Spatial and Temporal Language Understanding: Representation, Reasoning, and Grounding

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Qiang Ning
AWS
qiangning.01@gmail.com

James Pustejovsky
Brandeis University
jamesp@brandeis.edu

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

1 Description

This tutorial provides an overview of the *cutting edge* research on *spatial and temporal language understanding*. We also cover some essential background material from various subdisciplines to this topic, which we believe will enrich the CL community's appreciation of the complexity of spatiotemporal reasoning.

One of the essential functions of natural language is to express spatial and temporal relationships between objects and events. Linguistic constructs can encode highly complex, relational structures of objects, events, and spatiotemporal relations between them. Spatiotemporal language understanding is useful in many research areas and real-world applications. Extending two past tutorials on spatial language in EMNLP-2020 and COLING-2022, we propose this new tutorial that jointly discusses both spatial and temporal semantics for the first time; we also want to take this opportunity to showcase the challenges we still face today in spatiotemporal reasoning, even with state-of-the-art large language models.

This topic recently has attracted the attention of various sub-communities in the intersection of Natural Language, Computer Vision, and Robotics. The complexity of spatiotemporal language understanding and its importance in downstream tasks that involve grounding the language in the physical world has become evident to the NLP research community. The recent evaluation results on large generative language models such as ChatGPT show these models struggle with spatial and temporal reasoning while comparatively spatial reasoning appeared harder than temporal reasoning for these models (Bang et al., 2023).

While these two aspects of semantics are highly related, there are rare efforts with a focus on a combination of these two semantic aspects. We hope such a tutorial makes the connections more explicit and inspires new ideas for future research in the intersection of spatial and temporal semantic understanding in language and when language is combined with vision and action.

Similar to various aspects of symbolic semantic representations of language, standardizing tasks related to spatiotemporal language is challenging. It has been rather hard to obtain a set of concepts and relationships together with a formal meaning representation that applies to all real-world situations (Pustejovsky et al., 2003a,b; Pustejovsky, 2017; Pustejovsky et al., 2011; Kordjamshidi et al., 2010; Mani, 2009; Dan et al., 2020; Chambers et al., 2014; Ning et al., 2018a, 2020).

This has resulted in research on spatiotemporal language learning and reasoning becoming diverse, task-specific, and, to some extent, not comparable. While formal meaning representation is a general issue for language understanding, formalizing spatiotemporal concepts and building formal reasoning and machine learning models based on these concepts have a wealth of prior foundational work that can be exploited and linked to language understanding.

In this tutorial, we overview five main themes: **1) Spatiotemporal Semantic Representation; 2) Spatiotemporal Information Extraction and; 3) Spatiotemporal qualitative representation and reasoning; 4) Reasoning over spatial and temporal information with pre-trained and large generative language models; 5) Downstream applications that require Spatiotemporal reasoning including language grounding, robotics, navigation, dialogue systems and other tasks that require combining vision and language.** These are detailed in three categories in the detailed outline provided in a later section.

We cover the research on using spatial concepts for language grounding using spatial commonsense about object affordances (Pustejovsky and Krishnaswamy, 2021; Krajovic et al., 2020), composi-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), pages 39–46
June 16-21, 2024 ©2024 Association for Computational Linguistics

tional referring expressions, and robotic navigation (Francis et al., 2021; Mogadala et al., 2021).

The semantic representation section covers the research that attempted to arrive at a common set of basic concepts and relationships (Pustejovsky et al., 2003a; Bateman, 2010; Hois and Kutz, 2011) as well as making existing corpora interoperable (Pustejovsky et al., 2011; Mani and Pustejovsky, 2012; Kordjamshidi et al., 2010, 2017; Ning et al., 2018a, 2020). We discuss the existing qualitative and quantitative representation and reasoning models that can be used for the investigation of interoperability of machine learning and reasoning over spatial and temporal semantics (Cohn et al., 1997; Allen, 1984). Spatiotemporal language meaning representation includes research on cognitive and linguistically motivated semantic representations, knowledge representation and ontologies, qualitative and quantitative representation models used for formal meaning representation, and various annotation schemas and efforts for creating specialized corpora. We discuss various datasets that either focus on spatiotemporal annotations or downstream tasks that need spatial and temporal language learning and reasoning. Particularly, natural language visual reasoning data (Suhr et al., 2017, 2018) and question-answering data (Ning et al., 2020; Han et al., 2021). Moreover, we highlight the lack of research on learning representations that are spatiotemporally rich and point to a few sparse works in this area. We refer to meaning representations and foundation models currently being developed when processing video data which might be inspiring (Villegas et al., 2022; Fei et al., 2023; Ning et al., 2022; Bagad et al., 2023).

