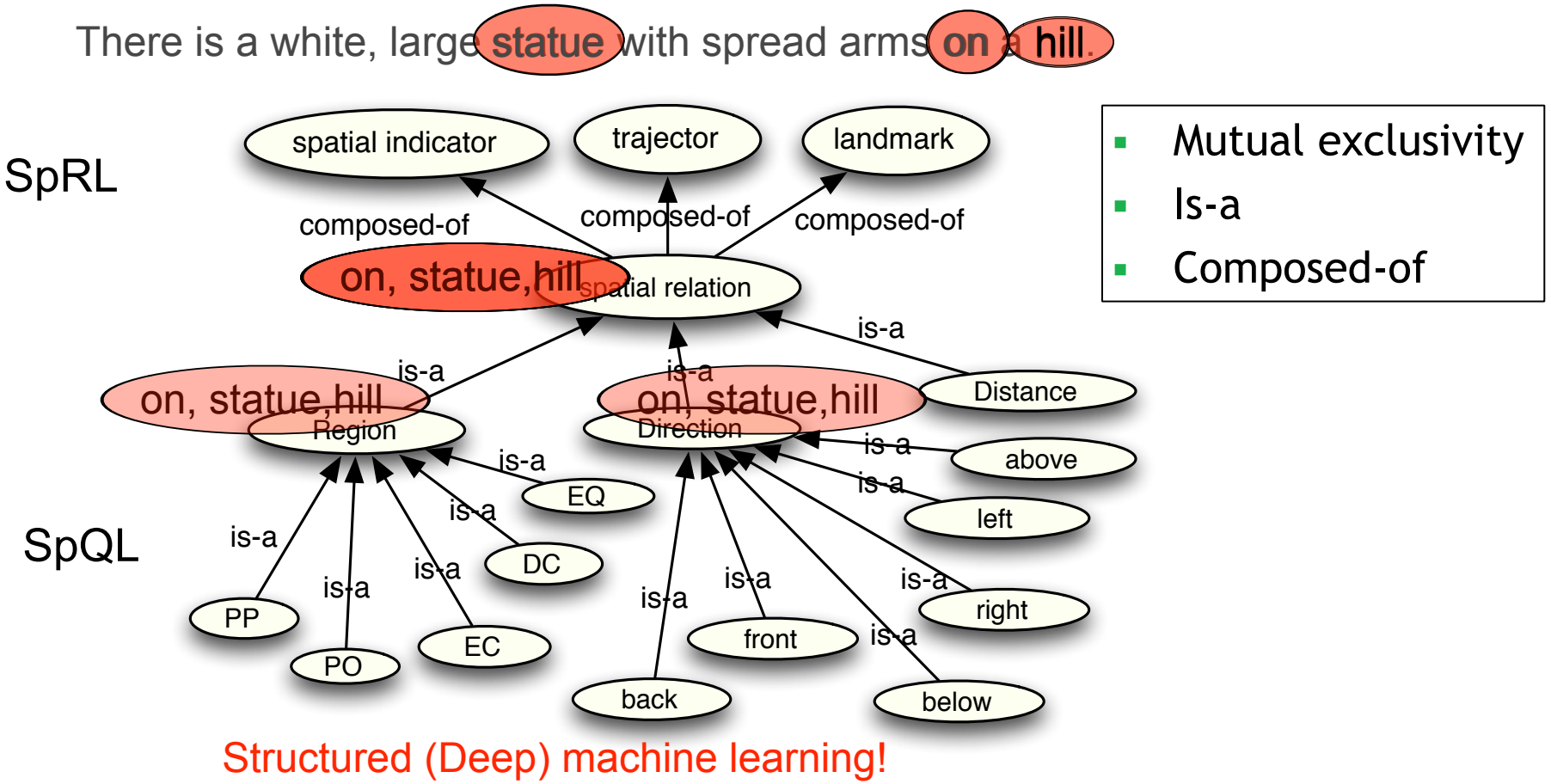
SemEval-2012/2013/2015 and CLEF/mSpRL-2017 benchmarks.

[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



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

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*

Spatial Reasoning:
Evaluation and Improving Language Models

Spatial Reasoning

Spatial Reasoning

- Nowadays: Reasoning is often evaluated by Question Answering.

Spatial Reasoning

- Nowadays: Reasoning is often evaluated by Question Answering.
- In generative models: Evaluated by generated Natural language/ Vision (image, video).

Spatial Reasoning

- Nowadays: Reasoning is often evaluated by Question Answering.
- In generative models: Evaluated by generated Natural language/ Vision (image, video).

Chris Manning in AAAI-2024 panel: we should not work on Information extraction anymore (solved!?)

Spatial Reasoning

- Nowadays: Reasoning is often evaluated by Question Answering.
- In generative models: Evaluated by generated Natural language/ Vision (image, video).

Chris Manning in AAAI-2024 panel: we should not work on Information extraction anymore (solved!?)

BUT

Spatial Reasoning

- Nowadays: Reasoning is often evaluated by Question Answering.
- In generative models: Evaluated by generated Natural language/ Vision (image, video).

Chris Manning in AAAI-2024 panel: we should not work on Information extraction anymore (solved!?)

BUT

- Depending on the task:
  - Reasoning is needed for Extraction
  - Extraction is needed for Reasoning

Spatial Question Answering

Do we have relevant corpora to evaluate spatial meaning representations and their impact on downstream tasks? (A question in 2020)

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

Checking samples of these datasets, we realized:

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

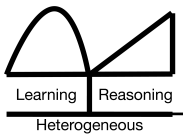
Spatial Reasoning: Lack of Benchmarks

[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Spatial Reasoning: Lack of Benchmarks

- Synthetic data using toy visual environment (SpaRTQA extended to SpaRTUN)
- Human annotations explaining toy visual environment (SpaRTQA-human)
- Realistic data curation using truistic photos for spatial QA (ReSQ)

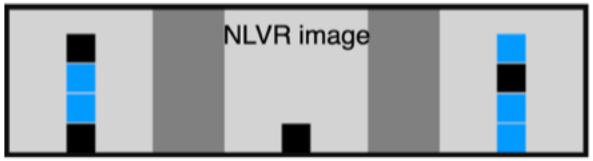


[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Spatial Reasoning: Lack of Benchmarks

- Synthetic data using toy visual environment (SpaRTQA extended to SpaRTUN)
- Human annotations explaining toy visual environment (SpaRTQA-human)
- Realistic data curation using truistic photos for spatial QA (ReSQ)



We have three block A, B and C….

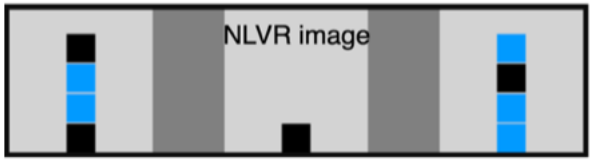
[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Learning | Reasoning
Heterogeneous

Spatial Reasoning: Lack of Benchmarks

- Synthetic data using toy visual environment (SpaRTQA extended to SpaRTUN)
- Human annotations explaining toy visual environment (SpaRTQA-human)
- Realistic data curation using truistic photos for spatial QA (ReSQ)



We have three block A, B and C….

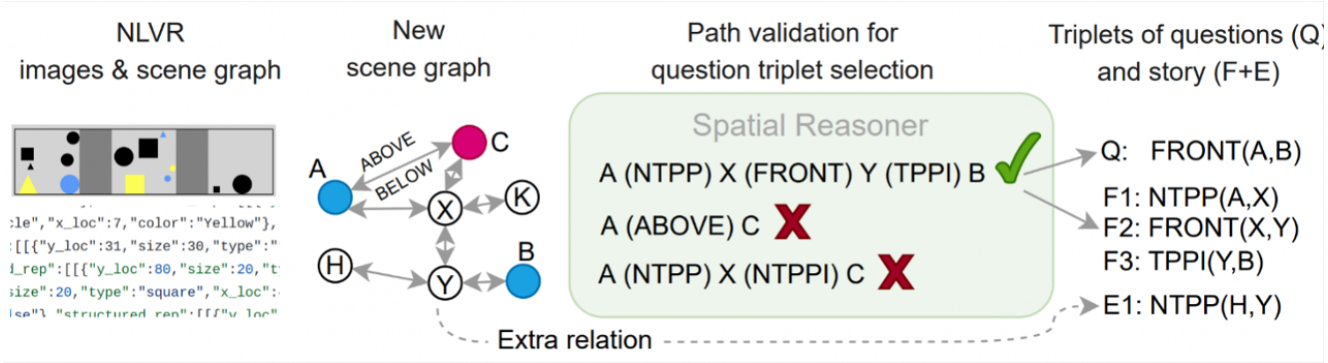
A statue with spread arms on a hill …

[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Learning Reasoning Heterogeneous

Spatial Reasoning over Text

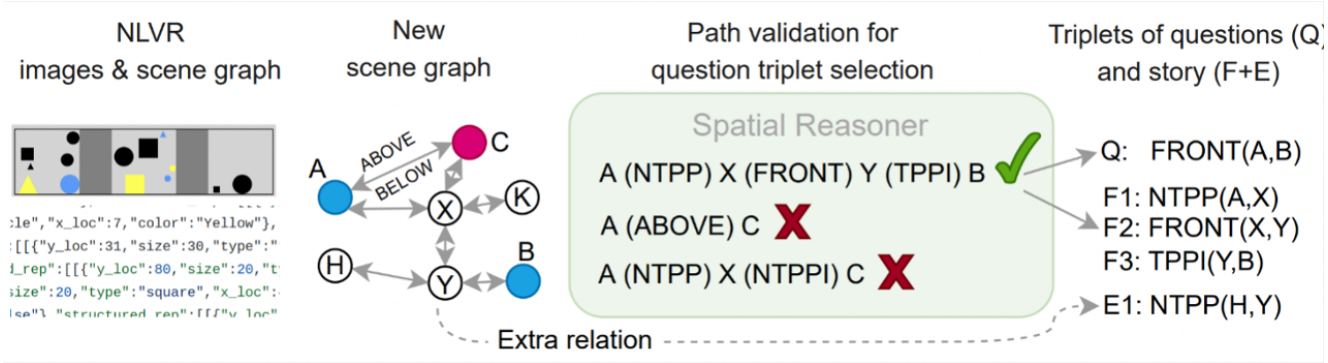


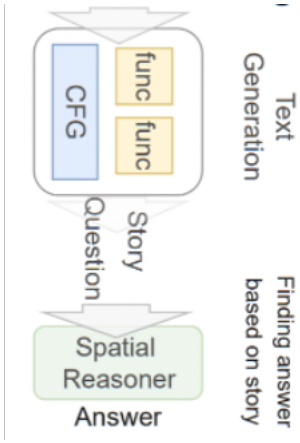
[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Learning | Reasoning
Heterogeneous

