

Coherence and Grounding in Multimodal Communication

Malihe Alikhani

University of Pittsburgh

December 1, 2020

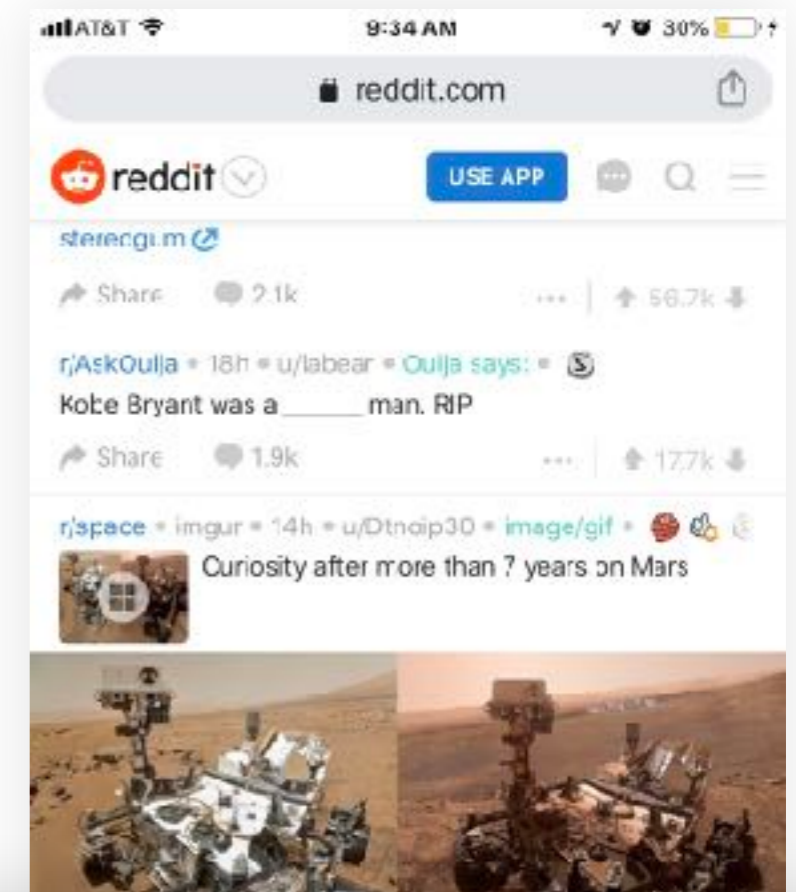




Communication is multimodal!



Communication is multimodal, especially on the internet!



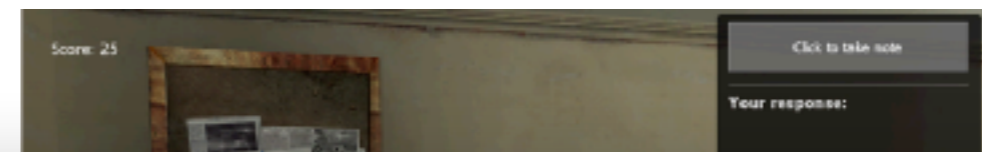
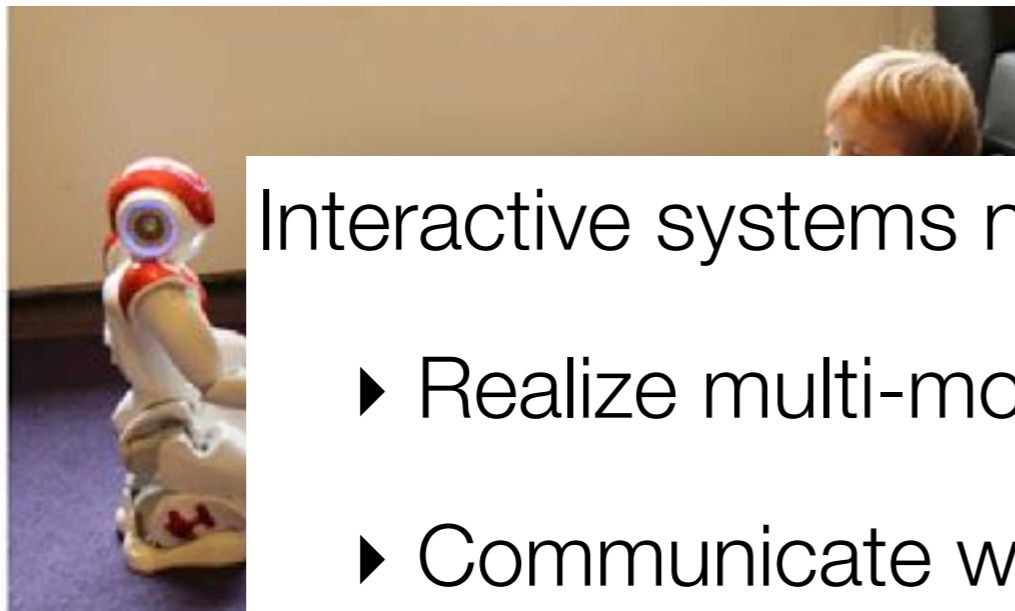
Every classic problem in NLP and information management needs to be generalized to handle multimodal datasets.

Machines are trying to catch up!



Credits: Amazon, GettyImages, AFP, Google

Machines are trying to catch up!



Interactive systems need to

- ▶ Realize multi-modal contributions.
- ▶ Communicate with people using a broad range of appropriate modalities.

Part 1: Commonsense and Coherence

Computational Challenges

Integrating space, visual presentations and language requires **learning commonsense inferences.**

Computational Challenges

Integrating space, visual presentations and language requires **learning commonsense inferences.**



Visual and linguistic communication have similar intentional, contextual and inferential properties.

Commonsense Inference in Text and Imagery

A wide range of background knowledge needs to be integrated with visual presentations.



A view from the bridge

Photo credit: Garden-party Limeui/Alamy



A man is sitting in front of a bunch of fruits.

Photo credit: Carol Mitchell

Surface level models that don't take into account these inferences have systematic problems.

Surface level models that don't take into account these inferences have systematic problems.

Content Hallucination



Model



A close up of a stuffed animal on a plate.

Example from Lu et al. 2018

Surface level models that don't take into account these inferences have systematic problems.

Content Hallucination



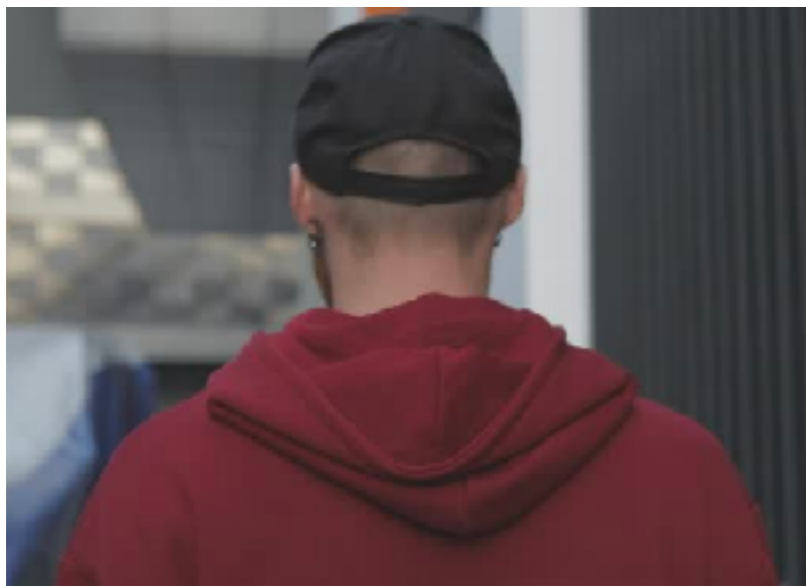
Model



A close up of a stuffed animal on a plate.

Example from Lu et al. 2018

Context Hallucination



Model



This is the new manager of the team.

Example from Sharma et al. 2019

Architecture

Classic Architecture

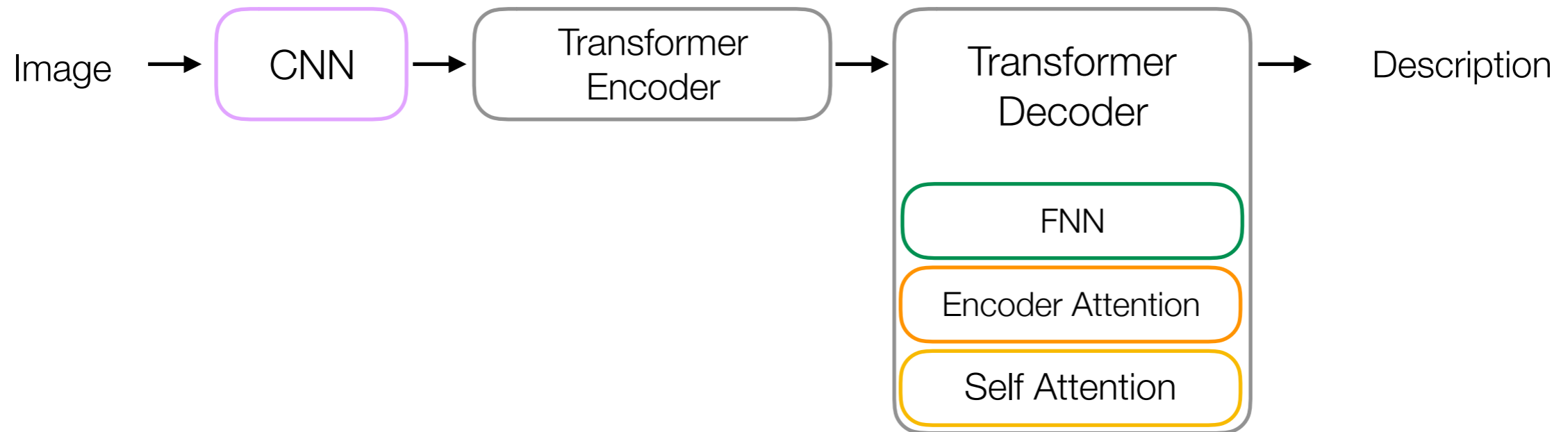


Architecture

Classic Architecture



SotA by Zhao et al. 2019



Generating Descriptions for Images

How can we address these issues?

