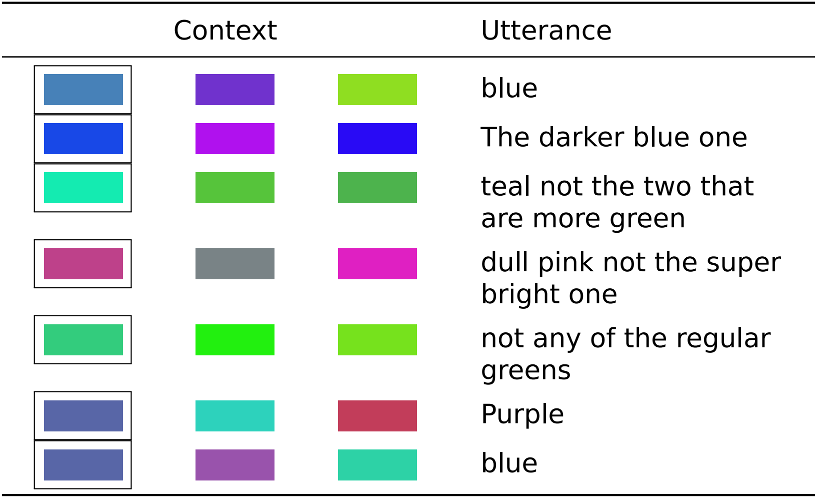
Grounded language understanding: Listeners: From language to the world

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

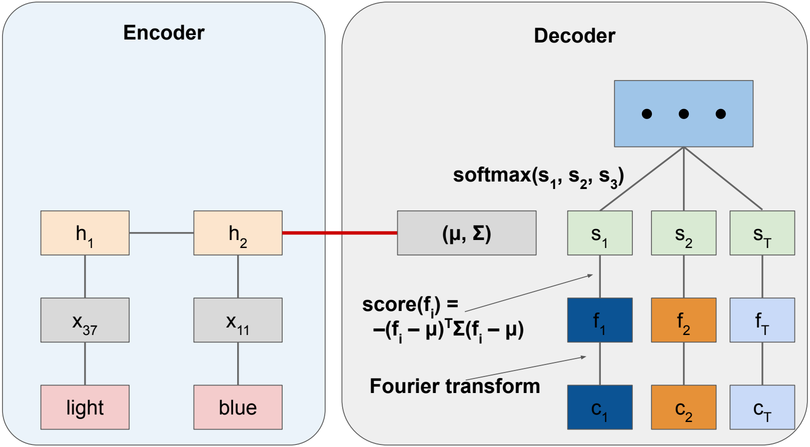
Color interpreter: Task formulation and data



Context			Utterance
🟦	🟪	🟩	blue
🟦	🟪	🟦	The darker blue one
🟩	🟩	🟩	teal not the two that are more green
🟥	⬜	🟪	dull pink not the super bright one
🟩	🟩	🟩	not any of the regular greens
🟦	🟦	🟥	Purple
🟦	🟪	🟩	blue

Stanford Colors in Context corpus
(Monroe et al. 2017)

A neural listener model



Other ideas and datasets

- NLU classifiers are very simple listeners: they consume language and make an inference in a structured space.
- Semantic parsers are very complex listeners: they consume language, construct rich latent representations, and predict into structured output spaces.
- Scene generation is the task of mapping language to structured representations of visual scenes (Seversky and Yin 2006; Chang et al. 2014, 2015).
- Young et al. (2014) seek to learn visual denotations for linguistic expressions.
- Mei et al. (2015) develop essentially a seq2seq version of the above model: given a linguistic input, they predict action sequences. (Kai Sheng Tai did his 2015 CS224u project on this, working at the same time as Mei et al.!)
- Suhr et al. (2019): Released the CerealBar data and game engine for learning to execute instructions.

References I

Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. 2015. Text to 3d scene generation with rich lexical grounding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 53–62, Stroudsburg, PA. Association for Computational Linguistics.

Angel Chang, Manolis Savva, and Christopher D. Manning. 2014. Learning spatial knowledge for text to 3D scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038, Doha, Qatar. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2015. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. ArXiv:1506.04089.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Lee M Seversky and Lijun Yin. 2006. Real-time automatic 3D scene generation from natural language voice and text descriptions. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 61–64. ACM.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.