Spatial Reasoning over Text



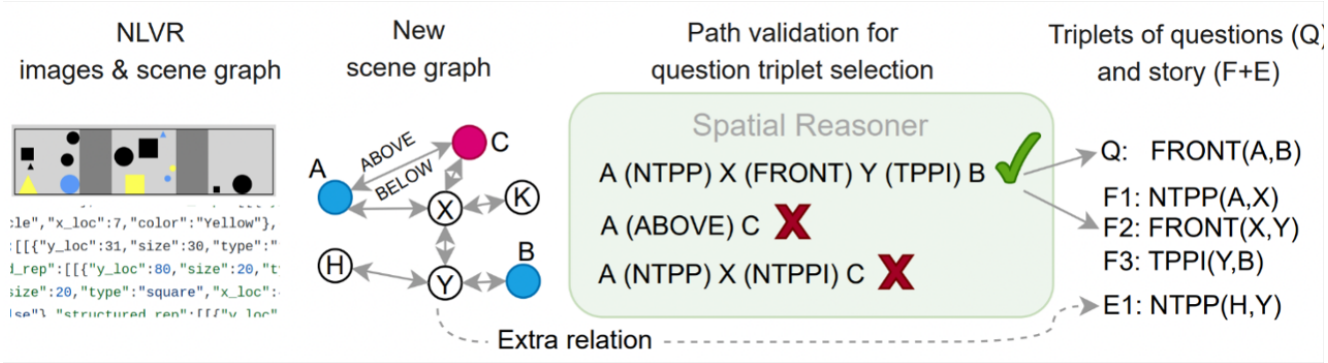


[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Learning | Reasoning
Heterogeneous

Spatial Reasoning over Text



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:

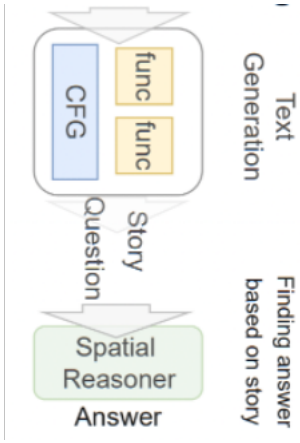
FB: Which block(s) has a medium thing that is below a black square? A, B, C

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

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

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

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



[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Learning Reasoning
Heterogeneous

SpaRTUN & ReSQ

- Spatial Question Answering
- Spatial Role Labeling
  - Spatial concepts (landmarks, trajectors, …)
  - Spatial relationships (topological relation, directional relations, …)

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

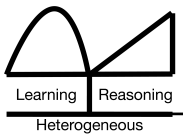
(a) SPARTUN - A synthetic large dataset provided as source of supervision

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



[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

SpaRTUN & ReSQ

- Spatial Question Answering
- Spatial Role Labeling
  - Spatial concepts (landmarks, trajectors, …)
  - Spatial relationships (topological relation, directional relations, …)

Sets	FB	FR	YN	CO	Total
SPARTQA-HUMAN:					
Test	104	106	194	107	511
Train	154	149	162	151	616
SPARTQA-AUTO:					
Seen Test	3560	3445	3584	3464	14053
Unseen Test	3560	3443	3584	3465	14052
Dev	3534	3467	3560	3487	14048
Train	23502	23233	23776	23332	93843

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

Learning | Reasoning
Heterogeneous

[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

SpaRTUN & ReSQ

- Spatial Question Answering
- Spatial Role Labeling
  - Spatial concepts (landmarks, trajectors, …)
  - Spatial relationships (topological relation, directional relations, …)

Sets	FB	FR	YN	CO	Total
SPARTQA-HUMAN:					
Test	104	106	194	107	511
Train	154	149	162	151	616
SPARTQA-AUTO:					
Seen Test	3560	3445	3584	3464	14053
Unseen Test	3560	3443	3584	3465	14052
Dev	3534	3467	3560	3487	14048
Train	23502	23233	23776	23332	93843

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

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

Learning | Reasoning
Heterogeneous

[**R. Mirzaee**, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

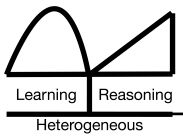
Spatial Reasoning over Text

SpartQA

#	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 Blocks (FB), Find Relations (FR), Choose Objects (CO), Yes/No (YN)

Concluding message



[R. Mirzaee, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Spatial Reasoning over Text

SpartQA

#	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

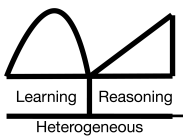
- Find Blocks (FB), Find Relations (FR), Choose Objects (CO), Yes/No (YN)

Concluding message

-Fine-tuning with SpartQA improved tests on different domains such as BoolQA, bAbI.

-Fine-tuning with the new extension SPARTUN, improved results on complex reasoning: StepGame.

-Fine-tuning with the new extension SPARTUN, improved results on realistic domain ResQ.



[R. Mirzaee, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Spatial Reasoning over Text

SpartQA

#	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 Blocks (FB), Find Relations (FR), Choose Objects (CO), Yes/No (YN)

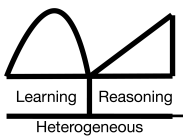
Concluding message

-Fine-tuning with SpartQA improved tests on different domains such as BoolQA, bAbI.

-Fine-tuning with the new extension SPARTUN, improved results on complex reasoning: StepGame.

-Fine-tuning with the new extension SPARTUN, improved results on realistic domain ResQ.

SpartQA and SPARTUN help transferring spatial reasoning knowledge.



[R. Mirzaee, P. Kordjamshidi. **Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning,** EMNLP-2022]

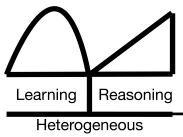
[**R. Mirzaee,** H. Faghihi, Q. Ning and P. Kordjamshidi, ***SpaRTQA: A Textual Question Answering Benchmark for Spatial Reasoning***, NAACL-2021.]

Spatial Reasoning over Text

- Related benchmarking and evaluation efforts:

StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts, Zhengxiang ShiZhengxiang ShiQiang ZhangAldo LipaniAldo Lipani, AAAI-2022.

Advancing Spatial Reasoning in Large Language Models: An In-depth Evaluation and Enhancement Using the StepGame Benchmark, Fangjun Li, David C. Hogg, Anthony G. Cohn, AAAI-24.



Integration of Spatial Logic in Training

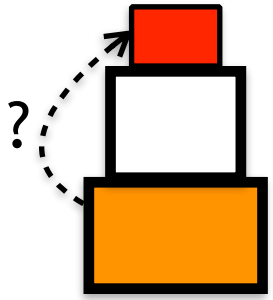
Context:
The white rectangle is on an orange rectangle. There is also the red rectangle, which is above the white rectangle.
Question:
Is the orange object below the red rectangle?

Initial facts: above(red-r, white-r), above(white-r, orange-r)

Rules:

- Converse
  - right(a, b) => left(b, a)
- Symmetric
  - near(b, c) => near(c, b)
- Transitive
  - above(x, b), above(b, c) => above(x, c)
  - -Transitive + topological
  - inside(x, y), inside(h, z), front(y, z) => front(x, z)

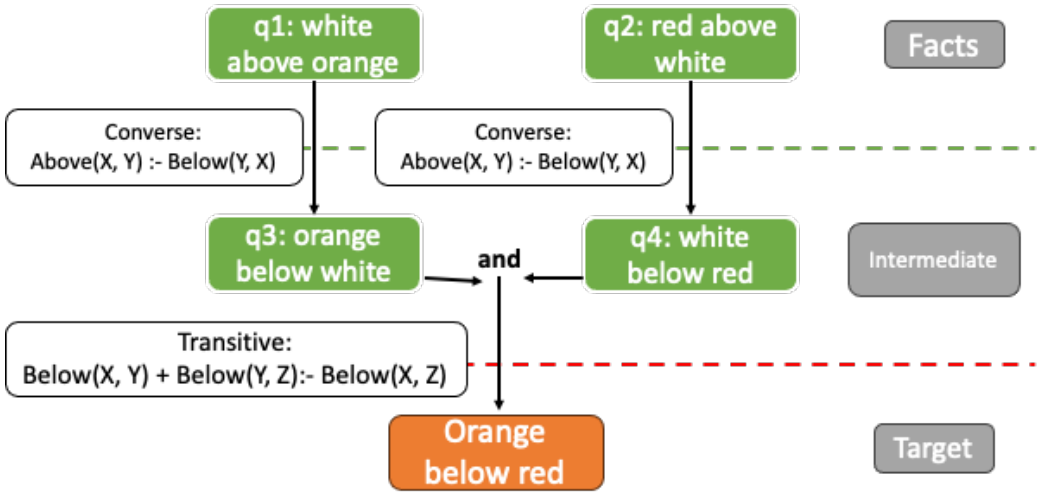
Target Query: below(orange-r, red-r)



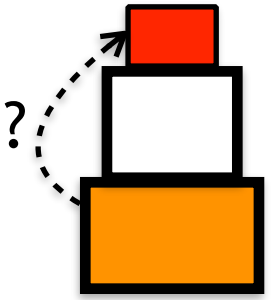
Learning | Reasoning
Heterogeneous

[**Tanawan Premsri,** P. Kordjamshidi, **Transferring Spatial Reasoning Knowledge by Neuro-symbolic Training, 2024,** under review.]

Integration of Spatial Logic in Training



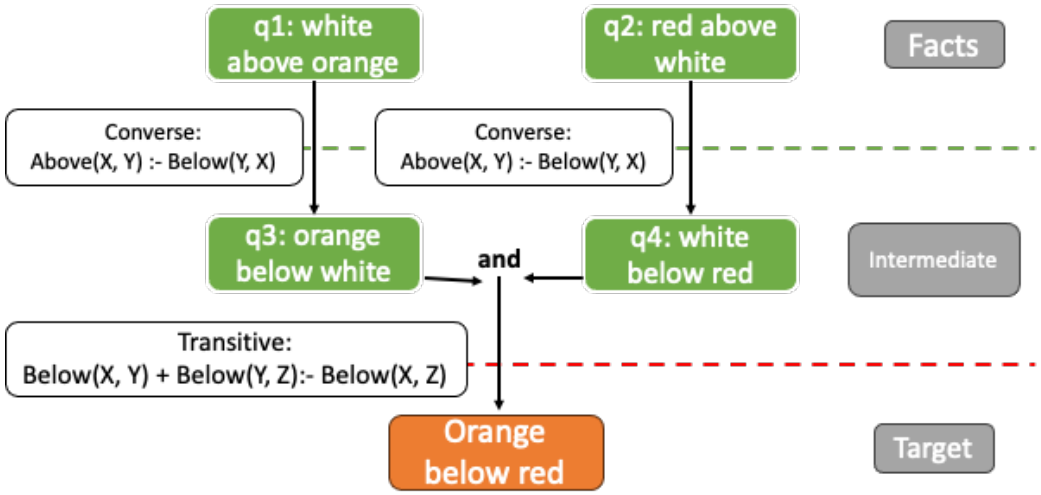
- Converse
  - right(a, b) => left(b, a)
- Symmetric
  - near(b, c) => near(c, b)
- Transitive
  - above(x, b), above(b, c) => above(x, c)
  - -Transitive + topological
  - inside(x, y), inside(h, z), front(y, z) => front(x, z)