- ▶ Commonsense understanding doesn't fall out of machine learning methods.
- ▶ Systems needs to recognize that image descriptions convey different kinds of information and fulfill different purposes.

The police refused the women a permit because they **advocated** violence.

The police refused the women a permit because they **feared** violence.

John can open Bill's safe. He knows the combination.

John can open Bill's safe. He should change the combination.

?

John can open Bill's safe. He knows the combination.

?

John can open Bill's safe. He should change the combination.



Connection among ideas first articulated by the philosopher David Hume:
Resemblance, Cause-Effect, and Contiguity.

2. Discourse Coherence

Discourse

- Discourse is made of sequences of individual segments which can form a more complex meaningful unit.
- Central features of discourse:
 1. Structure: relations that link segments
 2. Dynamics: overall content of the discourse grows, expands and gets enriched

Discourse

- Discourse is made of sequences of individual segments which can form a more complex meaningful unit.
- Central features of discourse:
 1. Structure: relations that link segments
 2. Dynamics: overall content of the discourse grows, expands and gets enriched

Discourse

- Discourse is made of sequences of individual segments which can form a more complex meaningful unit.
- Central features of discourse:
 1. Structure: relations that link segments
 2. Dynamics: overall content of the discourse grows, expands and gets enriched

Coherence Relations

- Segments of discourse are logically related to one another. (From Hobbs 1979):
 1. John took a train from Paris to Istanbul. He has family there.
 2. John took a train from Paris to Istanbul. He likes spinach.

Coherence Relations

- Segments of discourse are logically related to one another. (From Hobbs 1979):
 1. John took a train from Paris to Istanbul. He has family there.
 2. John took a train from Paris to Istanbul. He likes spinach.

Coherence Relations

- Pronoun interpretation(Hobbs 1985):

Explanation

1. **John** can open Bill's safe. **He** knows the combination.

Result

2. John can open **Bill's** safe. **He** should change the combination.

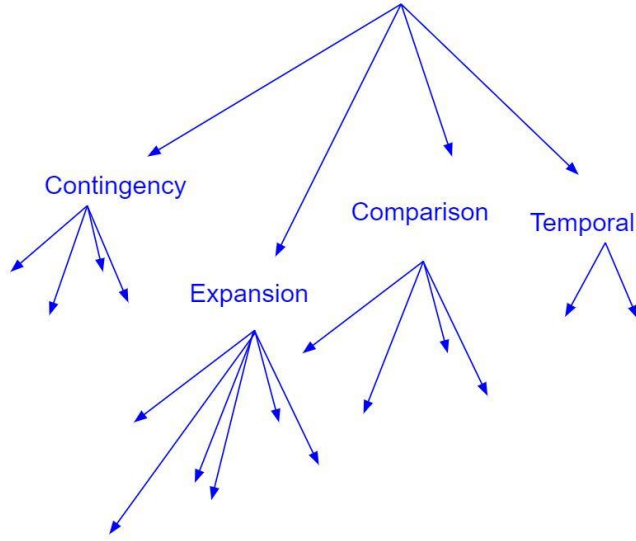
Computational Methods

1. Supervised
2. Unsupervised

Supervised

Datasets: Penn Discourse Treebank(PDTB)

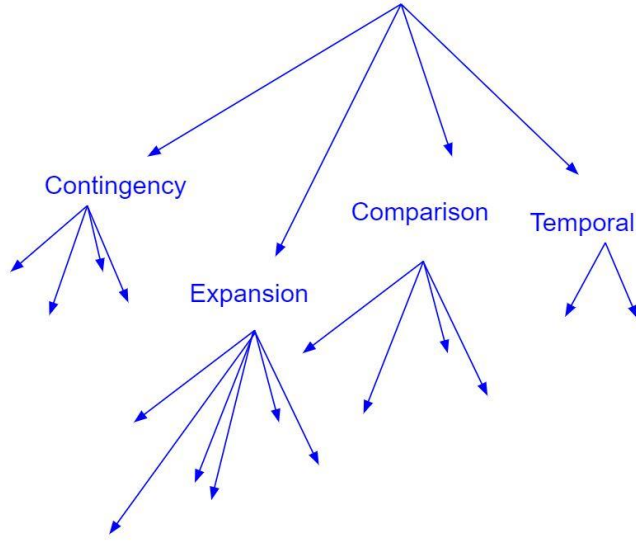
PDTB annotates [relations](#) between adjacent text spans in Wall Street Journal articles.



Supervised

Datasets: Penn Discourse Treebank(PDTB)

PDTB annotates **relations** between adjacent text spans in Wall Street Journal articles.



Example:

__Arg1__

Mr. Smith said the company's sales pace has been picking up

__Arg2__

because the effect of unfavorable exchange rates has been easing.

__Explicit__

because, Contingency.Cause.Reason

Supervised

Linguistically informed features

First-last

The first and last words of arguments are indicative discourse relations. (Pitler et al 2008)

Verbs

The tense of the main verbs of arguments can be good indicators for Temporal and Causal relations. (Pitler et al 2008)

Neural Nets

Convolutional methods, Adversarial models(Biran and McKeown, 2013;Qin et al., 2017)

Supervised

Linguistically informed features

First-last

The first and last words of arguments are indicative of discourse relations. (Pitler et al 2008)

Expansion(restatement)

She thought the story was predictable. **In other words**, She found it boring.

Neural Nets

Convolutional methods, Adversarial models(Biran and McKeown, 2013;Qin et al., 2017)

Supervised

Linguistically informed features

First-last

The first and last words of arguments are indicative discourse relations. (Pitler et al 2008)

Verbs

The tense of the main verbs of arguments can be good indicators for Temporal and Causal relations. (Pitler et al 2008)

Neural Nets

Convolutional methods, Adversarial models (Biran and McKeown, 2013; Qin et al., 2017)

Supervised

Linguistically informed features

First-last

The
an

Temporal

I **invited** Susan to my party yesterday. She **will
bring** her guitar.

N

Convolut

Verbs

The tense of the main verbs of arguments can be good indicators for Temporal and Causal relations.

Supervised

Linguistically informed features

First-last

The first and last words of arguments are indicative discourse relations.

Verbs

The tense of the main verbs of arguments can be good indicators for Temporal and Causal relations.

Neural Nets

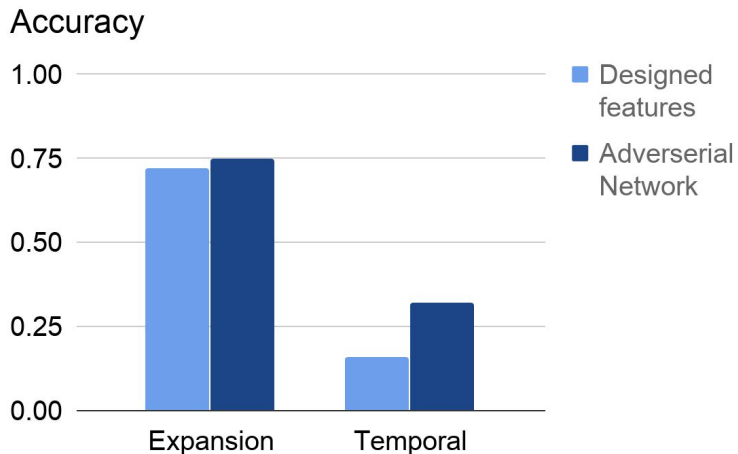
Convolutional methods, Adversarial models (Biran and McKeown, 2013;Qin et al., 2017)

Supervised

Linguistically informed features

First-last

The first and last arguments are discourse related
'However' for instance marks a contrast