We overview the existing models for extraction of spatial and temporal information from language, both the abstract semantic extraction (Kordjamshidi et al., 2011; Kordjamshidi and Moens, 2015; Ning et al., 2018b; Leeuwenberg and Moens, 2018, 2020) and extractions driven by various target tasks and applications. We will discuss the recent datasets and results that are probing language models' ability in spatial language understanding using spatial question answering, visual questions answering (Mirzaee et al., 2021; Collell et al., 2021; Mirzaee and Kordjamshidi, 2022; Bang et al., 2023; Shi et al., 2022; Chen et al., 2024; Liu et al., 2023) and also in the recent diffusion models (Cho et al., 2023).

Finally, we overview the usage of spatiotemporal semantics by various downstream tasks and killer applications including language grounding (Alikhani and Stone, 2020), navigation (Zhang and Kordjamshidi; Zhang et al., 2024), self-driving cars (Deruyttere et al., 2021; Grujicic et al., 2022) robotics (Tellex et al., 2011; Kollar et al., 2010; Zheng et al., 2021), dialogue systems (Degand and Muller, 2020; Li et al., 2023) and human-machine interaction, and geographical information systems and knowledge graphs (Stock et al., 2013; Mai et al., 2020).

Spatiotemopral semantics is very closely connected and relevant to the visualization of natural language and grounding language into perception, central to dealing with configurations in the physical world and motivating a combination of vision and language for a richer understanding of time and space. The related tasks include text-to-scene, text-to-video, conversion; image captioning; spatial and visual (image/video) question answering; and spatial understanding in multimodal settings (Rahgooy et al., 2018) for robotics and navigation tasks and language grounding (Thomason et al., 2018; Pustejovsky and Krishnaswamy, 2021).

The current research using end-to-end monolithic deep models fails to solve complex tasks that need deep language understanding and reasoning capabilities (Hudson and Manning, 2019). Throughout this tutorial, we will highlight the importance of combining learning and reasoning for spatiotemporal language understanding and its influence on the semantic representation and type of the learning models as well as the performance on various applications. Regarding the question of reasoning, we (a) point out the role of qualitative and quantitative formal representations in helping spatiotemporal reasoning based on natural language and the possibility of learning such representations from data to support compositionality and inference (Hudson and Manning, 2018; Hu et al., 2017); and (b) examine how continuous representations contribute to supporting reasoning and alternative hypothesis formation in learning (Krishnaswamy et al., 2019). We point to the cutting-edge research that shows the influence of explicit representation of concepts (Hu et al., 2019; Liu et al., 2019). The main goal of this tutorial is to combine these current related efforts from different communities and application domains into one unified treatment, to identify the challenges, problems and future directions for spatiotemporal language understanding.

2 Outline

The tutorial will cover the following syllabus.

1. Spatial-Temporal Symbolic Representations and Extraction
   - Annotation schemes and symbolic semantic representation of space.
   - Annotation schemes and symbolic semantic representation of time.
   - Spatial Information Extraction from Language
   - Temporal Information Extraction from Language
2. Spatial and Temporal Reasoning and Grounding
   - Spatial and Temporal Reasoning and Evaluation with Language Models
   - Evaluations with Spatial QA, VQA, and Diffusion Models
   - Spatial and Temporal Reasoning with Formal Logical Representations
   - Multimodal spatial reasoning and dense paraphrasing
   - Grounding language into physical 2D and 3D coordinates
   - Grounding events into 1D timelines
   - Commonsense LLMs
3. Downstream Applications
   - Vision and Language Navigation
   - Motion planning for robots
   - Situated Grounding and multimodal dialogues
   - Self-driving cars, Clinical reports timeline

Duration: 3 hours, we estimate to present 50% our research work and 50% other related research. **Diversity Considerations:** The organizing committee, is diverse from the gender perspective of the instructors, coming from industry and academia, covering the research that is done in European Union as well as national US projects. It includes both junior and senior instructors affiliated with different organizations and countries. **Special requirements:** No specific equipment, other than video projector and internet access. **Number of attendees:** The topics, potentially are interesting for a large audience. This research direction has been paid a lot of attention recently, particularly the application areas that we cover in this tutorial and the research on the evaluation of large language models. We estimate 100 attendees. **Venue:** This Tutorial is presented at NAACL-2024. **Open access:** We make all the teaching material publicly available[1] and allow ACL to publish the slides and the video recording of the tutorial in the ACL Anthology.