Learning | Reasoning

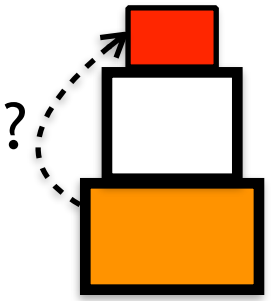
Heterogeneous

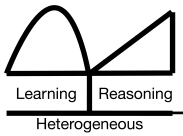
Integration of Spatial Logic in Training



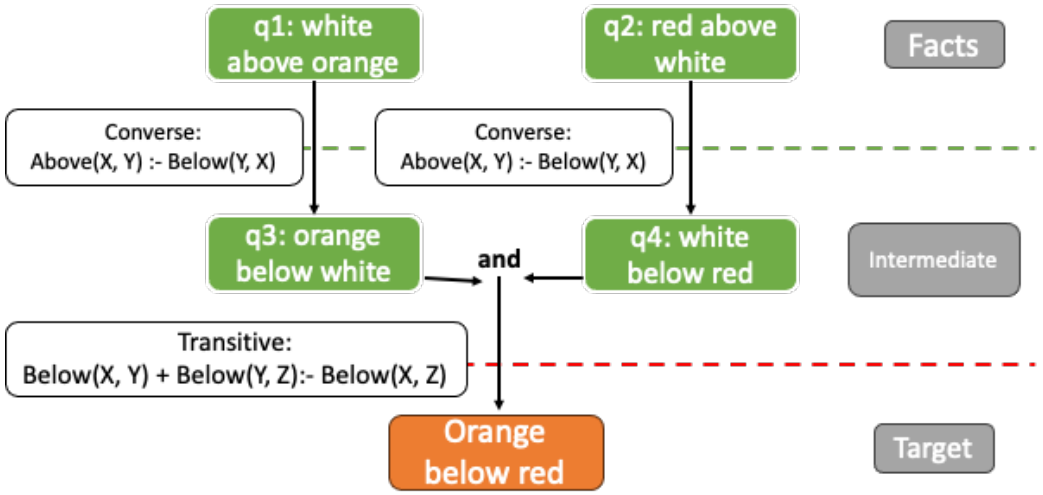
- Converse
  - right(a, b) => left(b, a)
- Symmetric
  - near(b, c) => near(c, b)
- Transitive
  - above(x, b), above(b, c) => above(x, c)
  - -Transitive + topological
  - inside(x, y), inside(h, z), front(y, z) => front(x, z)

- **Constraints for this example:**
  - **C1: (q1=> q3)**
  - **C2: (q2 => q4)**
  - **C3: (q3 and q4 => target)**





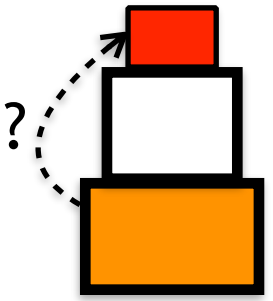
Integration of Spatial Logic in Training

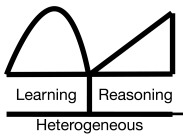


The intermediate annotations are not needed for inference, the entailment rules are translated to consistency constraints for every specific situation. The constraints are used only during training.

- Converse
  - right(a, b) => left(b, a)
- Symmetric
  - near(b, c) => near(c, b)
- Transitive
  - above(x, b), above(b, c) => above(x, c)
  - -Transitive + topological
  - inside(x, y), inside(h, z), front(y, z) => front(x, z)

- **Constraints for this example:**
  - **C1: (q1=> q3)**
  - **C2: (q2 => q4)**
  - **C3: (q3 and q4 => target)**





Integration of Spatial Logic in Training

Training

$$Loss = TaskLoss + \sum_{i=1}^{m} \lambda_j * C_i$$

Constraint Loss

- This training objective can be used for tuning both generative and encoder-based language models when adapted to the classification task.

Integration of Spatial Logic in Training

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 ; ChatGPT: No

Q: Are the **plants** **in** the **house**? Yes; ChatGPT: No

Overall performance on **Realistic ResQ** dataset (All). Performance per reasoning steps (k) required for the question to be answered. Null indicates the cases in which k was not clear.

Model	SPARTQA-Human	ResQ			
		k=1	k=2	unclassified	All
BERT	54.54	70.67	56.85	60.66	60.98
BERT-T	55.94	**76.00**	54.79	**61.18**	61.15
BERT-T+Q-Chain	**59.44**	72.00	**58.90**	59.90	**61.31**
T5	54.54	74.67	56.16	61.44	61.80
T5-T	49.65	81.33	54.79	61.44	62.30
T5-T+Q-Chain	**55.94**	**81.33**	**57.53**	**63.75**	**64.43**
GPT-3 (zero-shot)	58.04	74.67	60.95	66.58	66.22
GPT-3 (few-shot)	62.23	84.00	68.49	68.12	70.16
GPT-3 (COT)	65.73	**86.67**	67.12	68.64	70.49
GPT-4 (zero-shot)	**77.62**	84.00	**73.97**	**76.86**	**77.05**
Llama-3	61.54	80.00	64.38	67.35	68.20
Llama-3 (few)	62.94	82.67	69.86	**71.46**	**72.46**
Llama-3 (COT)	**67.83**	**82.76**	**76.03**	67.10	71.15

GPT3 (few-shot)	55.00	37.00	25.00	30.00	32.00	29.00	21.00	22.00	34.00	31.00
GPT3 (COT)	61.00	45.00	30.00	35.00	35.00	27.00	22.00	24.00	23.00	25.00

Learning | Reasoning
Heterogeneous

[**Tanawan Premsri,** P. Kordjamshidi, **Transferring Spatial Reasoning Knowledge by Neuro-symbolic Training, 2024,** under review.]

Integration of Spatial Logic in Training

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 ; ChatGPT: No

Q: Are the **plants in** the **house**? Yes; ChatGPT: No

Overall performance on **Realistic ResQ** dataset (All). Performance per reasoning steps (k) required for the question to be answered. Null indicates the cases in which k was not clear.

Model	SPARTQA-Human	ResQ			
		k=1	k=2	unclassified	All
BERT	54.54	70.67	56.85	60.66	60.98
BERT-T	55.94	**76.00**	54.79	**61.18**	61.15
BERT-T+Q-Chain	**59.44**	72.00	**58.90**	59.90	**61.31**
T5	54.54	74.67	56.16	61.44	61.80
T5-T	49.65	81.33	54.79	61.44	62.30
T5-T+Q-Chain	**55.94**	**81.33**	**57.53**	**63.75**	**64.43**
GPT-3 (zero-shot)	58.04	74.67	60.95	66.58	66.22
GPT-3 (few-shot)	62.23	84.00	68.49	68.12	70.16
GPT-3 (COT)	65.73	**86.67**	67.12	68.64	70.49
GPT-4 (zero-shot)	**77.62**	84.00	**73.97**	**76.86**	**77.05**
Llama-3	61.54	80.00	64.38	67.35	68.20
Llama-3 (few)	62.94	82.67	69.86	**71.46**	**72.46**
Llama-3 (COT)	**67.83**	**82.76**	**76.03**	67.10	71.15

Model	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
BERT	98.51	95.53	91.68	66.71	49.11	41.47	41.47	32.09	28.94	28.16
BERT-T	98.50	95.32	93.26	76.78	66.36	58.76	53.70	46.27	42.71	40.12
BERT-T+Q-Chain	98.70	96.45	93.03	74.58	64.95	59.04	54.38	49.23	45.36	44.05
GPT3 (few-shot)	55.00	37.00	25.00	30.00	32.00	29.00	21.00	22.00	34.00	31.00
GPT3 (COT)	61.00	45.00	30.00	35.00	35.00	27.00	22.00	24.00	23.00	25.00

Accuracy of various models on STEPGAME including result of GPT3 model reported in Zhun Y., et.al,. 2023. Coupling large language models with logic programming for robust and general reasoning from text.

[**Tanawan Premsri,** P. Kordjamshidi, **Transferring Spatial Reasoning Knowledge by Neuro-symbolic Training, 2024,** under review.]

Learning Reasoning Heterogeneous

Spatial QA

Current Results on text-based spatial QA

- Models developed that learned to obey spatial logic during training
- Models developed that do extraction and pass the formal representations to reasoning engine.
- Logic-based training helped complex multi-hop reasoning but GPT models are better in commonsense.
- Pipeline for symbolic reasoning less applicable to the realistic domains.

Spatial QA

Current Results on text-based spatial QA

- Models developed that learned to obey spatial logic during training
- Models developed that do extraction and pass the formal representations to reasoning engine.
- Logic-based training helped complex multi-hop reasoning but GPT models are better in commonsense.
- Pipeline for symbolic reasoning less applicable to the realistic domains.

Ongoing Next Questions…

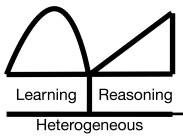
Spatial QA

Current Results on text-based spatial QA

- Models developed that learned to obey spatial logic during training
- Models developed that do extraction and pass the formal representations to reasoning engine.
- Logic-based training helped complex multi-hop reasoning but GPT models are better in commonsense.
- Pipeline for symbolic reasoning less applicable to the realistic domains.

Ongoing Next Questions…

- Looking into a variety of downstream tasks for spatial language understanding, in particular image generation with diffusion models and deep understanding of compositional spatial configurations, frame of references.



Spatial Reasoning on Vision and Language

- VAQ datasets until 5 years ago did not evaluate spatial understanding deeply
- Recent efforts for LLM evaluations for multimodal spatial reasoning are becoming more popular

Figure 1: Caption: *The potted plant is at the right side of the bench.* Label: `True`.

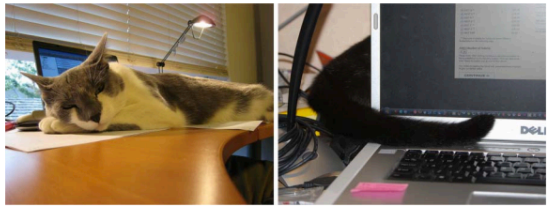
Figure 2: Caption: *The cow is ahead of the person.* Label: `False`.

- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Spatial Reasoning on Vision and Language

Category	Spatial Relations
Adjacency	Adjacent to, alongside, at the side of, at the right side of, at the left side of, attached to, at the back of, ahead of, against, at the edge of
Directional	Off, past, toward, down, deep down*, up*, away from, along, around, from*, into, to*, across, across from, through, down from
Orientation	Facing, facing away from, parallel to, perpendicular to
Projective	On top of, beneath, beside, behind, left of, right of, under, in front of, below, above, over, in the middle of
Proximity	By, close to, near, far from, far away from
Topological	Connected to, detached from, has as a part, part of, contains, within, at, on, in, with, surrounding, among, consists of, out of, between, inside, outside, touching
Unallocated	Beyond, next to, opposite to, after*, among, enclosed by

Table 1: The 71 available spatial relations; 66 of them appear in our final dataset (∗ indicates not used).



