Points, Paths, and Playscapes: Large-scale Spatial Language Understanding Tasks Set in the Real World

Jason Baldridge Tania Bedrax-Weiss Daphne Luong Srini Narayanan
Bo Pang Fernando Pereira Radu Soricut Michael Tseng Yuan Zhang
Google Inc.
Mountain View, CA
{jridge,tbedrax,daphnel,srinin,bopang,pereira,
rsoricut,michaeltseng,zhangyua}@google.com

Abstract

Spatial language understanding is important for practical applications and as a building block for better abstract language understanding. Much progress has been made through work on understanding spatial relations and values in images and texts as well as on giving and following navigation instructions in restricted domains. We argue that the next big advances in spatial language understanding can be best supported by creating large-scale datasets that focus on points and paths based in the real world, and then extending these to create online, persistent playscapes that mix human and bot players, where the bot players must learn, evolve, and survive according to their depth of understanding of scenes, navigation, and interactions.

1 Introduction

Language is not sealed in a textual medium disconnected from the world. People use language to talk about people, places and things that exist both in time and space. Abstract ideas are typically conveyed through metaphors that are grounded in embodied concepts from the domains of spatial movement, forces, and manipulation (Narayanan, 1999). Mental simulation involving motor and perceptual content likely plays a crucial role in sentence comprehension (Bergen et al., 2010). Natural language understanding thus requires the ability to analyze complex descriptions that relate referents spatially and temporally and connect them to grounded locations and times.

One of the richest domains for encountering such language is that of providing and following navigational instructions involving both named and vague references and relationships in both indoor and outdoor contexts. Spatial navigation itself is one of the better understood aspects of cognitive function, including extensive research into cells that encode grids, boundaries and directions (Chersi and Burgess, 2015; Moser et al., 2017). This indicates that work on spatial tasks in language has the potential to lead to a virtuous cycle between modeling of language and understanding of the brain and cognition.

No current systems adequately support natural language interactions for spatial tasks. Geospatial mapping applications (such as Google Maps) provide algorithmic, route-based instruction at a global scale. However, they rely on explicitly named roads, paths, and addresses, and they assume a large database as a model of the world, which includes mappings between names and geo locations. Such systems give instructions but cannot interpret them, much less interact with a human user. They also typically do poorly at providing contextual descriptions, especially for buildings, bridges, and other salient landmarks.

Understanding spatial references from natural language must handle inherent spatial vagueness and other features of the figure, and ground objects or trajectories in a coordinate system. Spatial grounding is relative—it depends on size, shape, and function of the figure and ground objects. Furthermore, it is identified by transforming location with respect to reference frames in language (Levinson, 2003; Tenbrink and Kuhn, 2011) to a ground. Languages have many options for describing the spatial relationships between different participants and objects and these must be reconciled with the ground- truth scene or map.

We argue that the next big advances in spatial language understanding can be best enabled by first creating large-scale datasets (hundreds of thousand to millions of examples) that require spatial understanding of real world points and paths, and next, building on these to create persistent, online playscapes that enable both automated agents and people to interact in virtual and augmented re-

ality environments.

Navigation involves traversal through a series of points, and each point can involve detailed scene understanding needs. Navigation is also an excellent link between the local (e.g., within a building) and the global (e.g., across a continent) variants of spatial tasks. Scene understanding—in both images and texts—is needed at both ends of this scale and in between. We expect that such a project provides challenges of high complexity, while also linking in to rich, already-available resources that connect both text and images to each other and to key metadata, including coordinates in both space and time.

2 Pillars and Principles

Here are some considerations as we begin a multi-year effort to create these resources.

2.1 Data and annotation

Our goal is to create large-scale resources that encompass natural spatially oriented tasks that ordinary people accomplish every day.

Scale To be able to work with diverse locations (e.g., cities, theme parks, natural settings) across the world, we need large datasets associating language with spatially relevant points and paths—on the scale of *at least hundreds of thousands*.

Multilinguality For both theoretical and practical reasons, we cannot focus on just one language. Different languages have different spatial relations, often involving the three different frames of reference—relative, intrinsic, and absolute (Levinson, 2003)—in different ways. Navigational systems supporting vague reference off the grid are needed even more in locations where English and other majority languages are not spoken.