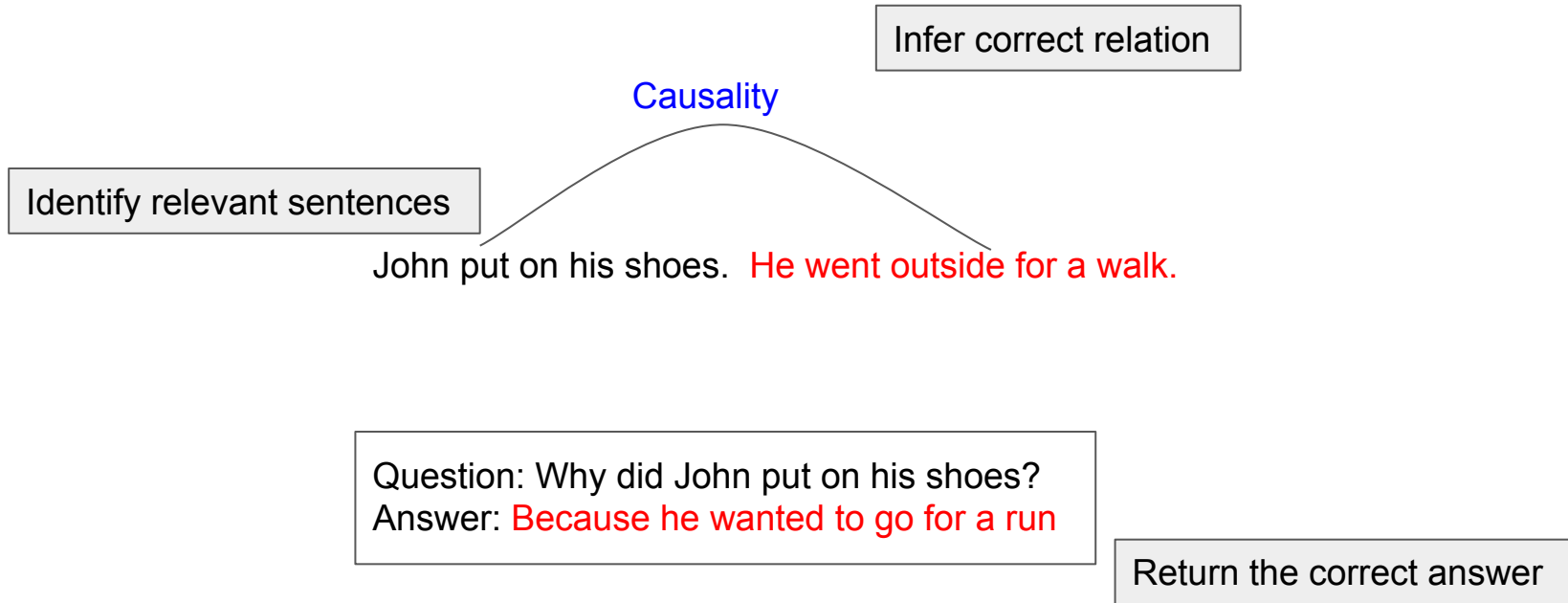


Neural Nets

Convolutional methods, Adversarial models (Biran and McKeown, 2013; Qin et al., 2017)

Unsupervised

Learning a task-specific representations of discourse coherence without annotated or indirect supervision. E.g. machine comprehension(Narassimhan et al 2016):



Discourse Connectives

1. The Mountain View, Calif., company has been receiving 1,000 calls a day about the product **since** it was demonstrated at a computer publishing conference several weeks ago.
2. It was a far safer deal for lenders **since** NWA had a healthier cash flow and more collateral on hand.
3. Domestic car sales have plunged 19% **since** the Big Three ended many of their programs Sept. 30.

Discourse Connectives

1. The Mountain View, Calif., company has been receiving 1,000 calls a day about the product **since** it was demonstrated at a computer publishing conference several weeks ago. (Temporal)
2. It was a far safer deal for lenders **since** NWA had a healthier cash flow and more collateral on hand. (Causal)
3. Domestic car sales have plunged 19% **since** the Big Three ended many of their programs Sept. 30. (Temporal and Causal)

Discourse Connectives

Connectives can be **modified** by adverbs and focus particles:

- *That power can sometimes be abused*, (particularly) since jurists in smaller jurisdictions operate without many of the restraints that serve as corrective measures in urban areas.
- *You can do all this* (even) if you're not a reporter or a researcher or a scholar or a member of Congress.

Implicit/explicit

- ▶ Deduction of implicit information from juxtaposed sentences

It's too far to walk. Let's take the bus.

Infer alternatives: walk/bus as means of transport
Infer causal relation: too far, therefore bus

It's too far to walk **so let's take the bus.**

- ▶ **Assumption:** A passage marks its coherence relation either explicitly or implicitly — i.e., if explicit connective is present, no need for pragmatic inference about additional relations.

It's too far to walk. ^{so?} ~~N~~instead let's take the bus.

Fill-in-the-blank study

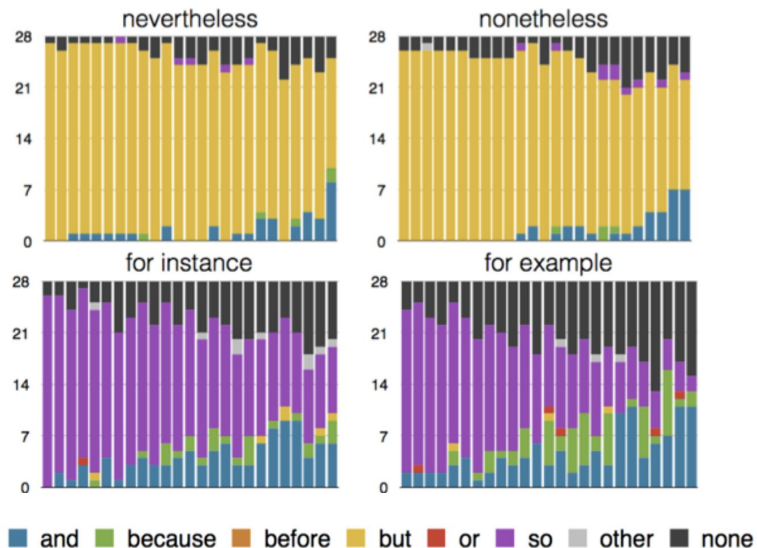
The screenshot shows a web interface for a trial. At the top, there is a blue header with 'ConnText' on the left and 'University of Edinburgh' on the right. Below the header, the word 'Trial' is displayed next to a purple button labeled 'Show Instructions'. The main content area features a grey box with the sentence: 'I don't mind walking // in fact it's good exercise'. Below this, under the heading 'Conjunction:', there is a list of radio button options: Or, But, Because, None at all, So, And, Before, and Other word or phrase. To the right of the options, there is a line of text: 'Once you have made your selections, press submit to complete the trial. To share additional comments about this trial, please [click here](#).' At the bottom right of the form is a blue 'Submit' button.

→ Dataset of judgments for 50 adverbials, each in 50+ passages, each passage judged by 28 people... 70,000+ data points

<http://people.cs.georgetown.edu/nschneid/p/disadv-gurt-slides.pdf>

Implicit passages

- ▶ On one hand, we see some consistency in semantically related adverbial pairs.



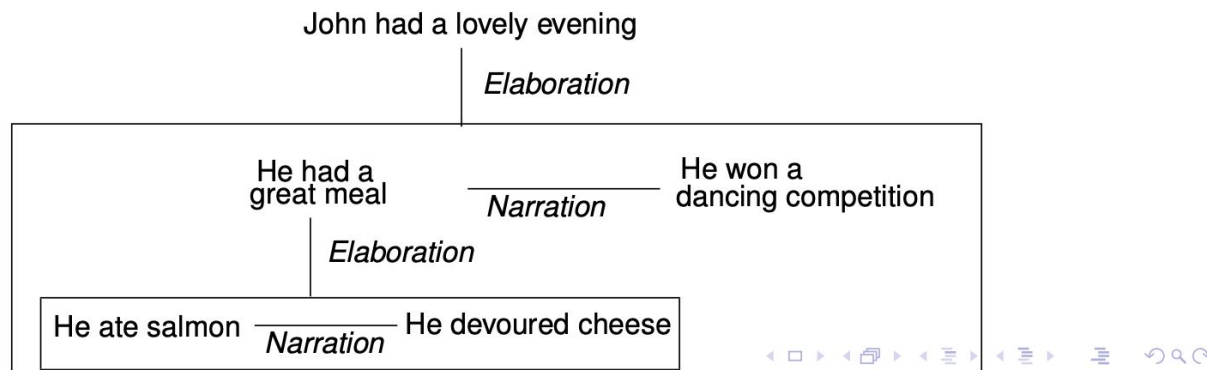
Other Datasets

- RST
 - Theory
 - [Corpus](#)
- SDRT
 - Theory
 - [Settlers of Catan \(STAC\)](#)
- [The GUM Corpus](#)
 - RST framework
 - Reddit
- PDTB v3
- RST for Diagrams
- CITE and CLUE for text and images

Need Rhetorical Relations: Some Motivating Data

Pronouns

- (2)
- a. John had a great evening last night.
 - b. He had a fantastic meal.
 - c. He ate salmon.
 - d. He devoured lots of cheese.
 - e. He won a dancing competition.
 - f. ??It was a beautiful pink.

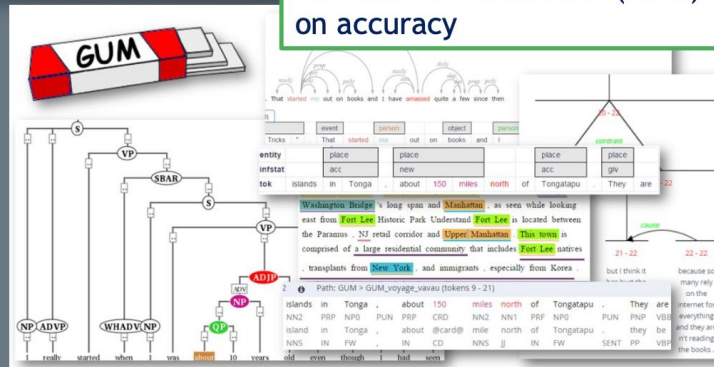


Georgetown University Multilayer corpus

15

See Zeldes & Simonson (2016)
on accuracy

- POS tagging (PTB, CLAWS, TT)
- Sentence type (SPAAC++)
- Document structure (TEI)
- Syntax trees (PTB + Stanford)
- Information status (SFB632)
- (Non-) named entity types
- **Coreference + bridging**
- **Rhetorical Structure Theory**
- Speaker information, ISO time...