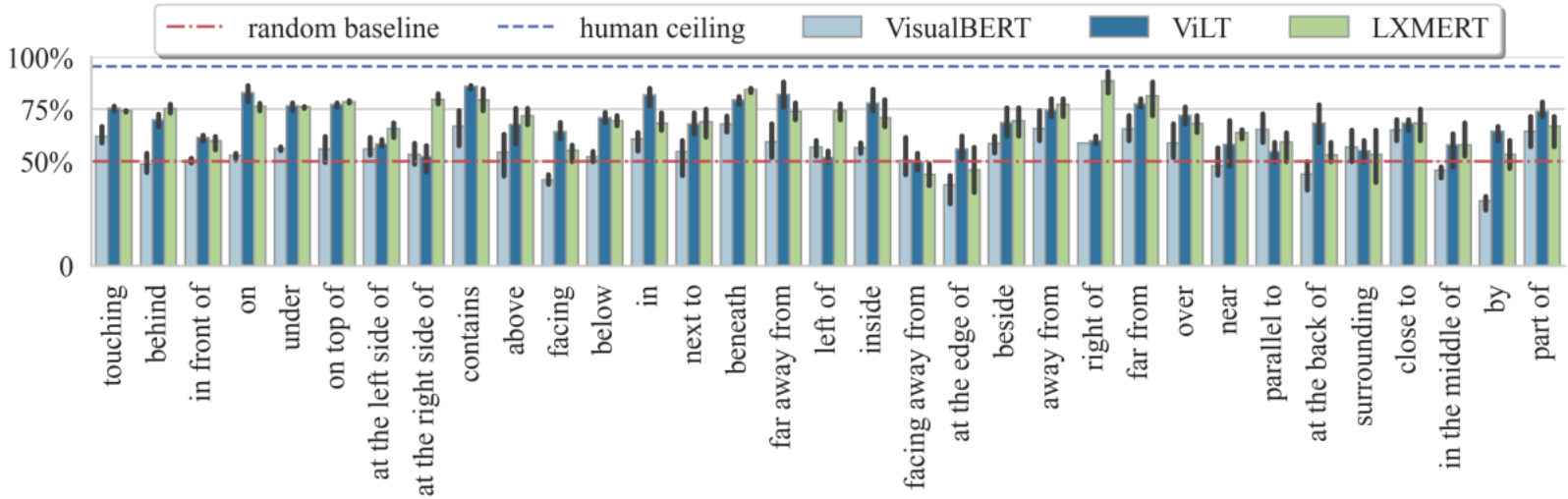
Caption: *The cat (is) ___ the laptop.*

Choose a relation to complete the caption:
☐ *above* ☐ *beneath* ☑ *behind* .....

The completed caption is true for:
☐ image 1 ☑ image 2

- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Spatial Reasoning on Vision and Language



- Results of multiple pre-trained VLMs for accuracy on spatial relations (random splits).
- There is still a large gap to achieve human level spatial reasoning

- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

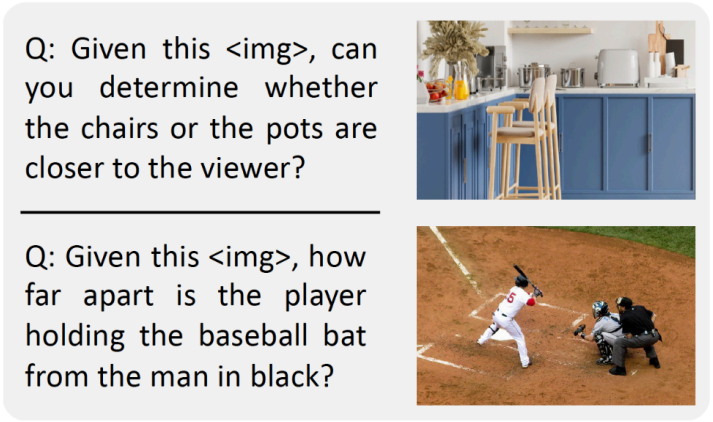
Spatial Reasoning on Vision and Language

Why such a gap? What is the Challenge?

- One hypothesis: Pre-trained model did not observe enough variations of spatial information compared to other types of information.

SpatialVLM

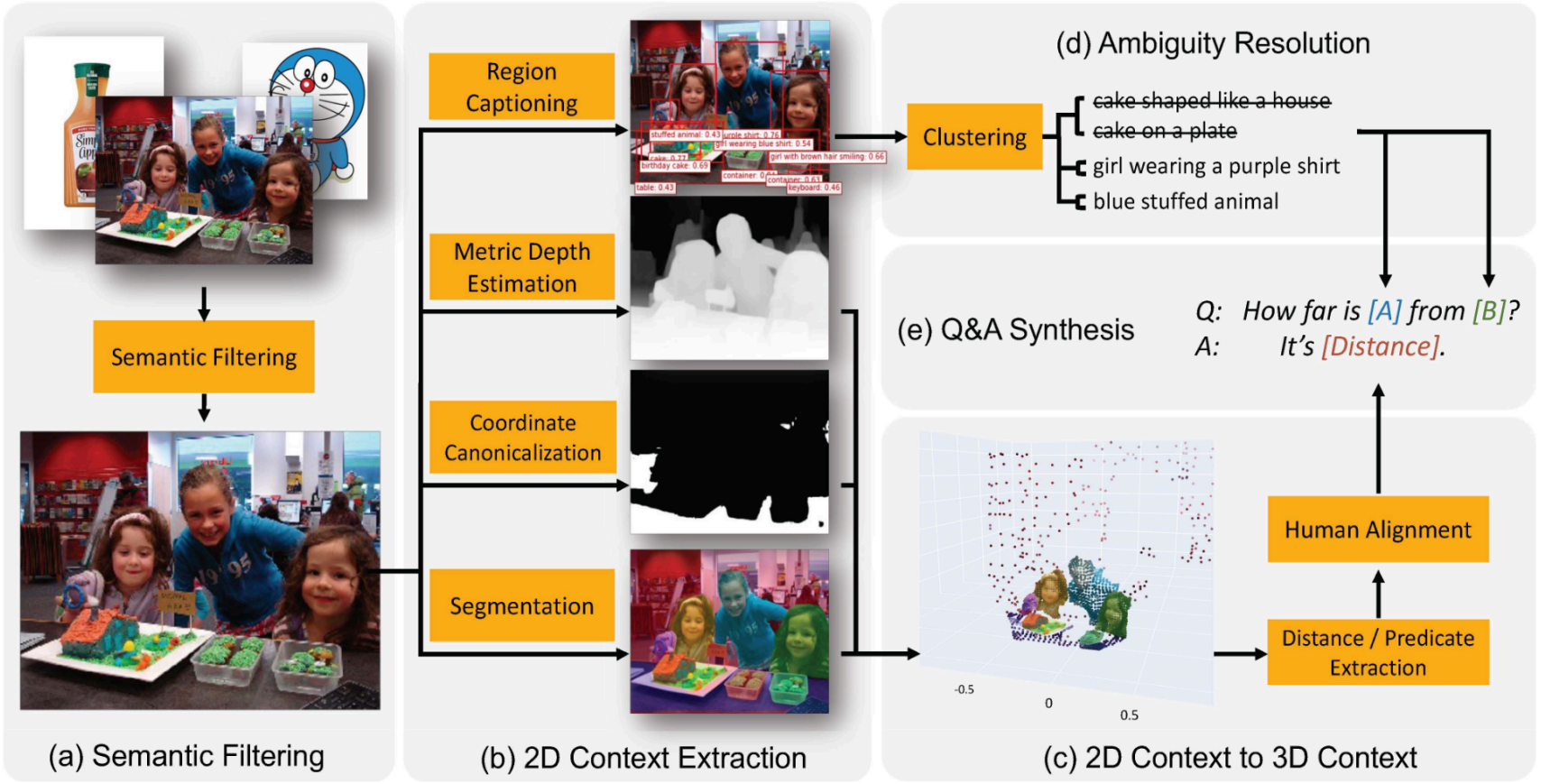
- Problem: The main challenge is lack of training data at large scale



- Solution: create internet-scale synthetic data!!

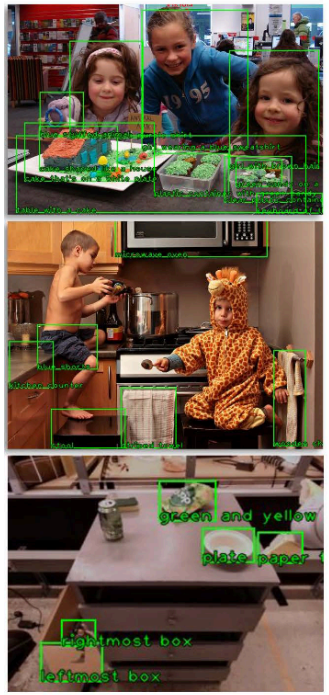
Spatial VLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities, Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, Fei Xia, CVPR 2024

SpatialVLM



Automatic 3D spatial VQA data generation framework that lifts 2D images into metric scale 3D point clouds. Scales the data pipeline up to 2 billion VQA examples on 10 million real-world images.

SpatialVLM

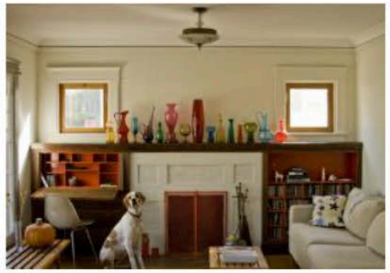


Qualitative Q & A	
Find out if the girl wearing a purple shirt is situated behind or in front of the girl wearing a blue sweatshirt.	It's the woman wearing white shirt.
Does the stool lie in front of the microwave oven?	Yes, the stool is in front of the microwave oven.
Is the plate positioned on the left or right side of the green and yellow napkin?	The plate is to the right.