One way we already target multilinguality is via community-driven crowd-sourcing (Funk et al., 2018). In our approach, we intentionally cycle our iterations throughout the world and we involve developers from each locale because they have insights into how the local context affects how language is used and how the task is performed.

User-driven annotation We seek to complement previous efforts that have focused on fine-grained linguistic annotation, such as Iso-Space (Pustejovsky, 2017). We will obtain scale through both crowd-sourcing and gaming environments—that is, annotations that can be derived from competent language speakers (Chang et al., 2016). This places an emphasis on task evaluations with implicit feedback rather than prediction and evaluation of labels on text and images. Spatial tasks are natural fits for this strategy, since both evaluation metrics and reward functions (in reinforcement learning) can use spatial proximity to an end location (MacMahon et al., 2006; Chen and Mooney, 2011; Vogel and Jurafsky, 2010; Artzi and Zettlemoyer, 2013) or spatial configuration (Bisk et al., 2018; Misra et al., 2017; Tan and Bansal, 2018).

There are trade-offs between *model-driven* and *user-driven* corpus building. The former defines inventories of spatial relations and generating assignments that will cover them. This may omit phenomena or distinctions not covered in the model and requires considerable expertise and tooling—both of which increase cost and limit scale. User-driven annotation is more exploratory and may be limited by the preferences and tendencies of contributors. We will mitigate such effects by composing diverse crowds from various locales (Funk et al., 2018). Ultimately, we seek to create resources that contain language grounded in spatial relations that, by construction, include extra-linguistic factors like vantage point, shared context, and other location-dependent world knowledge. We also expect this setting to support complementary non-linguistic spatial understanding approaches, such as Simultaneous Localization and Mapping (Cadena et al., 2016).

Sharing and privacy To facilitate accessibility and reproducibility, the source material used for building resources should be, wherever possible, unencumbered by copyright and be acquired with full permission from content creators. Location information brings with it significant privacy and ethical considerations. We will thus focus on locations in shared public spaces that avoid close connections to any person who helps create the data. We will develop our datasets using open resources such as Wikipedia and Open Street Maps combined with materials produced by (both paid and volunteer) crowd contributors who have granted permission in advance. Overall, our datasets and environments will be built—from start to finish—to be compliant with the European Union's General Data Protection Regulation (Council of European Union, 2016).[1]

[1] https://www.eugdpr.org/

2.2 Task considerations

We emphasize the real world as the basis for spatial language understanding tasks, while allowing for a spectrum of resources from digitized real world artifacts to virtual environments to augmented real world interactions.

Real world emphasis. The natural starting point for building spatial language understanding capabilities is the real world itself. For example, spoken interfaces to mobile robots necessarily integrate vague reference and learning a local map through exploration (Thomason et al., 2015; Hanheide et al., 2017; Arkin et al., 2017). Unfortunately, working with physical robots brings additional challenges such as dealing with hardware calibration and failure. Thus, many researchers have opted instead to work with simulated environments that enable faster iteration on modeling and learning (Jänner et al., 2017; Hermann et al., 2017; Bisk et al., 2018), and some support both movement and manipulation (Yan et al., 2018).

Simulated environments, however, do not represent full real world messiness. It is thus interesting to consider a middle ground: working with high-fidelity simulations of the real world. For example, Anderson et al. (2017) introduce a visually grounded navigation task set in 3D simulations of actual houses. This requires both rich scene understanding and difficult language interpretation. We intend to work in this same mode, gathering digitized artifacts relating to real world locations—including databases, texts, images, and more—to support complex and compelling tasks that can impact real world applications. In particular, we expect to achieve considerable scale on navigational tasks for walking through a campus or park.

First-person perspective. For at least some of the tasks we envision, human and machine agents will not have access to a God's eye view, like that available to mapping applications (with access to full geographic features via databases). Instead, such tasks must be solvable by interpreting visual and textual stimuli relevant to the locations. This should put a greater emphasis on challenging spatial descriptions and relationships rather than known and named routes. Nonetheless, maps as visual artifacts (e.g., PDFs) may be incorporated in some cases, giving automated agents the ability to use them as a hiker might use a paper map without access to a GPS-based mapping application.