<http://corpling.uis.georgetown.edu/gum/>

text type	source	texts	tokens
Interviews (conversational)	Wikinews	19	18037
News (narrative)	Wikinews	21	14093
Travel guides (informative)	Wikivoyage	17	14955
How-tos (instructional)	wikiHow	19	16920
Total		76	64005



CC creative commons

A Multilayer View of Discourse Relation Graphs / A. Zeldes

CMU LTI Colloquium

https://corpling.uis.georgetown.edu/amir/pdf/RST_CMU2017.pdf

3. Discourse Coherence and Visual Explanations

Coherence in Multimodal Explanations

1. Image-text
 - a. Temporal
 - b. Illustration
 - c. Exemplification
 - d. Summary

2. Image-image
 - a. Narration
 - b. T-constraint

Coherence in Multimodal Explanations

Image-text

a. **Temporal:** temporal links between text and image

- Inclusion: text describes a process and the picture gives us a moment in the process.
- Result: The image illustrates the result of the action that is described in the text.



1- Score a small x at the end of each peach with a paring knife.

4- Using paring knife, remove strips of loosened peel, starting at X on base of each peach.

Coherence in Multimodal Explanations

Image-text

b. Illustration relation: relations from part of the description to a particular image region. E.g. small X, ice water.



1- Score a small x at the end of each peach with a paring knife.

3- Transfer peaches immediately to the ice water and let cool for 1 minute.

Coherence in Multimodal Explanations

Image-text

- c. **Exemplification:** visual information often shows just one case of a generalization presented in accompanying text.



1- Score a small x at the end of each peach with a paring knife.

Coherence in Multimodal Explanations

Image-text

- d. **Summary:** utterances summarize the information that should have been visible.(cite)



1- Score a small x at the end of each peach with a paring knife.



2- Lower peaches into the boiling water and simmer until skin loosen, 30 to 60 seconds.



3- Transfer peaches immediately to the ice water and let cool for 1 minute.



4- Using paring knife, remove strips of loosened peel, starting at X on base of each peach.

Coherence in Multimodal Explanations

Image-image

- a. **Narration:** Sequence of images describe sequence of actions in a temporal manner.



Coherence in Multimodal Explanations

Image-image

- b. **T-constraint:** Image 1 zooms in on the scene while image 2 pans slightly to the right to show all the peaches. (cite)



Why studying coherence in images and text is challenging?

Identifying coherence relations in text mainly relies on textual cues. These cues are missing in the image-text presentations.

I missed my meeting today **because** my car broke down.

Data Collection: CLUE

10,000 image–text pairs annotated by expert annotators with a high agreement.

Data Collection: CLUE

10,000 image–text pairs annotated by expert annotators with a high agreement.

- ▶ 5,000 from Conceptual Captions (Sharma et al., 2018)
- ▶ 5,000 from machine-authored captions from the state of the art models in 2019



Daily **Mail**.com

alamy



gettyimages

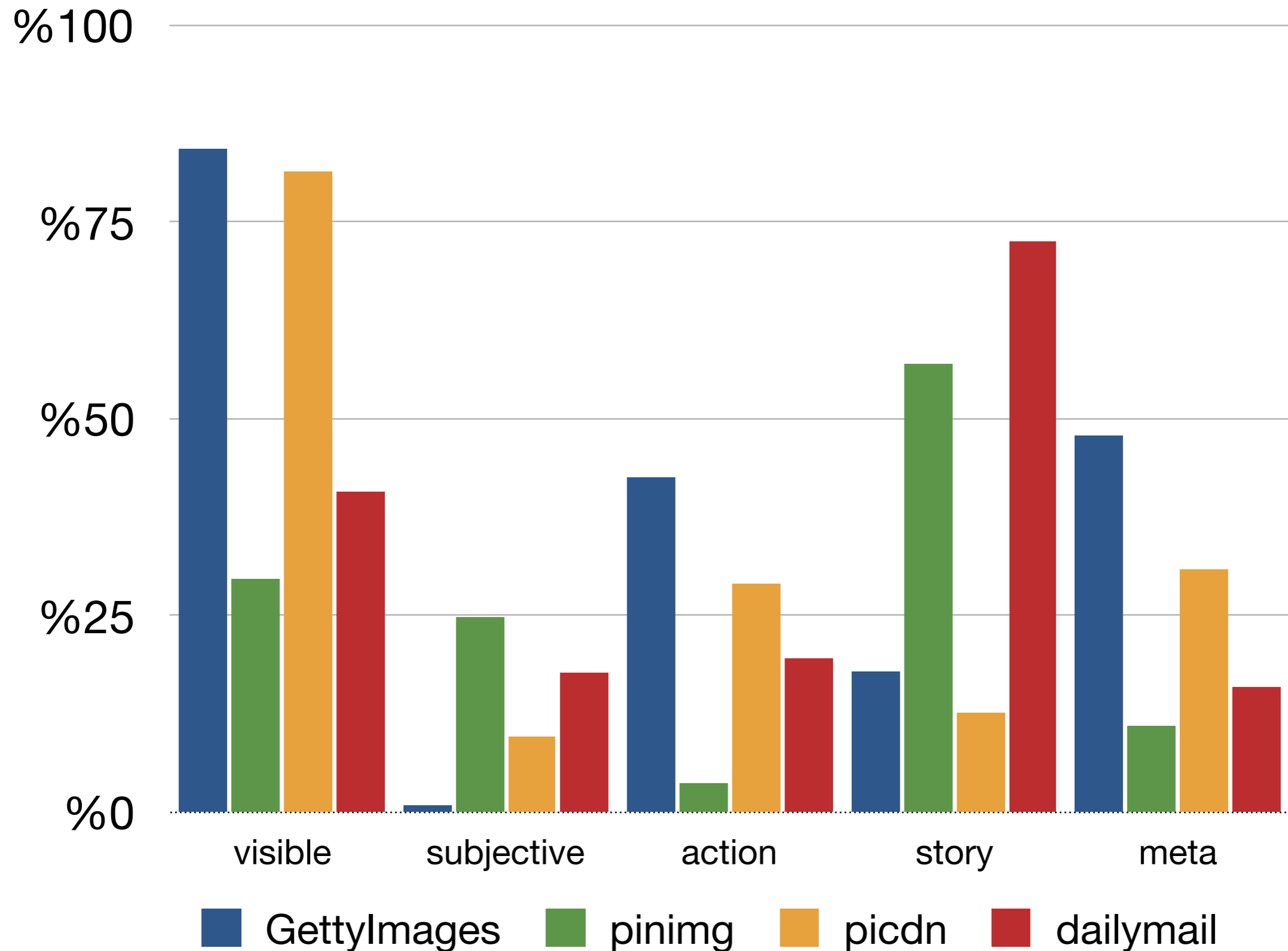
Story

The text is understood as providing a free-standing description of the circumstances depicted in the image.