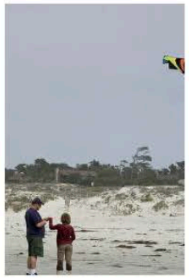
Quantitative Q & A	
Measure the distance from the cake shaped like a house to the girl wearing a purple shirt?	Around half a meter
Can you provide the distance measurement between the microwave oven and the kitchen counter?	They are 1 meter apart.
How wide is the plate?	20 centimeters

Spatial VLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities, Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, Fei Xia, CVPR 2024

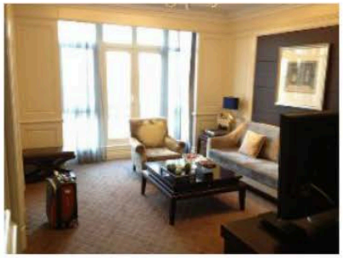
SpatialVLM



Q: Is the fireplace screen with red doors smaller than the dog standing on the floor in width?
A: no



Q: Compared to the little boy in a red shirt, which side is the man wearing a blue shirt on?
A: left

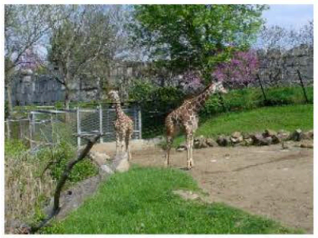


Q: Are the windows positioned to the left of the black television?
A: yes

Q: Is the palm tree in distance taller than the parked white car?
A: yes

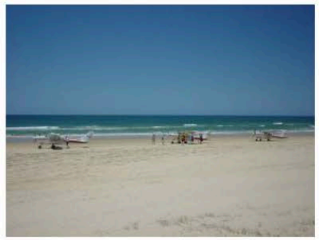
Method	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
Accuracy	68.0%	71.3%	60.4%	60.7%	50.2%	50.4%	**75.2%**

Q: Could you provide the distance between the sign and the motorcyclist?
A: about 0.5 meter



Q: Determine the distance of the fence from the giraffes in a zoo relative to the camera.
A: about 5 meters

Q: How far is the striped tie towards the left from the black cell phone?
A: about 0.2 meter



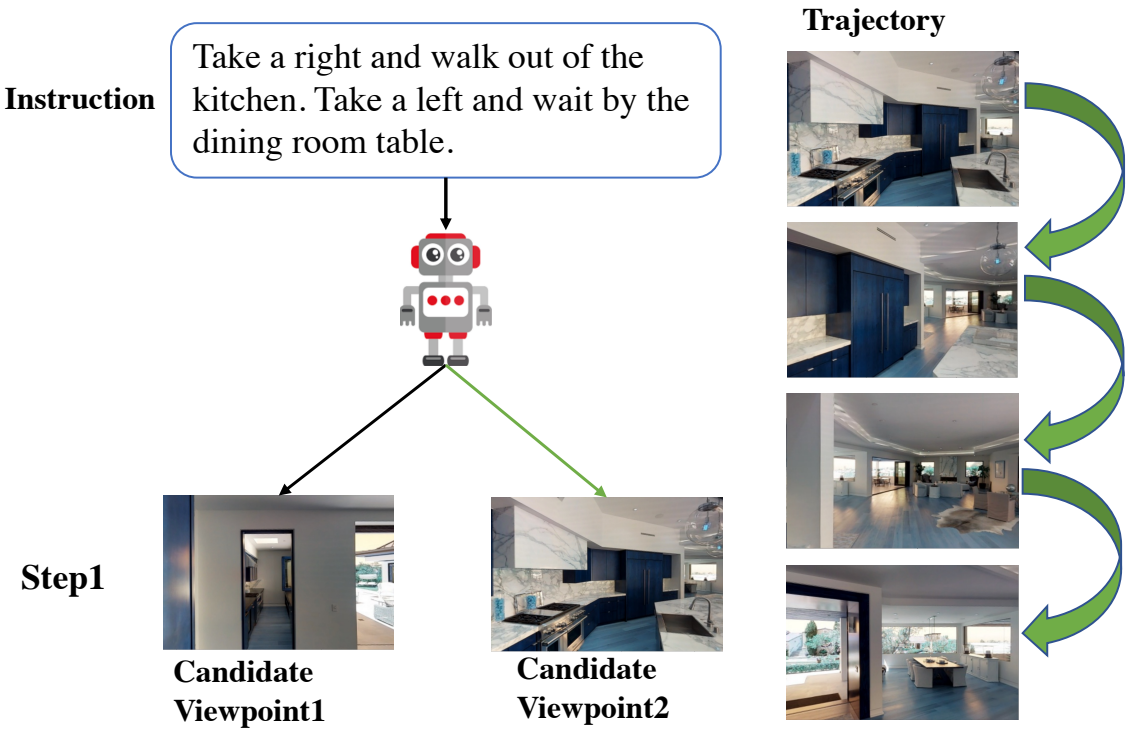
Q: How much distance is there between the sand and the people that are standing on the beach?
A: 0, as the people are directly standing on the sand

	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
number%	1.0%	20.9%	26.0%	52.0%	83.2%	88.8%	**99.0%**
In range [50, 200]%	0.0%	13.0%	7.9%	5.3%	23.7%	33.9%	**37.2%**

Back to the Outline

- **Spatial Annotation Schemes**
- **Spatial Information Extraction**
- **Spatial Reasoning over Text: Spatial QA**
- **Spatial Reasoning over Vision and Language**
- **Downstream Application: Navigation**

Vision and Language Navigation

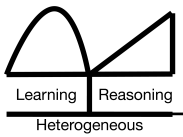


VLN-Trans: Translator for the Vision and Language Navigation Agent. Yu Zhang, Parisa Kordjamshidi. (ACL- 2023)

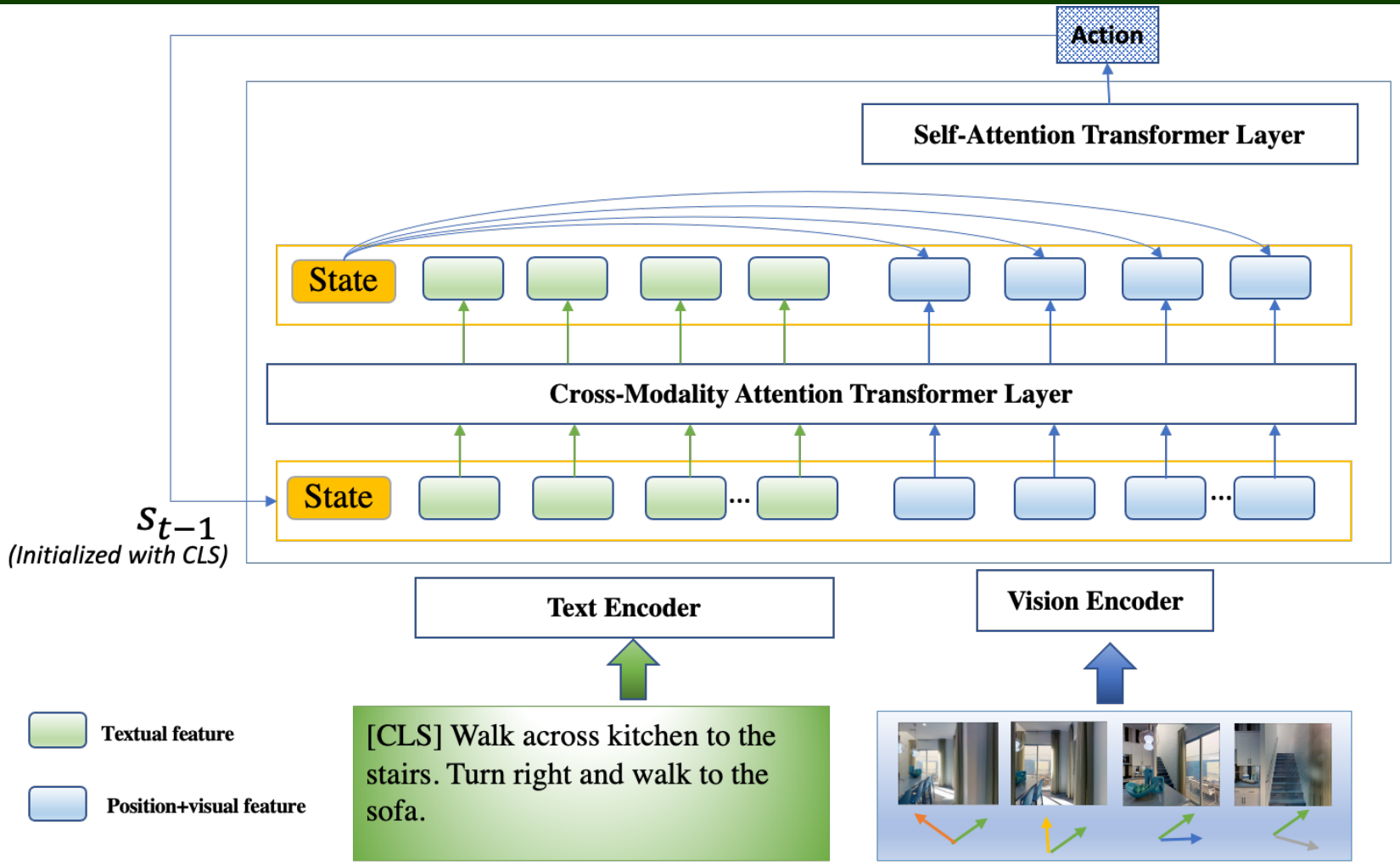
LOViS: Learning Orientation and Visual Signals for Vision and Language Navigation. Yue Zhang, Parisa Kordjamshidi. (COLING-2022)

Explicit Object Relation Alignment for Vision and Language Navigation. Yue Zhang, Parisa Kordjamshidi. (ACL SRW 2022)

Towards Navigation by Reasoning over Spatial Configurations. Yue Zhang, Quan Guo, Parisa Kordjamshidi. (ACL-2021 workshop on SpLU-RoboNLP)