Mirowski et al. (2018) is a recent example that takes a first-person perspective in a real world simulation, though one that does not incorporate language. They learn a model for navigating the Google Street View graph via reinforcement learning, where the goal location is specified via its distance to several other landmark locations and no explicit maps are used. Two especially interesting aspects of their approach are their use of curriculum learning (start with nearby goals and then tackle more distant ones) and showing successful adaptation from one city to another. These ideas are complementary to those that use language as a component in learning to navigate, so it should be possible to effectively integrate linguistic inputs (e.g., directions and descriptions of the goal) into the approach.

Human–machine interaction Thomason et al. (2015) demonstrate a robot that interacts with people and incrementally expands its language understanding capabilities. In this vein, we seek to create simulated (real world) environments that support spatial language tasks in which bots and humans mix, collaborate, and compete. In such settings, there is no annotation: instead, players–both bot and human–gain points, status, and bounty (e.g., compute credits) by accomplishing goals.

This approach opens up opportunities to transition from static tasks such as following a particular set of navigational instructions to dynamic interactions such as following instructions made in the moment and in context by another player. If successful, this dynamism could create far greater scale for iterating on modeling ideas—with the evaluation measure (success in the game) as a built-in feature. This approach not only frees us from the need for costly, one-off annotation efforts, but also creates an ecologically compelling environment where progress is forced on and by the bots: they must perform well to get rewards to stay alive and maintain their status in the playscape (such as compute credits). As importantly, this survival criterion also entails the need to attend to representational and computational efficiency (FLOPS) on top of overall ability.

Building playscapes also plots a path from virtual real world to augmented reality applications and games that include linguistic interactions between human and bot players, and manipulation of virtual objects that have real world locations. Capturing Pokémon characters and interacting with

gyms in Pokémon Go are examples of such manipulations.

3 Tasks

Our focus on real world spatial language artifacts provides a natural and mutually reinforcing progression from points to paths to playscapes.

Points Scene understanding—building a model for a point in space—is the bedrock of real world spatial language tasks. We must be able to observe and describe visible objects and the spatial relationships between them. Before addressing paths and navigation tasks, we can make considerable progress by improving our data and modeling for spatial relations in tasks like image segmentation and image captioning (Hall et al., 2011; Hürlimann and Bos, 2016), grounding referential expressions (Kazemzadeh et al., 2014; Mao et al., 2016; Hu et al., 2017), relative positioning of objects (Kitaev and Klein, 2017) and image geolocation (Hays and Efros, 2008; Zamir et al., 2016). We will create collaborative image identification and description tasks that emphasize spatial relations and geographically salient landmarks.

There has also been much work on annotating and calculating spatial relations in text (Pustejovsky et al., 2015; Pustejovsky, 2017), resolving toponyms (Leidner, 2007; DeLozier et al., 2015), and text geolocation (Wing and Baldridge, 2014; Rahimi et al., 2017). There are further opportunities for building or exploiting annotations on spatially focused texts—e.g., identifying vague regions (DeLozier et al., 2016) or writing a WikiVoyage page for a city given all available information in Wikipedia, akin to Liu et al. (2018).

Most importantly, the extensive mappings we have between texts and images and their corresponding locations motivate a focus on simulations of the real world. Learning spatial relations within massive amounts of images and texts can serve as a pretraining step to building components of models that solve real world navigation tasks.

Paths Understanding salient features and spatial relations in images and text naturally extends into navigation tasks that connect such points. To avoid biases, we will create navigation challenges through several different means, with an emphasis on domains that require a mix of named features, salient landmarks, and general features that necessitate relational, imprecise reference.

Harvesting and extending: There are numerous, extensive walking tours of public spaces. For example, universities typically provide self-guided campus tours that include text, images, and maps. Considerable work is required to standardize the specification and formatting of the tours, organize the associated artifacts (such as pictures), and convert the analog paths to digital ones (or create them) so that they could be used in experiments.

Descriptions to paths: In other cases, we have human descriptions of journeys in resources like WikiVoyage, such as from airports to city centers or how to get into Grand Canyon by car from different directions. We can have multiple people follow the directions in a resource like Google Street View to establish both ground truth and capture variation in human performance.

Paths to descriptions: Many volunteers on OpenStreetMaps produce GPS traces,[2] and we can elicit navigational instructions covering them.