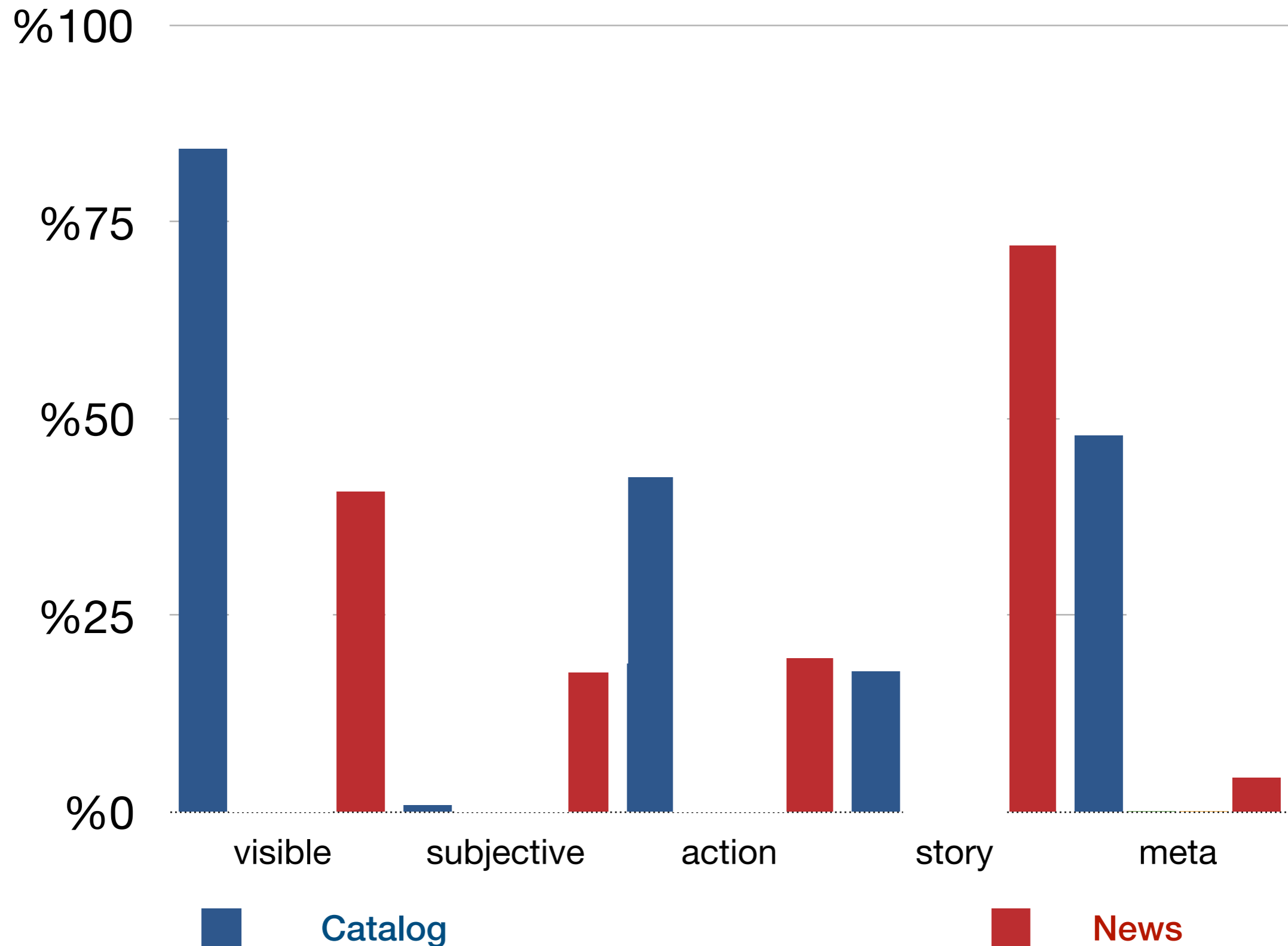


Room of a Nobel prize winner, looted during the revolution.

Coherence relations predict genre!



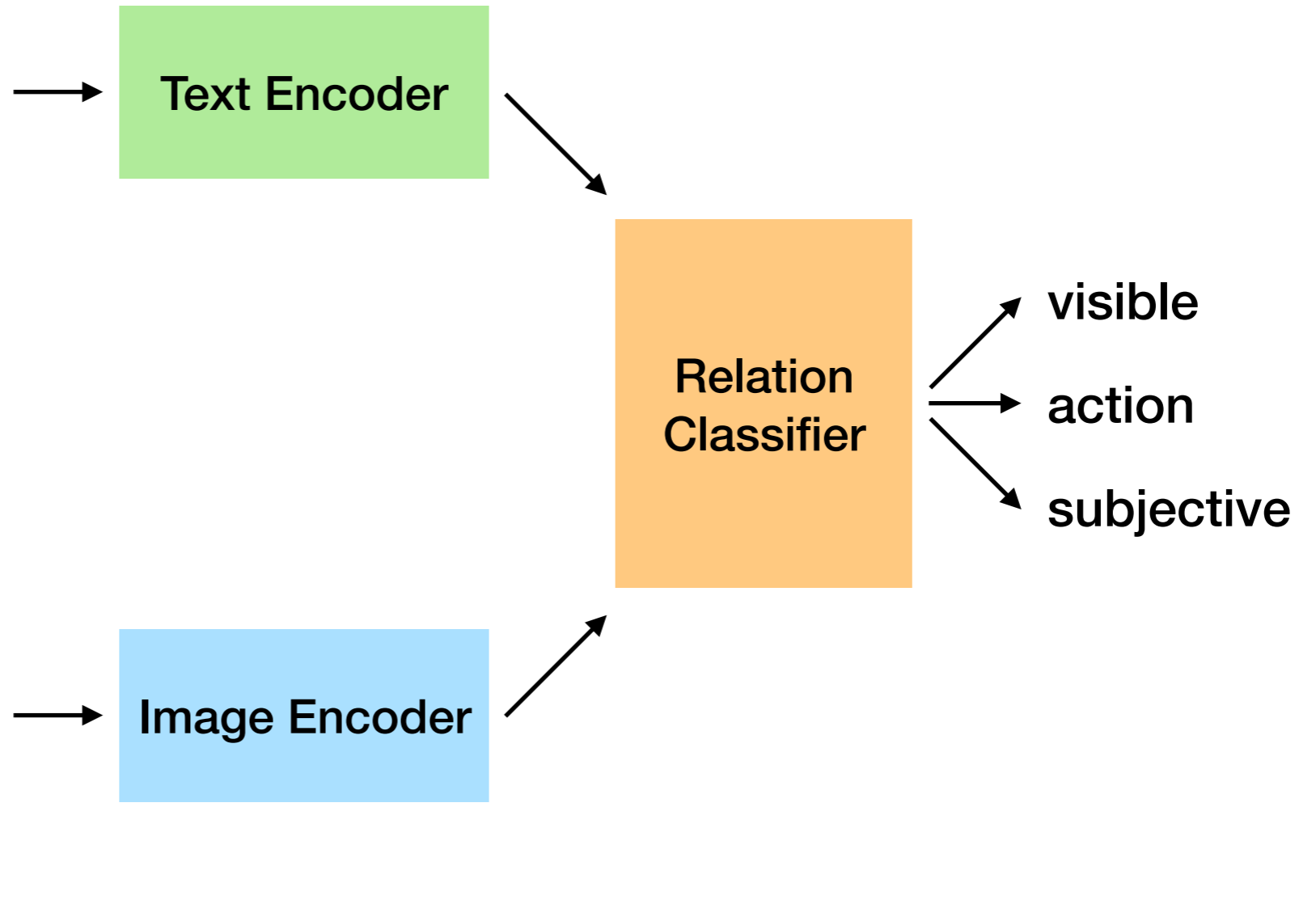
Coherence relations predict genre!



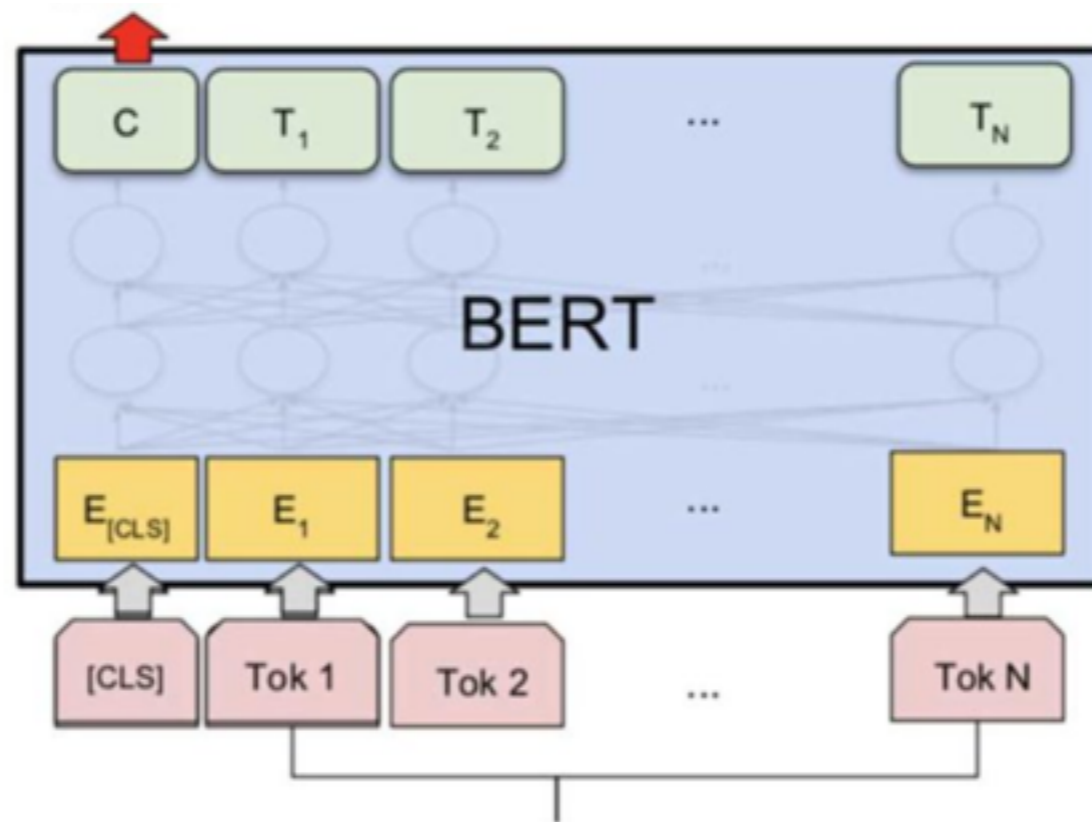
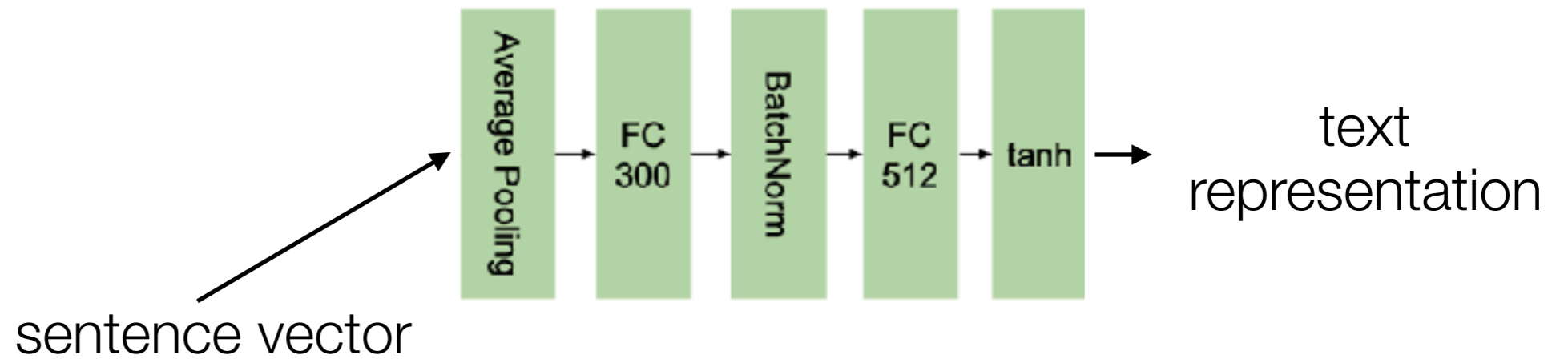
Can we learn to predict these relations?