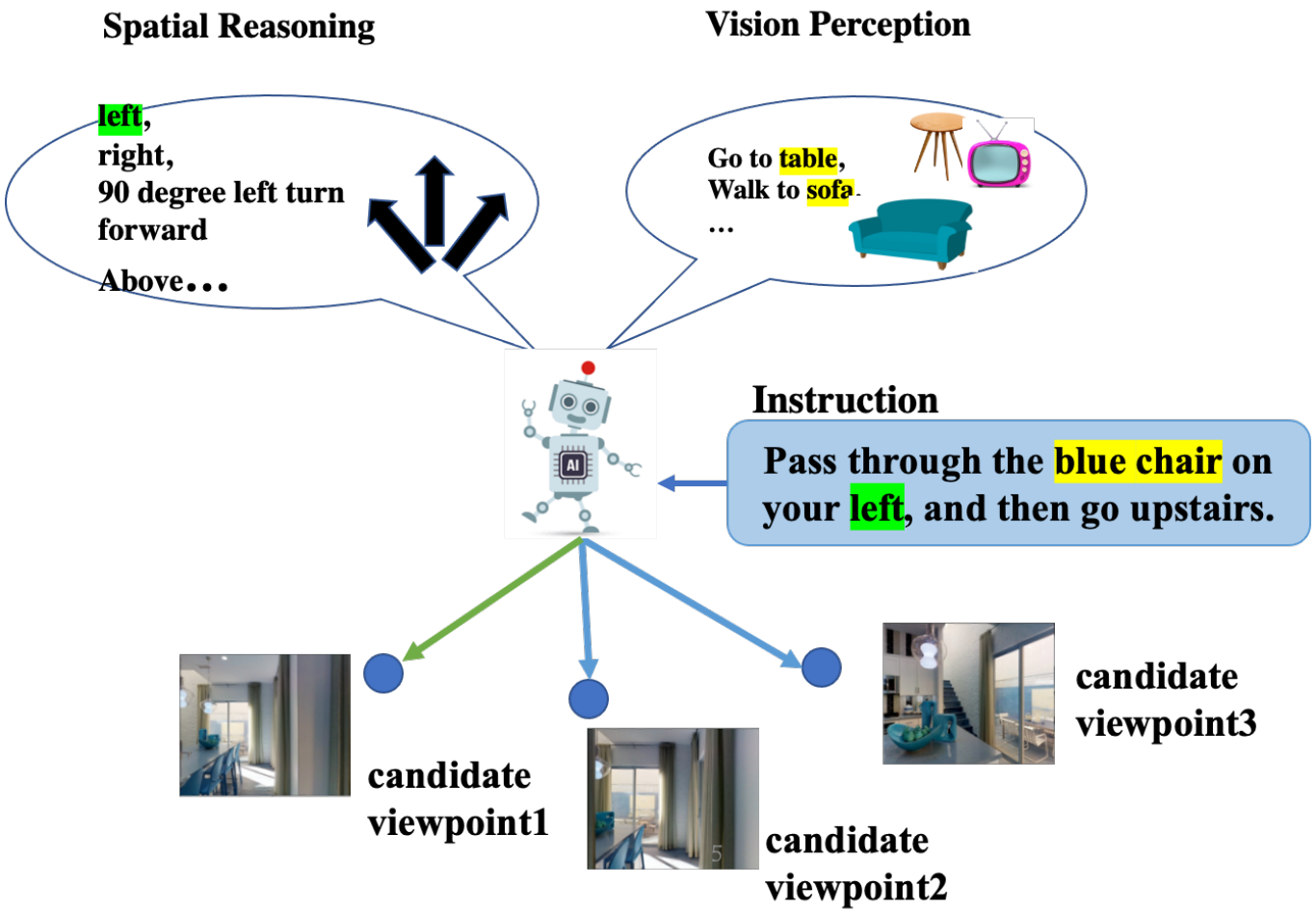


Transformer-based Baseline: Recurrent VLN



A Recurrent Vision-and-Language BERT for Navigation, Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, Stephen Gould, 2020.

Reasoning ability of VLN

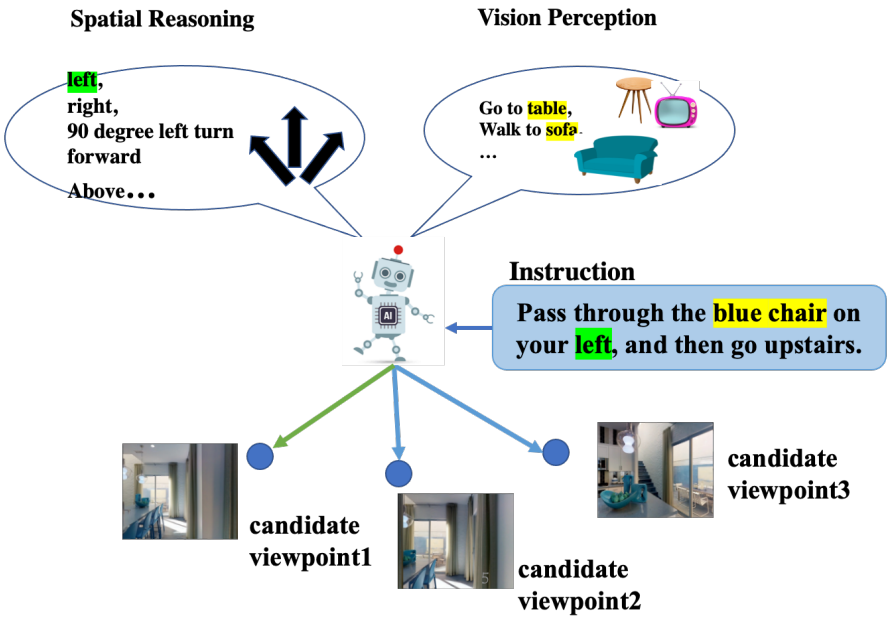


LovIS model

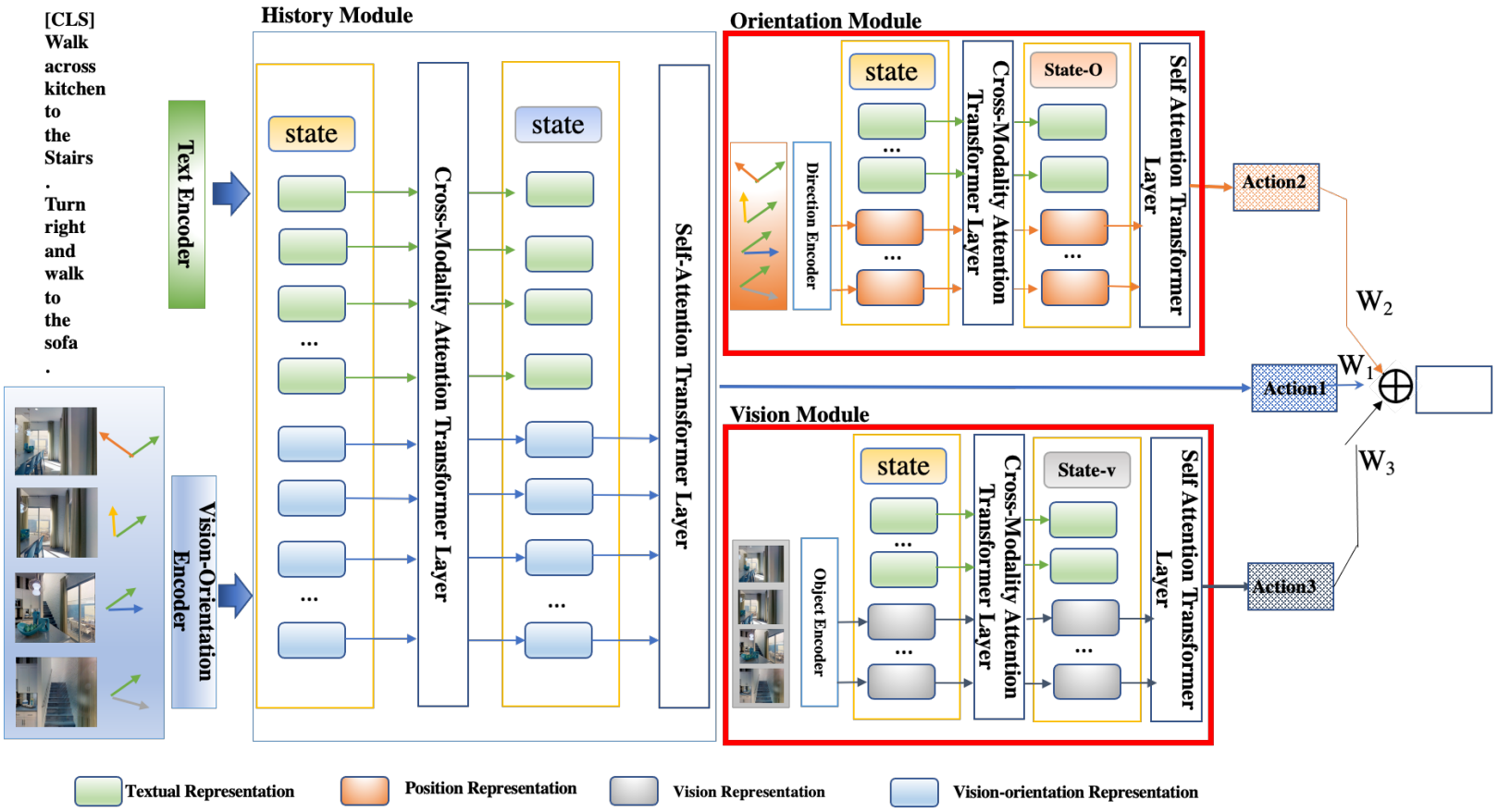
- **Transformer-based navigation agent**

- **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

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
  - Navigation Error [NE], Success rate [SR], Success rate normalized and weighted by length

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**

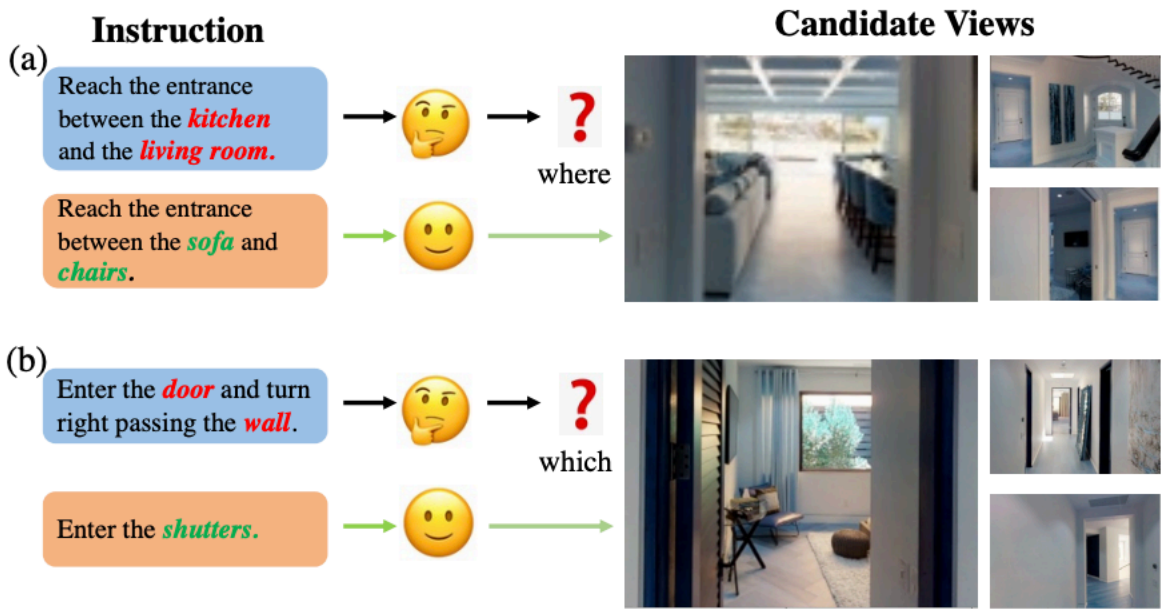
LOViS: Learning Orientation and Visual Signals for Vision and Language Navigation, **Yue Zhang** and Parisa Kordjamshidi. (COLING-2022)

Further Challenges: Ambiguity in Instructions

- Distinctive and distinguishable landmarks and towards an interactive setting.