Points to paths and descriptions: Given points, we can generate random paths, elicit navigational instructions for them, and then have others generate paths following instructions. This setup does not depend on existing data and gives more control over variables such as the number of points, length of the descriptions, and more. It can also tie into existing point-based data, such as the Google Landmarks,[3] so that point and path models that reinforce each other can be explored.

This is the strategy we are beginning with: focusing on collecting navigational instructions in city centers, resorts and college campuses for itineraries that include three to ten points of interest. Itineraries will be generated both by sampling paths connecting waypoints drawn from gazeeteers and Wikipedia and by generating travel itineraries from real world trips (Friggstad et al., 2018). We will collect instructions given both by people who are physically on the ground as well as others visiting the points virtually via Google Street View. We expect that this effort will go through several iterations as we discover the pain points and better understand which approaches work best.

Playscapes Collecting datasets with paths and corresponding navigation instructions can provide a valuable source for learning and evalu-

[2] https://www.openstreetmap.org/traces

[3] https://research.googleblog.com/2018/03/google-landmarks-new-dataset-and.html

ating models. The HCRC MapTask (Anderson et al., 1991) is a launching-off point for creating collaborative games where participants help each other complete a virtual road rally. This naturally extends the path-oriented efforts discussed above, but mixes in collaboration and competition while providing motivation through in-game rewards (e.g., status, points, and compute credits). Such games could take a variety of forms: one possibility is to provide a series of waypoints drawn from a WikiVoyage page to one player who then uses the page and resources like Google Street View to write instructions. Another player (or players) must then follow the instructions and possibly solve additional puzzles or tasks along the way.

It would be even more powerful to create online, persistent games in which human and bot players need to understand multi-step natural language cues in order to find target locations and accomplish other in-game objectives. This moves us from creating datasets to establishing ecologically interesting playscapes, such as ones in which bots must solve navigation tasks in order to gain the rewards needed for their survival.

Here we focus on spatial motion and relations necessary for navigation and scene understanding. By embedding our tasks and playscapes in digitized versions of the real world, however, we provide a natural launching-off point for eventually adding manipulation via augmented reality applications. The recently released Google Maps gaming API[4] can be a significant enabling technology for creating such playscapes. A tantalizing prospect would be to create games akin to Ingress and Pokémon Go that furthermore involve language. The key would be to design them to be relevant, compelling and fun while ensuring privacy and safety.

Gamification also makes the playscape more compelling and fun for human participants. It gives a reason for participants to engage more with with other players and negotiate the spatial environment to achieve their in-game goals. We will likely assign asymmetric capabilities for both human and machine players. That is, players will take on different roles with different abilities—e.g., some could be scouts who have a wider range of (augmented) perception, while others could be manipulators who can acquire objects and solve puzzles requiring interaction with virtual objects at game-relevant real world locations. Machine agents could play many different roles, such as fast virtual scouts, helpful carriers of virtual objects, and translators who help interactions between human players who speak different languages. Such an environment should also provide a rich substrate for exploring approaches that incorporate pragmatic inference for giving and following instructions (Fried et al., 2018).

In designing such playscapes, we will avoid violent themes and combatitive gameplay. Instead, we seek to design them in the mold of collaborative board games like Forbidden Island. Players may still compete for overall higher individual rankings with respect to status and points, but we envision that they will do this by individually contributing to collaborative group efforts.

4 Conclusion

We seek to create large-scale datasets that thread together tasks that present challenges from points to paths and ultimately provide the basis upon which we create playscapes that incorporate real world data and interactions. The annotations for these will be in the form of language and behaviors rather than detailed formal linguistic representations. However, we believe it is likely that successful models will avail themselves of structured information around ideas like reference frames, structural biases in planning and navigation, and more. We also would welcome additional layers of analysis on the data we release.

In sum, we seek to produce richly associated data that ties text and images to locations at local, global, and scene-level resolutions. We hope to get feedback from the community and build collaborations as we begin this endeavor.

Acknowledgments

We thank Igor Karpov, Slav Petrov, Michael Ringgaard, Chris Waterson, David Weiss and the anonymous reviewers for their valuable feedback.

References

A Anderson, M Bader, E Bard, E Boyle, G. M Doherty, S Garrod, S Isard, J Kowtko, J McAllister, J Miller, C Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.

[4] https://developers.google.com/maps/gaming/

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2017. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *CoRR*, abs/1711.07280.