3 Prerequisites and reading list

Familiarity with machine learning and natural language processing will be helpful for tutorial attendees. Our selected reading list is as follows.

- Qualitative spatial representation and reasoning. Anthony G. Cohn, and Jochen Renz. Foundations of Artificial Intelligence 3 (2008): 551-596.

- A linguistic ontology of space for natural language processing. John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. Artificial Intelligence 174, no. 14 (2010): 1027-1071.

- Spatial Role Labeling: Task Definition and Annotation Scheme. Parisa Kordjamshidi, Marie-Francine Moens, Martijn van Otterlo, (2010). Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).

- The qualitative spatial dynamics of motion in language. James Pustejovsky, and Jessica L. Moszkowicz. Spatial Cognition Computation 11, no. 1 (2011): 15-44.

- Interpreting Motion: Grounded Representations for Spatial Language. Inderjeet Mani and James Pustejovsky (2012), Explorations in language and space. Oxford University Press.

- Changing perspective: Local alignment of reference frames in dialogue, Simon Dobnik, Christine Howes, JD Kelleher, Proceedings of SEMDIAL (goDIAL), 24-32, 2015.

[1] Slides: https://spatial-language-tutorial.github.io/

- Global machine learning for spatial ontology population. Parisa Kordjamshidi, Marie-Francine Moens, (2015). Journal of Web Semantics, 30, 3-21.

- VoxML: A Visualization Modeling Language. James Pustejovsky, and Nikhil Krishnaswamy. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4606-4613. 2016.

- Do you see what I see? Effects of pov on spatial relation specifications. Nikhil Krishnaswamy, and James Pustejovsky. In Proc. 30th International Workshop on Qualitative Reasoning. 2017.

- ISO-Space: Annotating static and dynamic spatial information. James Pustejovsky (2017). In Handbook of Linguistic Annotation, pages 989–1024. Springer.

- Spatial role labeling annotation scheme. Parisa Kordjamshidi, Martijn van Otterlo, Marie-Francine Moens, (2017). In: Pustejovsky J., Ide N. (Eds.), Handbook of Linguistic Annotation Springer Verlag.

- Source-target inference models for spatial instruction understanding. Hao Tan and Mohit Bansal (2018). In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) (5504-5511).

- Acquiring common sense spatial knowledge through implicit spatial templates. Guillem Collell, Luc Van Gool and Marie-Francine Moens (2018). In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018) (pp. 6765-6772). AAAI.

- Generating a Novel Dataset of Multimodal Referring Expressions. Nikhil Krishnaswamy, and James Pustejovsky. In Proceedings of the 13th International Conference on Computational Semantics, pp. 44-51. 2019.

- StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. Zhengxiang Shi, Qiang Zhang, Aldo Lipani, Proceedings of the AAAI Conference on Artificial Intelligence, 36 (2022) 11321-11329.

- SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. In Proceedings NAACL-2021, pages 4582–4598, Online. Association for Computational Linguistics.

- A Multi-axis Annotation Scheme for Event Temporal Relations. Qiang Ning, Hao Wu, and Dan Roth. 2018. In Proceedings of ACL-2018, pages 1318-1328.

- TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. In Proceedings of EMNLP-2020, pages 1158–1172.

- A Meta-framework for Spatiotemporal Quantity Extraction from Text. Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, and Matt Gardner. In Proceedings of ACL-2022, pages 2736–2749.

- SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. Chen, Boyuan and Xu, Zhuo and Kirmani, Sean and Ichter, Brian and Driess, Danny and Florence, Pete and Sadigh, Dorsa and Guibas, Leonidas and Xia, Fei,arXiv preprint arXiv:2401.12168,2024.

- Visual Spatial Reasoning. Fangyu Liu, Guy Emerson, Nigel Collier; Transactions of the Association for Computational Linguistics 2023.

- Multi-agent Motion Planning from Signal Temporal Logic Specifications. Dawei Sun, Jingkai Chen, Sayan Mitra, Chuchu Fan. IEEE Robotics and Automation Letters (RA-L).

- NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Yongchao Chen, Rujul Gandhi, Yang Zhang, and Chuchu Fan. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

4 Instructors

- **Parisa Kordjamshidi** is an Assistant Professor of the Computer Science Department at Michigan State University. She has been working on spatial semantics extraction and annotation schemes, mapping language to formal spatial representations, spatial ontologies, structured output prediction models for information extraction, and combining vision and language for spatial language understanding. She has organized/co-organized shared tasks on Spatial role labeling, SpRL-2012, SpRL-2013, and the Space Evaluation workshop, SpaceEval-2015, in the SemEval Series and Multimodal spatial role labeling workshop mSpRL at CLEF-2017 intending to consider vision and language media for spatial information extraction. She organized SpLU at (NAACL-18, EMNLP-2020) and Robonlp-SpLU at (NAACL-2019, ACL-IJCNLP 2021). Email: kordjams@msu.edu. Webpage: http://www.cse.msu.edu/~kordjams.

- **Qiang Ning** is an applied scientist at AWS (2022-) leading the human alignment team for Titan LLMs. Prior to that, Qiang was an applied scientist at Alexa (2020-2022) and a research scientist at the Allen Institute for AI (2019-2020). Qiang received his Ph.D. from the University of Illinois at Urbana-Champaign in 2019 in Electrical and Computer Engineering. Qiang's research interests span in information extraction, question answering, and the application of weak supervision methods in these NLP problems in both theoretical and practical aspects. Email: qiangning.01@gmail.com. Webpage: https://www.qiangning.info/

- **James Pustejovsky** is the TJX Feldberg Chair in Computer Science at Brandeis University, where he is also Chair of the Linguistics Program, Chair of the Computational Linguistics MA Program, and Director of the Lab for Linguistics and Computation. He received his B.S. from MIT and his Ph.D. from UMASS at Amherst. He has worked on computational and lexical semantics for 25 years and is the chief developer of Generative Lexicon Theory. Since 2002, he has been working on the development of a platform for temporal reasoning in language, called TARSQI (www.tarsqi.org). Pustejovsky is the chief architect of TimeML and ISO-TimeML, a recently adopted ISO standard for temporal information in language, as well as the recently adopted standard, ISO-Space, a specification for spatial information in language. He has developed a modeling framework for representing linguistic expressions and interactions as multimodal simulations. This platform, VoxML, enables real-time communication between humans and computers or robots for joint tasks, utilizing speech, gesture, gaze, and action. He is currently working with robotics researchers in HRI to allow the VoxML platform to act as both a dialogue management system as well as a simulation environment that reveals real-time epistemic state and perceptual input to a computational agent. His areas of interest include Computational semantics, temporal and spatial reasoning, language annotation for machines. Email: jamesp@brandeis.edu. Webpage: http://www.pusto.com.

- **Marie-Francine Moens** is a Full Professor at the Department of Computer Science, KU Leuven. She has a special interest in machine learning for natural language understanding and in grounding language in a visual context. She is a holder of the prestigious ERC Advanced Grant CALCULUS (2018-2023) granted by the European Research Council on the topic of language understanding. She is currently associate editor of the journal IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). In 2011 and 2012 she was appointed as chair of the European Chapter of the Association for Computational Linguistics (EACL) and was a member of the executive board of the Association for Computational Linguistics (ACL). From 2014 to 2018 she was the scientific manager of the EU COST action iV&L Net (The European Network on Integrating Vision and Language). Email: sien.moens@cs.kuleuven.be. Webpage: https://people.cs.kuleuven.be/~sien.moens

References

Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15.

James F Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

Piyush Bagad, Makarand Tapaswi, and Cees G.M. Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2503–2516.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

J. A. Bateman. 2010. Language and space: A two-level semantic approach based on principles of ontological engineering. *International Journal of Speech Technology*, 13(1):29–48.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. 2:273–284.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3043–3054.

Anthony G. Cohn, Brandon Bennett, John Gooday, and Nicholas M. Gotts. 1997. Representing and reasoning with qualitative spatial relations. In Oliviero Stock, editor, *Spatial and Temporal Reasoning*, pages 97–132. Springer.

Guillem Collell, Thierry Deruyttere, and Marie-Francine Moens. 2021. Probing spatial clues: Canonical spatial templates for object relationship understanding. *IEEE Access*, 9:134298–134318.

Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Martha Palmer, and Dan Roth. 2020. In *Proceedings of Language Resources and Evaluation Conference, LREC-2020*.

Liesbeth Degand and Philippe Muller. 2020. Dialogue and dialogue systems. *TAL: revue internationale Traitement Automatique des Langues*, 61(3).

Thierry Deruyttere, Victor Milewski, and Marie-Francine Moens. 2021. Giving commands to a self-driving car: How to deal with uncertain situations? *Eng. Appl. Artif. Intell.*, 103:104257.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. 2023. Empowering dynamics-aware text-to-video diffusion with large language models.

Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. 2021. Core challenges in embodied vision-language planning.

Dusan Grujicic, Thierry Deruyttere, Marie-Francine Moens, and Matthew B. Blaschko. 2022. Predicting physical world destinations for commands given to self-driving cars. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 715–725. AAAI Press.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Johana Hois and Oliver Kutz. 2011. Towards linguistically-grounded spatial logics. In *Spatial Representation and Reasoning in Language: Ontologies and Logics of Space*, number 10131 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 804–813.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6551–6557.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

T. Kollar, S. Tellex, D. Roy, and N. Roy. 2010. Toward understanding natural language directions. In *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '10, pages 259–266. ACM.

Parisa Kordjamshidi and Marie-Francine Moens. 2015. Global machine learning for spatial ontology population. *Web Semant.*, 30(C):3–21.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2010. Spatial role labeling: task definition and annotation scheme. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 413–420.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing*, 8:1–36.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2017. Spatial role labeling annotation scheme. In N. Ide James Pustejovsky, editor, *Handbook of Linguistic Annotation*. Springer Verlag.

Katherine Krajovic, Nikhil Krishnaswamy, Nathaniel J Dimick, R Pito Salas, and James Pustejovsky. 2020. Situated multimodal control of a mobile robot: Navigation through a virtual environment. *arXiv e-prints*, pages arXiv–2007.

Nikhil Krishnaswamy, Scott Friedman, and James Pustejovsky. 2019. Combining deep learning and qualitative spatial reasoning to learn complex structures from sparse examples with noise. In *Proceedings of the thirty-third AAAI Conference on Artificial Intelligence*.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246, Brussels, Belgium. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2020. Towards extracting absolute event timelines from english clinical reports. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28:2710–2719.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023. Diplomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems*, 36:46856–46884.

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

Yunchao Liu, Jiajun Wu, Zheng Wu, Daniel Ritchie, William T. Freeman, and Joshua B. Tenenbaum. 2019. Learning to describe scenes with programs. In *International Conference on Learning Representations*.

Gengchen Mai, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, and Ni Lao. 2020. Se-kge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24(3):623–655.

I. Mani and James Pustejovsky. 2012. *Interpreting Motion: Grounded Representations for Spatial Language*. Explorations in language and space. Oxford University Press.

Inderjeet Mani. 2009. SpatialML: annotation scheme for marking spatial expression in natural language. Technical Report Version 3.0, The MITRE Corporation.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. SpartQA: A textual question answering benchmark for spatial reasoning. *NAACL*.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Int. Res.*, 71:1183–1317.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A Reading Comprehension Dataset of Temporal ORdering QUEstions. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. A multi-axis annotation scheme for event temporal relations. pages 1318–1328. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. CogCompTime: A Tool for Understanding Time in Natural Language. In *EMNLP (Demo Track)*, Brussels, Belgium. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, and Matt Gardner. 2022. A meta-framework for spatiotemporal quantity extraction from text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2736–2749, Dublin, Ireland. Association for Computational Linguistics.

J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5)*.

James Pustejovsky. 2017. Iso-space: Annotating static and dynamic spatial information. In *Handbook of Linguistic Annotation*, pages 989–1024. Springer.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The TIMEBANK corpus. In *Corpus Linguistics*, page 40.

James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz*, pages 1–21.

James Pustejovsky, J. Moszkowicz, and M. Verhagen. 2011. ISO-Space: The annotation of spatial information in language. In *ACL-ISO International Workshop on Semantic Annotation (ISA'6)*.

Taher Rahgooy, Umar Manzoor, and Parisa Kordjamshidi. 2018. Visually guided spatial relation extraction from text. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT 2018*.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11321–11329.

Kristin Stock, Robert C. Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone. 2013. Creating a corpus of geospatial natural language. In *Spatial Information Theory*, pages 279–298, Cham. Springer International Publishing.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *ACL*.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Jesse Thomason, Jivko Sinapov, Raymond Mooney, and Peter Stone. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. 2024. NavHint: Vision and language navigation agent with a hint generator. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 92–103, St. Julian's, Malta. Association for Computational Linguistics.

Yue Zhang and Parisa Kordjamshidi. Vln-trans, translator for the vision and language navigation agent. *The 61st Annual Meeting of the Association for Computational Linguistics (ACL-2023)*.

Kaiyu Zheng, Deniz Bayazit, Rebecca Mathew, Ellie Pavlick, and Stefanie Tellex. 2021. Spatial language understanding for object search in partially observed city-scale environments.