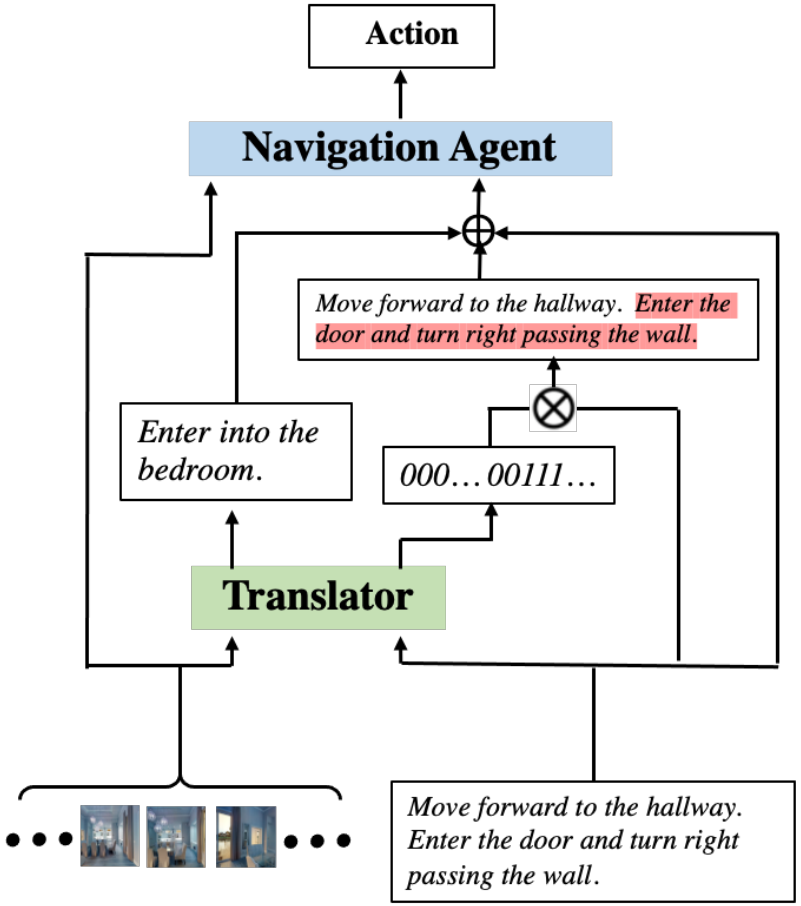
Unrecognizable

Non-distinctive

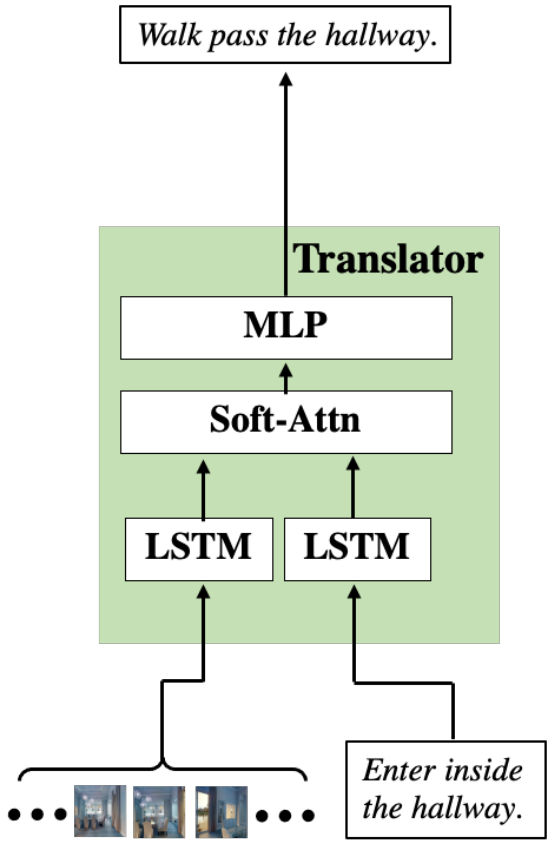


VLN-Trans: Translator for the Vision and Language Navigation Agent, Yue Zhang, Kordjamshidi, ACL-2023.

VLN-Trans



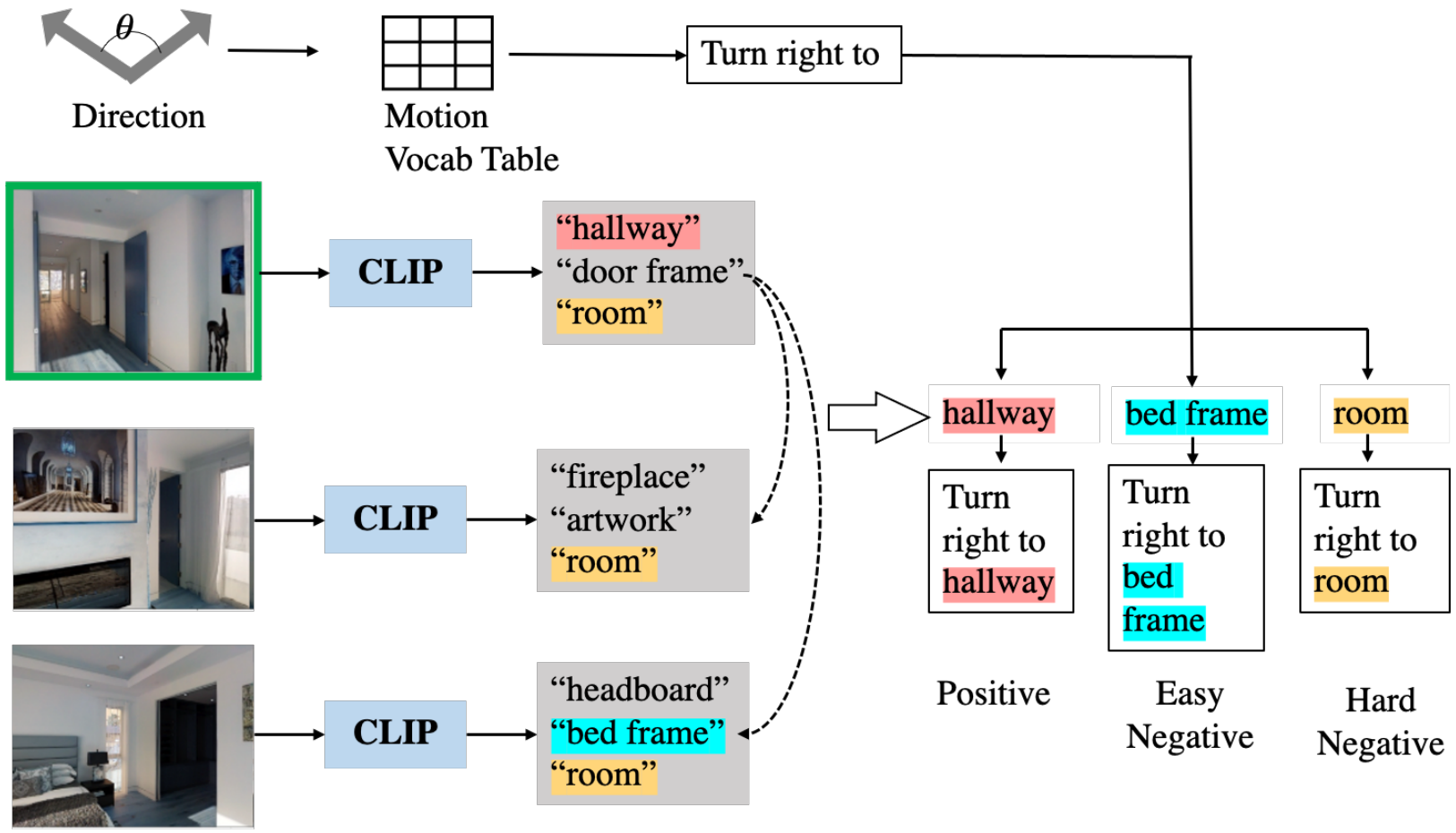
VLN-Trans



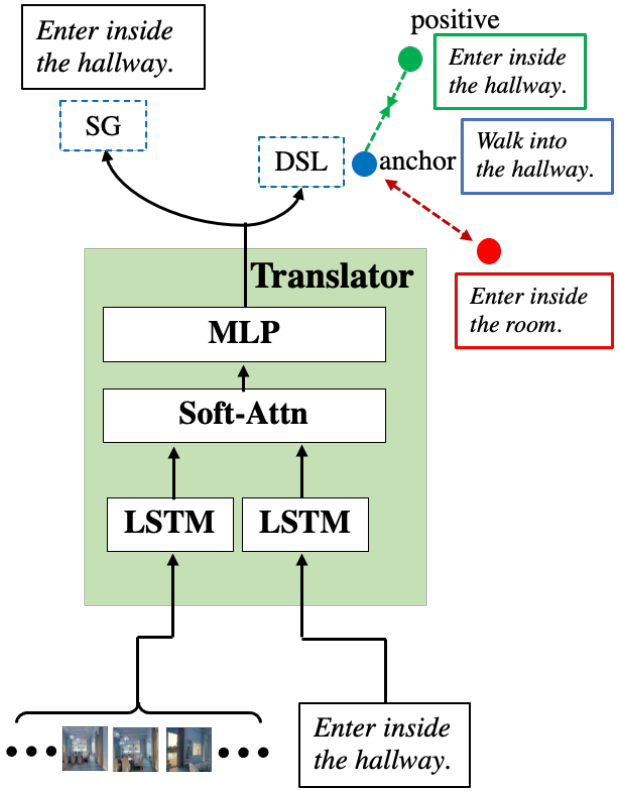
Translator Architecture

Learning | Reasoning
Heterogeneous

Translator Pre-training: Synthetic Sub-instruction Dataset



Translator Pre-traninig Task



Translator Pre-training

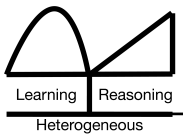
Sub-Instruction Generation (SG):
Train the translator to generate a sub-instruction, given the positive sub-instructions paired with the viewpoints.

Distinctive Sub-instruction Learning (DSL):
To encourage the translator to learn sub-instruction representations that are close to the positive sub-instructions with recognizable and distinctive landmarks, and are far from the negative sub-instructions.