Jacob Arkin, Matthew R. Walter, Adrian Boteanu, Michael E. Napoli, Harel Biggie, Hadas Kress-Gazit, and Thomas M. Howard. 2017. Contextual awareness: Understanding monologic natural language instructions for autonomous robots. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1(1):49–62.

Benjamin K. Bergen, Shane Lindsay, Teenie Matlock, and Srini Narayanan. 2010. Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, 31(5):733–764.

Yonatan Bisk, Kevin Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI-18)*, New Orleans, USA.

C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332.

Nancy Chang, Russell Lee-Goldman, and Michael Tseng. 2016. Linguistic wisdom from the crowd. In *Crowdsourcing Breakthroughs for Language Technology Applications*.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-2011)*, San Francisco, CA, USA.

Fabian Chersi and Neil Burgess. 2015. The cognitive architecture of spatial navigation: Hippocampal and striatal contributions. *Neuron*, 88(1):64 – 77.

Council of European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198, Berlin, Germany. Association for Computational Linguistics.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of NAACL-HLT 2018*.

Zachary Friggstad, Sreenivas Gollapudi, Kostas Kollias, Tamas Sarlos, Chaitanya Swamy, and Andrew Tomkins. 2018. Orienteering algorithms for generating travel itineraries. In *International Conference on Web Search and Data Mining (WSDM)*.

Christina Funk, Michael Tseng, Ravindran Rajakumar, and Linne Ha. 2018. Community-driven crowdsourcing: Data collection with local developers. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Mark Hall, Philip D Smart, and Christopher Jones. 2011. Interpreting spatial language in image captions. *Cognitive processing*, 12:67–94.

Marc Hanheide, Moritz Gbelbecker, Graham S. Horn, Andrzej Pronobis, Kristoffer Sj, Alper Aydemir, Patric Jensfelt, Charles Gretton, Richard Dearden, Miroslav Janicek, Hendrik Zender, Geert-Jan Kruijff, Nick Hawes, and Jeremy L. Wyatt. 2017. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence*, 247:119 – 150. Special Issue on AI and Robotics.

James Hays and Alexei A. Efros. 2008. IM2GPS: estimating geographic information from a single image. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. 2017. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4418–4427. IEEE.

Manuela Hürlimann and Johan Bos. 2016. Combining lexical and spatial knowledge to predict spatial relations between objects in images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 10–18, Berlin, Germany. Association for Computational Linguistics.

Michaela Jänner, Karthik Narasimhan, and Regina Barzilay. 2017. Representation learning for grounded spatial reasoning. *CoRR*, abs/1707.03938.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Nikita Kitaev and Dan Klein. 2017. Where is misty? interpreting spatial descriptors by modeling regions in space. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 157–166, Copenhagen, Denmark. Association for Computational Linguistics.

Jochen L. Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* Dissertations.com.

Stephen C. Levinson. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press.

Peter J. Liu, Mohammad Ahmad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006)*, pages 1475–1482, Boston, MA, USA.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. *CoRR*, abs/1804.00168.

Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1015–1026. Association for Computational Linguistics.

Edvard Moser, May-Britt Moser, and Bruce Mcnaughton. 2017. Spatial representation in the hippocampal formation: A history. 20:1448–1464.

Srini Naraynan. 1999. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the National Conference on Artificial Intelligence*, pages 121–128, Orlando, Florida. AAAI Press.

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

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216, Vancouver, Canada. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2018. Source-target inference models for spatial instruction understanding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana.

Thora Tenbrink and Werner Kuhn. 2011. A model of spatial reference frames in language. In *Spatial Information Theory*, pages 371–390, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jesse Thomason, Shiqi Zhang, Raymond J Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *IJCAI*, pages 1923–1929.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, Uppsala, Sweden. Association for Computational Linguistics.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348, Doha, Qatar. Association for Computational Linguistics.

Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. 2018. Chalet: Cornell house agent learning environment. *CoRR*, abs/1801.07357.

Amir R. Zamir, Asaad Hakeem, Luc Van Gool, Mubarak Shah, and Richard Szeliski. 2016. Introduction to large-scale visual geo-localization. In Amir R. Zamir, Asaad Hakeem, Luc Van Gool, Mubarak Shah, and Richard Szeliski, editors, *Large-Scale Visual Geo-Localization*, pages 1–18. Springer International Publishing.