Predicting Relations

Young happy boy swimming
in the lake.

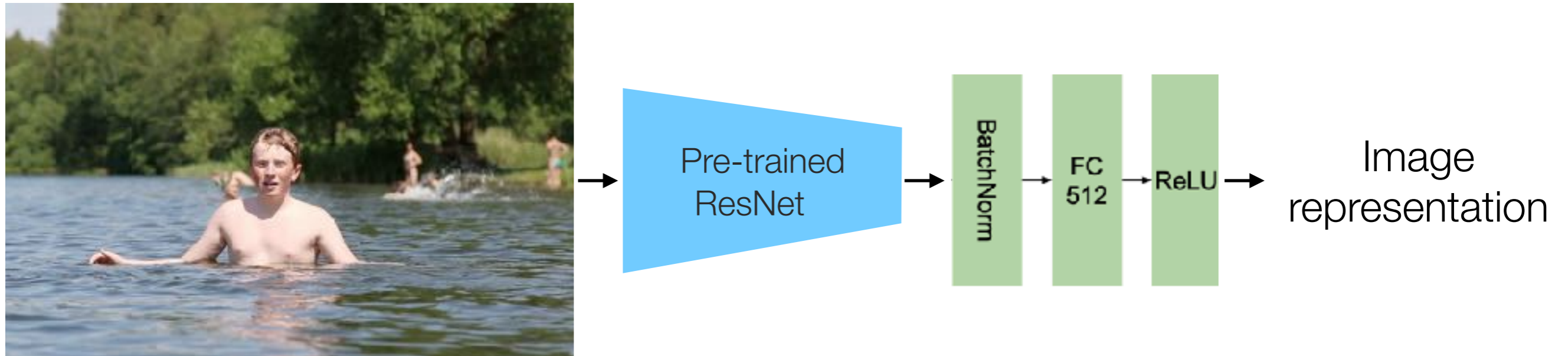


Text Encoder

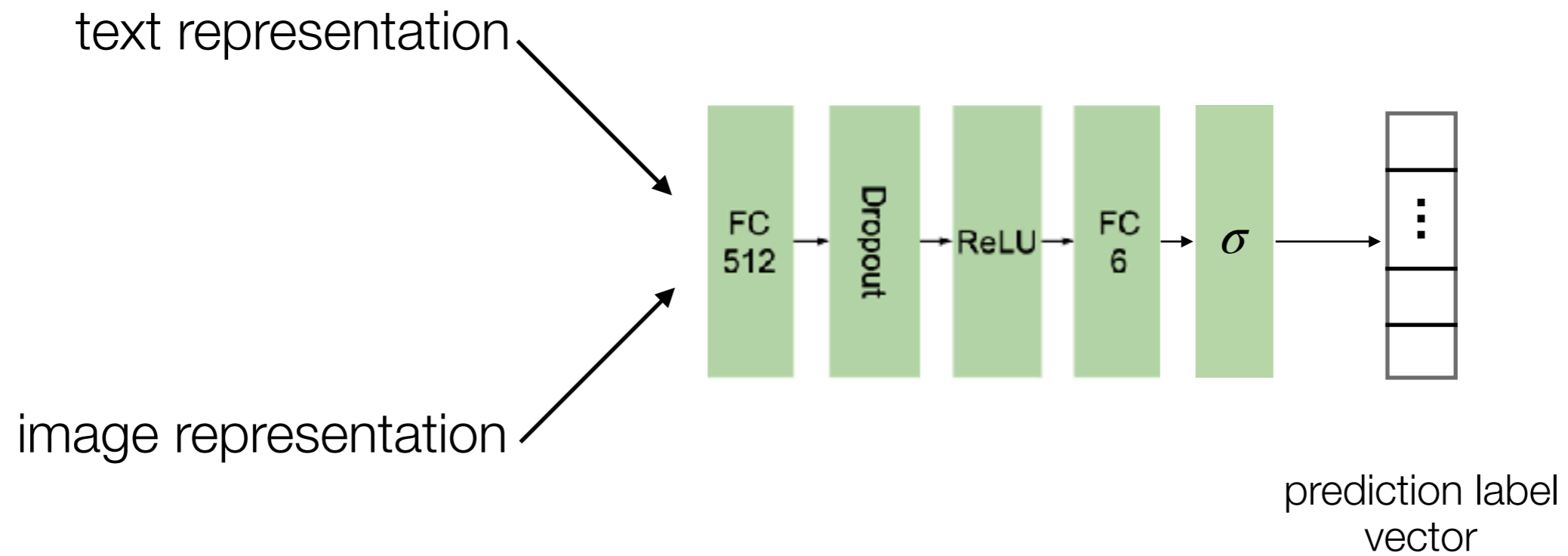


Young happy boy swimming in the lake.