Experimental Results

		Val seen			Val Unseen			Test Unseen		
	Method	NE ↓	SR ↑	SPL↑	NE ↓	SR ↑	SPL↑	NE ↓	SR ↑	SPL ↑
1	Env-Drop (Tan et al., 2019)	3.99	0.62	0.59	5.22	0.47	0.43	5.23	0.51	0.47
2	RelGraph (Hong et al., 2020a)	3.47	0.67	0.65	4.73	0.57	0.53	4.75	0.55	0.52
3	NvEM (An et al., 2021)	3.44	0.69	0.65	4.27	0.60	0.55	4.37	0.58	0.54
4	PREVALENT (Hao et al., 2020)	3.67	0.69	0.65	4.71	0.58	0.53	5.30	0.54	0.51
5	HAMT (ResNet) (Chen et al., 2021)	–	0.69	0.65	–	0.64	0.58	–	–	–
6	HAMT (ViT) (Chen et al., 2021)	2.51	0.76	0.72	–	0.66	0.61	**3.93**	0.65	**0.60**
7	CITL (Liang et al., 2022)	2.65	0.75	0.70	3.87	0.63	0.58	3.94	0.64	0.59
8	ADAPT (Lin et al., 2022)	2.70	0.74	0.69	3.66	0.66	0.59	4.11	0.63	0.57
9	LOViS (Zhang and Kordjamshidi, 2022b)	**2.40**	0.77	0.72	3.71	0.65	0.59	4.07	0.63	0.58
10	VLN⟳BERT (Hong et al., 2021)	2.90	0.72	0.68	3.93	0.63	0.57	4.09	0.63	0.57
11	VLN⟳BERT^{+} *(ours)*	2.72	0.75	0.70	3.65	0.65	0.60	4.09	0.63	0.57
12	VLN⟳BERT^{++} *(ours)*	2.51	0.77	0.72	3.40	0.67	0.61	4.02	0.63	0.58
13	VLN-Trans-R2R *(ours)*	**2.40**	**0.78**	**0.73**	3.37	0.67	**0.63**	3.94	0.65	0.59
14	VLN-Trans-FG-R2R *(ours)*	2.45	0.77	0.72	**3.34**	**0.69**	**0.63**	3.94	**0.66**	**0.60**

Table 1: Experimental results on R2R Benchmarks in a single-run setting. The best results are in bold font. + means we add RXR (Ku et al., 2020) and Marky-mT5 dataset (Wang et al., 2022b) as the extra data to pre-train the navigation agent. ++ means we further add SyFiS dataset to pre-train the navigation agent. ViT means Vision Transformer representations.



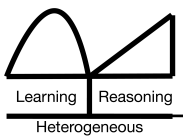
Experimental Results

		Val Seen					Val Unseen				
	Method	NE↑	SR↑	SPL↑	CLS↑	sDTW↑	NE↓	SR↑	SPL↑	CLS↑	sDTW↑
1	OAAM (Qi et al., 2020a)	-	0.56	0.49	0.54	-	0.32	0.29	0.18	0.34	0.11
2	RelGraph (Hong et al., 2020a)	5.14	0.55	0.50	0.51	0.35	7.55	0.35	0.25	0.37	0.18
3	NvEM (An et al., 2021)	5.38	0.54	0.47	0.51	0.35	6.80	0.38	0.28	0.41	0.20
4	VLN↻BERT* (Hong et al., 2021)	4.82	0.56	0.46	0.56	0.38	6.48	0.43	0.32	0.42	0.21
5	CITL (Liang et al., 2022)	**3.48**	0.67	0.57	0.56	**0.43**	6.42	0.44	0.35	0.39	0.23
6	LOViS (Zhang and Kordjamshidi, 2022b)	4.16	**0.67**	0.58	**0.58**	**0.43**	6.07	0.45	0.35	**0.45**	0.23
7	VLN-Trans	3.79	**0.67**	**0.59**	0.57	**0.43**	**5.87**	**0.46**	**0.36**	**0.45**	**0.25**

Table 2: Experimental results on R4R dataset in a single-run setting. * denotes our reproduced R4R results.

	Val Seen		Val Unseen	
Method	SR↑	SPL↑	SR↑	SPL↑
EnvDrop (Tan et al., 2019)	0.43	0.38	0.34	0.28
VLN↻BERT (Hong et al., 2020a)	0.50	0.46	0.42	0.37
HAMT (Chen et al., 2021)	0.53	0.50	0.45	0.41
VLN-Trans	**0.58**	**0.53**	**0.50**	**0.45**

Table 3: Experimental results on the R2R-Last dataset.



NavHint: Hint Generator

Instruction Turn around and go straight. Walk towards the wall and stop.

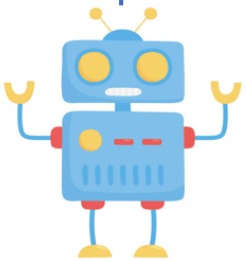
Candidate Viewpoints

view1

view2

view3 (target)

Hint Generator



Action Selection

Sub-Instruction

"**Walk towards the wall**" need to be executed.

Landmark Ambiguity

But I can see "**wall**" in all candidate viewpoints.

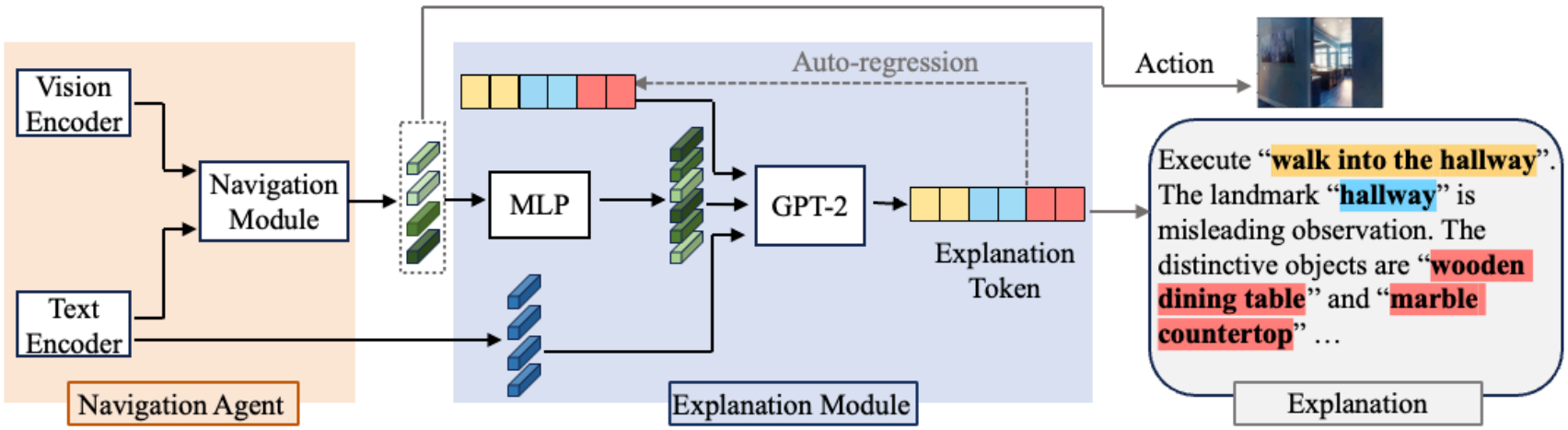
Target Distinctive Objects

However, there are "**large window with wooden blinds, glass table with white chairs, and a ceiling lamp**" that are specific to view3.

Learning | Reasoning
Heterogeneous

-Yue Zhang, Quan Guo, Parisa Kordjamshidi, **NavHint: Vision and Language Navigation Agent with a Hint Generator,** EACL-2024 Findings.

Model Architecture



Learning | Reasoning
Heterogeneous

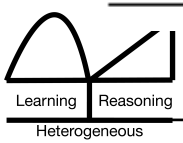
Experimental Results

Results on R2R

	Method	Validation Unseen					Test Unseen		
		NE ↓	SR ↑	SPL↑	sDTW↑	nDTW↑	NE ↓	SR ↑	SPL ↑
1	Seq-to-Seq (Anderson et al., 2018)	7.81	0.22	–	–	–	7.85	0.20	0.18
2	Speaker-follower (Fried et al., 2018)	6.62	0.36	–	–	–	6.62	0.35	0.28
3	Self-Monitor (Ma et al., 2019)	5.52	0.45	0.32	–	–	5.67	0.48	0.35
4	VLN◯BERT (Hong et al., 2020c)	3.93	0.63	0.57	–	–	4.09	0.63	0.57
5	HAMT (ViT) (Chen et al., 2021)	3.97	0.66	0.61	–	–	3.93	0.65	**0.60**
6	LANA (Wang et al., 2023)	–	0.68	0.62	–	–	–	0.65	**0.60**
7	VLN-SIG (ViT) (Li and Bansal, 2023)	3.37	0.68	0.62	0.59	0.70	–	0.65	**0.60**
8	VLN-trans (Zhang and Kordjamshidi, 2023)	3.34	**0.69**	0.63	0.60	0.70	3.94	**0.66**	**0.60**
9	EDrop* (Tan et al., 2019)	5.49	0.55	0.47	0.42	0.58	5.60	0.51	0.49
10	EDrop + Exp. (ours)	5.44	0.55	0.47	0.44	0.60	5.47	0.53	0.49
11	VLN◯BERT^{++} (Zhang and Kordjamshidi, 2023)	3.40	0.67	0.61	0.58	0.69	4.02	0.63	0.58
12	VLN◯BERT^{++} + Exp. (ours)	**3.23**	**0.69**	**0.65**	**0.61**	**0.72**	4.00	0.65	**0.60**

Results on R4R

	Method	NE↓	SR↑	SPL↑	CLS↑	sDTW↑
1	OAAM (Qi et al., 2020)	13.80	0.29	0.18	0.34	0.11
2	RelGraph (Hong et al., 2020a)	7.55	0.35	0.25	0.37	0.18
3	NvEM (An et al., 2021)	6.80	0.38	0.28	0.41	0.20
4	VLN◯BERT (Hong et al., 2020c)	6.48	0.43	0.32	0.42	0.21
5	CITL (Liang et al., 2022)	6.42	0.44	0.35	0.39	0.23
6	VLN-Trans (Zhang and Kordjamshidi, 2023)	**5.87**	**0.46**	**0.36**	**0.45**	**0.25**
7	VLN◯BERT^{++} (Zhang and Kordjamshidi, 2023)	6.33	0.44	0.34	0.43	0.23
8	VLN◯BERT^{++} + Exp. (ours)	6.04	**0.46**	**0.36**	**0.45**	**0.25**



-Yue Zhang, Quan Guo, Parisa Kordjamshidi, **NavHint: Vision and Language Navigation Agent with a Hint Generator,** EACL-2024 Findings.

Summary of Vision and Language Research

- **Modulating the orientation and visual capabilities**
- **Dealing with ambiguity of instructions**
- **Generating global and detailed explanations of views**

improves with interpretability and accuracy.

-Yue Zhang, Quan Guo, Parisa Kordjamshidi, **NavHint: Vision and Language Navigation Agent with a Hint Generator,** EACL-2024 Findings.

-Yue Zhang, Parisa Kordjamshidi, **VLN-Trans: Translator for the Vision and Language Navigation Agent**. (ACL- 2023)

-Yue Zhang, Parisa Kordjamshidi, **LOViS: Learning Orientation and Visual Signals for Vision and Language Navigation**. (COLING-2022)

-Yue Zhang, Parisa Kordjamshidi, **Explicit Object Relation Alignment for Vision and Language Navigation**. (ACL SRW 2022)

-Yue Zhang, Quan Guo and Parisa Kordjamshidi, **Towards Navigation by Reasoning over Spatial Configurations**. (ACL-2021 workshop on SpLU-RoboNLP)