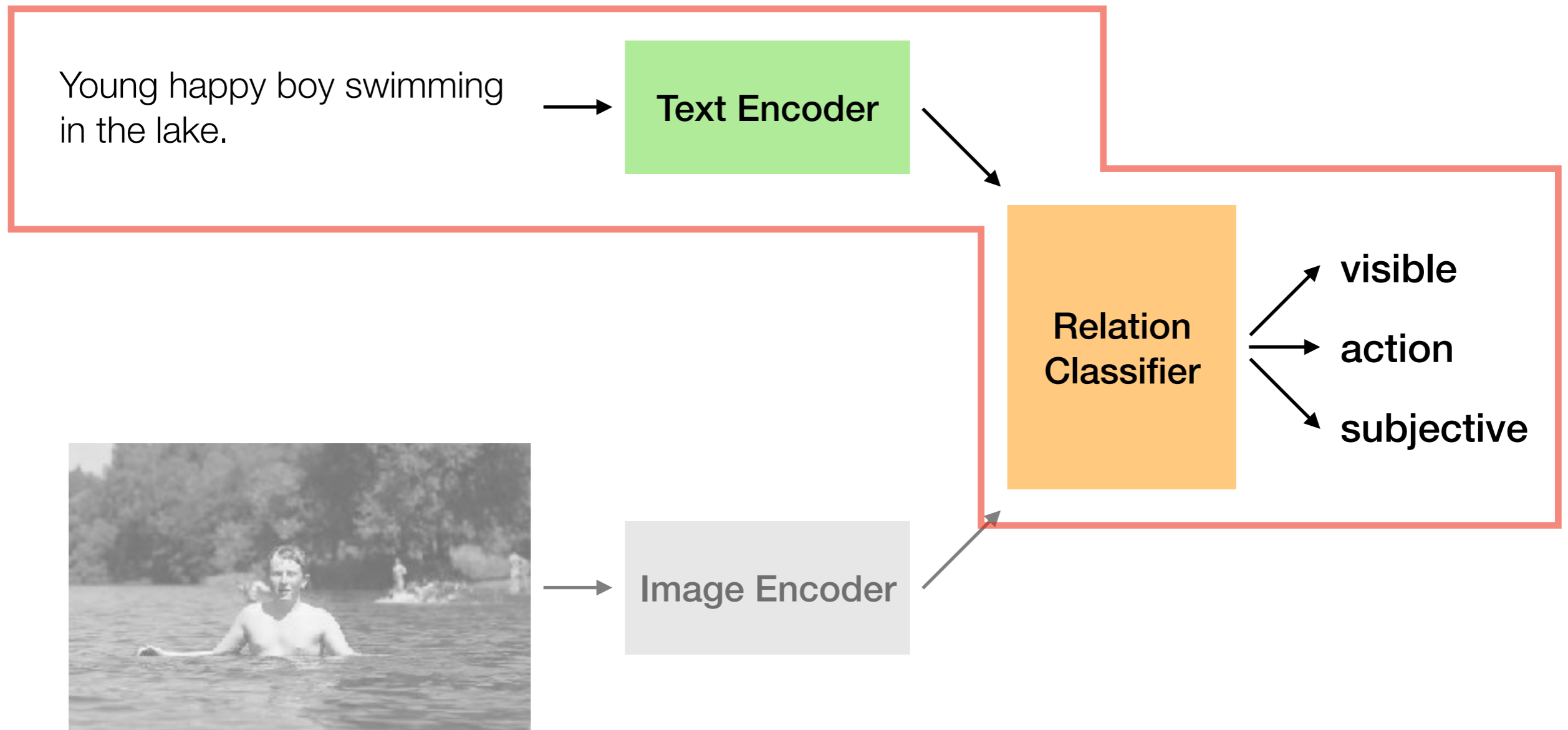
Image Encoder



Relation Classifier

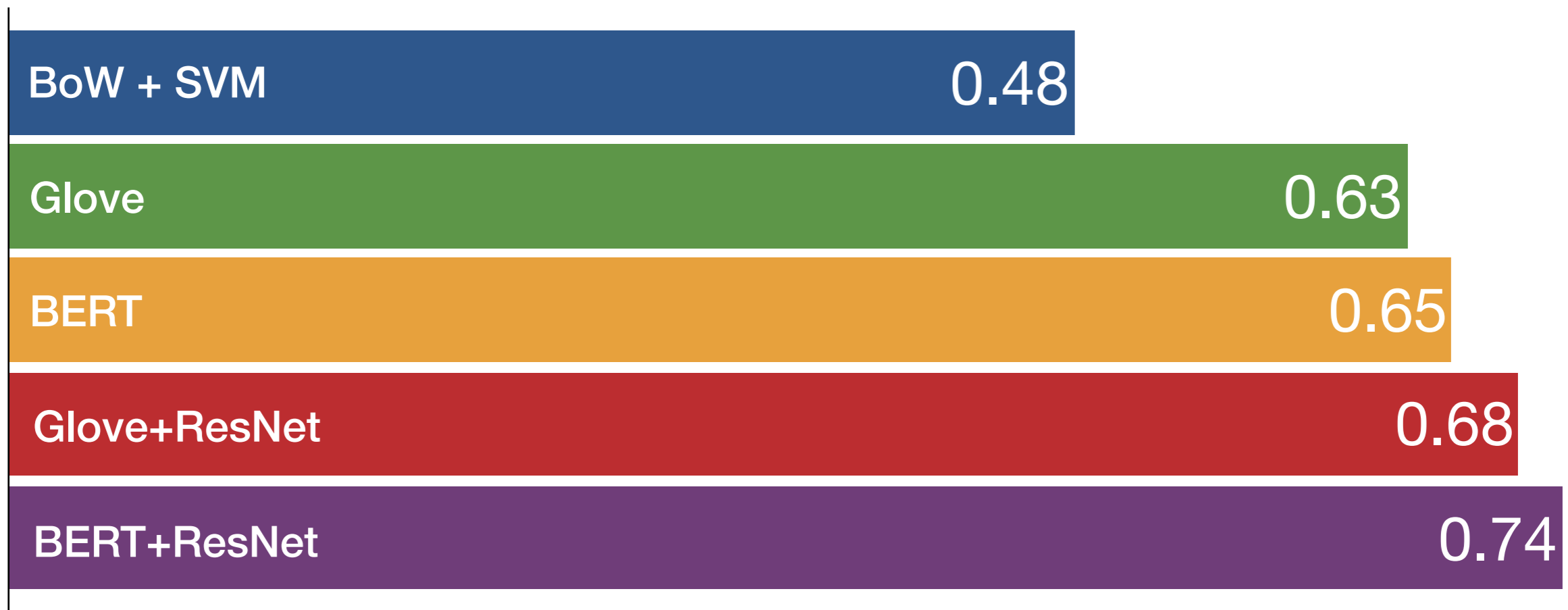


Predicting Relations



Predicting Relations

Machine learning models can reliably predict coherence relations.



Weighted F1 scores: Multi-label

Outline

- Introduction
- Connecting Text and Imagery
 - Coherence modeling in images and text
 - Temporal and logical inferences
 - Linguistic structure
 - Generating informed descriptions for images
- Multimodal Decisions in Conversational AI
- Judging the Intent of Pointing Actions with Robotic Arms
- Discussion and Conclusion

Data Collection

Instead of starting from a predefined taxonomy, we study the key elements of logical, temporal and elaboration-like relations.



Crowdsourcing Experiment

2,047 Image-text pairs annotated with a high rate of inter-rater agreement



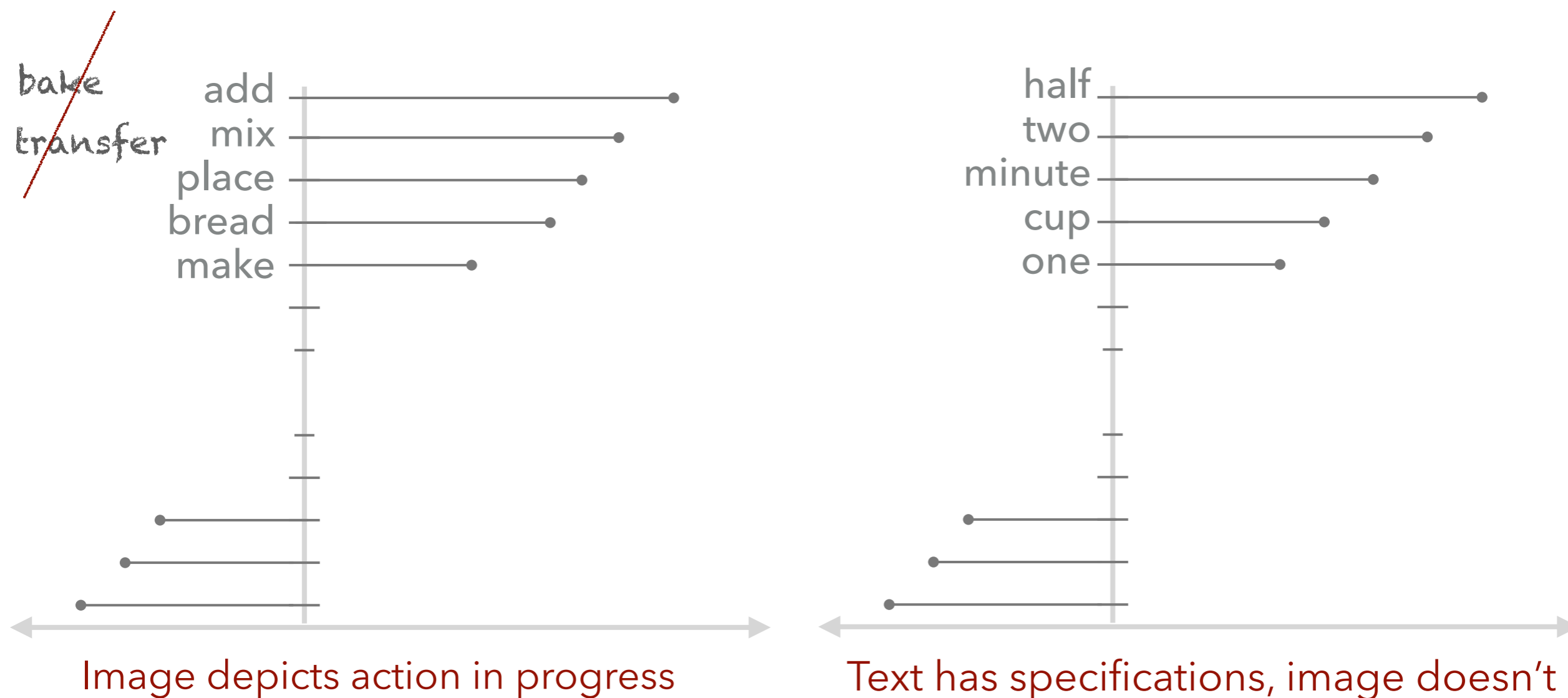
Top with half the spinach half the mozzarella and a third of the remaining sauce.

- Image shows action in progress.*
- Action in image need repetition.*
- Image shows tools not in the text.*
- Text has quantities, image doesn't.*
- and 6 more.*

Highlight the part of the text that is most related to the image.

Information Across Modalities

Classifiers can learn textual cues that signal the contributions of each mode to discourse.



Top Naive Bayes Features

Outline

- Introduction
- Connecting Text and Imagery
 - Coherence modeling in images and text
 - Temporal and logical inferences
 - Linguistic structure
 - Generating informed descriptions for images
- Multimodal Decisions in Conversational AI
- Judging the Intent of Pointing Actions with Robotic Arms
- Discussion and Conclusion

What distinguishes captions from other descriptions?



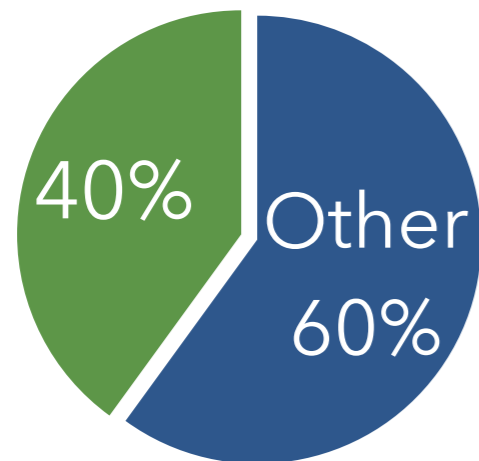
A man is sitting in front
of a bunch of fruits.

By Carol Mitchell

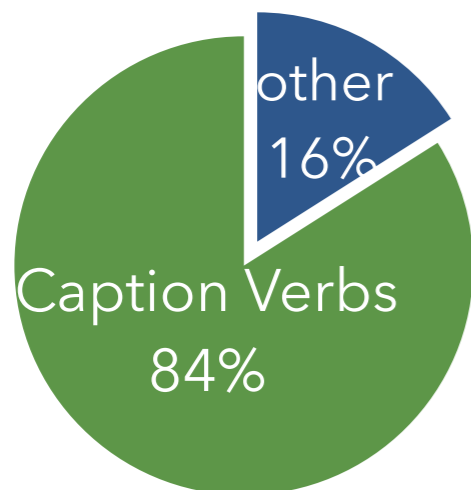
What distinguishes captions from other descriptions?

Captions show a distinctively limited distribution of verbs, with strong preferences for specific tense, aspect, and lexical aspect.

Caption verbs



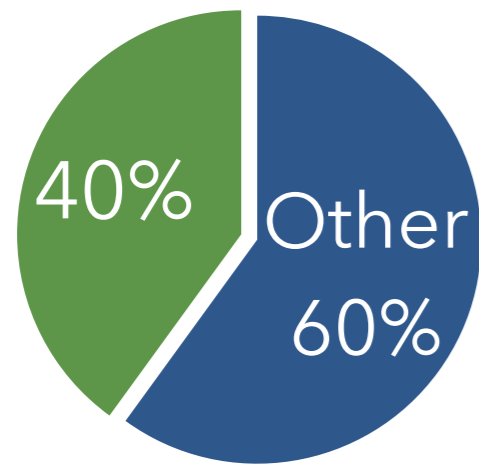
Corpora



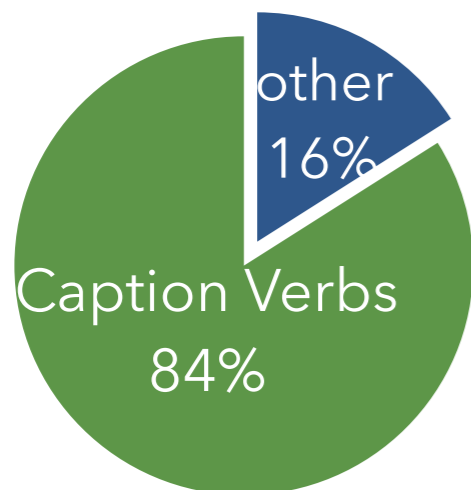
Machine-authored

What distinguishes captions from other descriptions?

Captions show a distinctively limited distribution of verbs, with strong preferences for specific tense, aspect, and lexical aspect.



Corpora



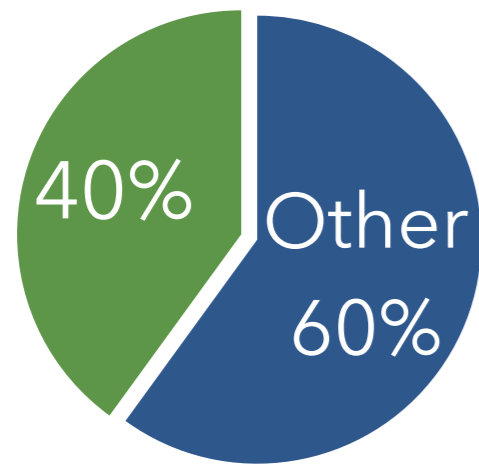
Machine-authored

Lexical aspect

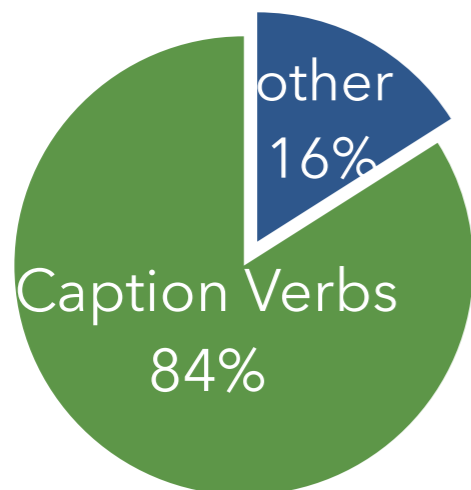
- ▶ A man is running in the park. (Atelic)
- ▶ A woman arrived at a party. (Telic)

What distinguishes captions from other descriptions?

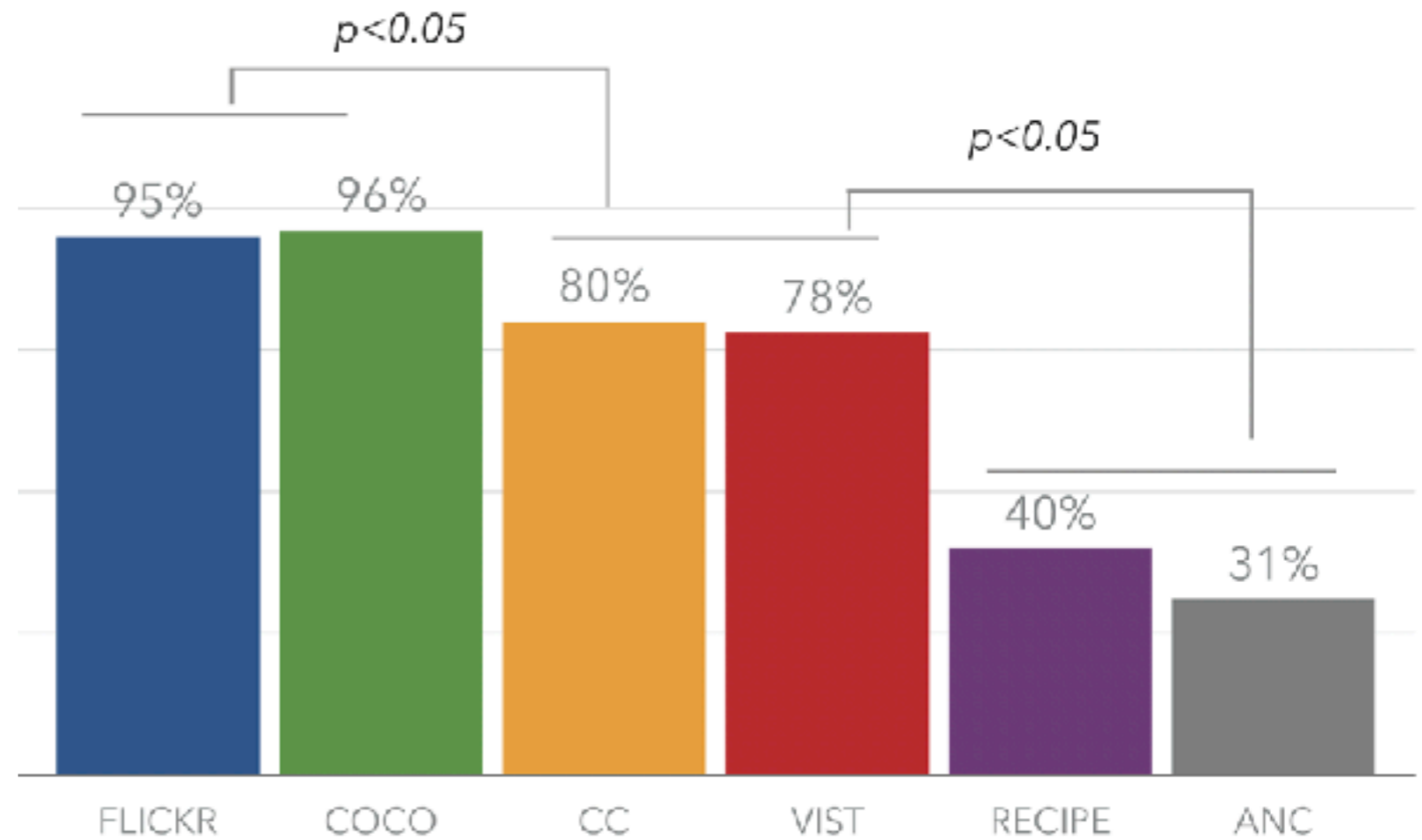
Captions show a distinctively limited distribution of verbs, with strong preferences for specific tense, aspect, and lexical aspect.



Corpora



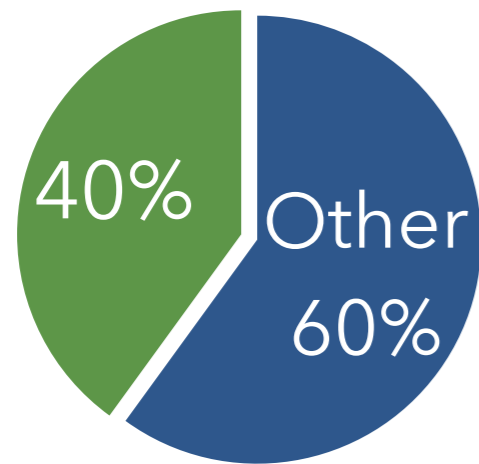
Machine-authored



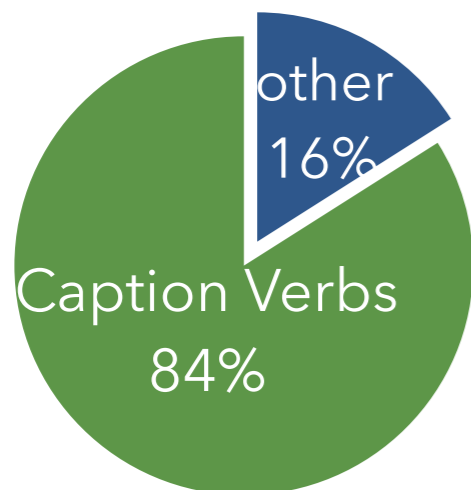
Captions describe indefinite temporal events.

What distinguishes captions from other descriptions?

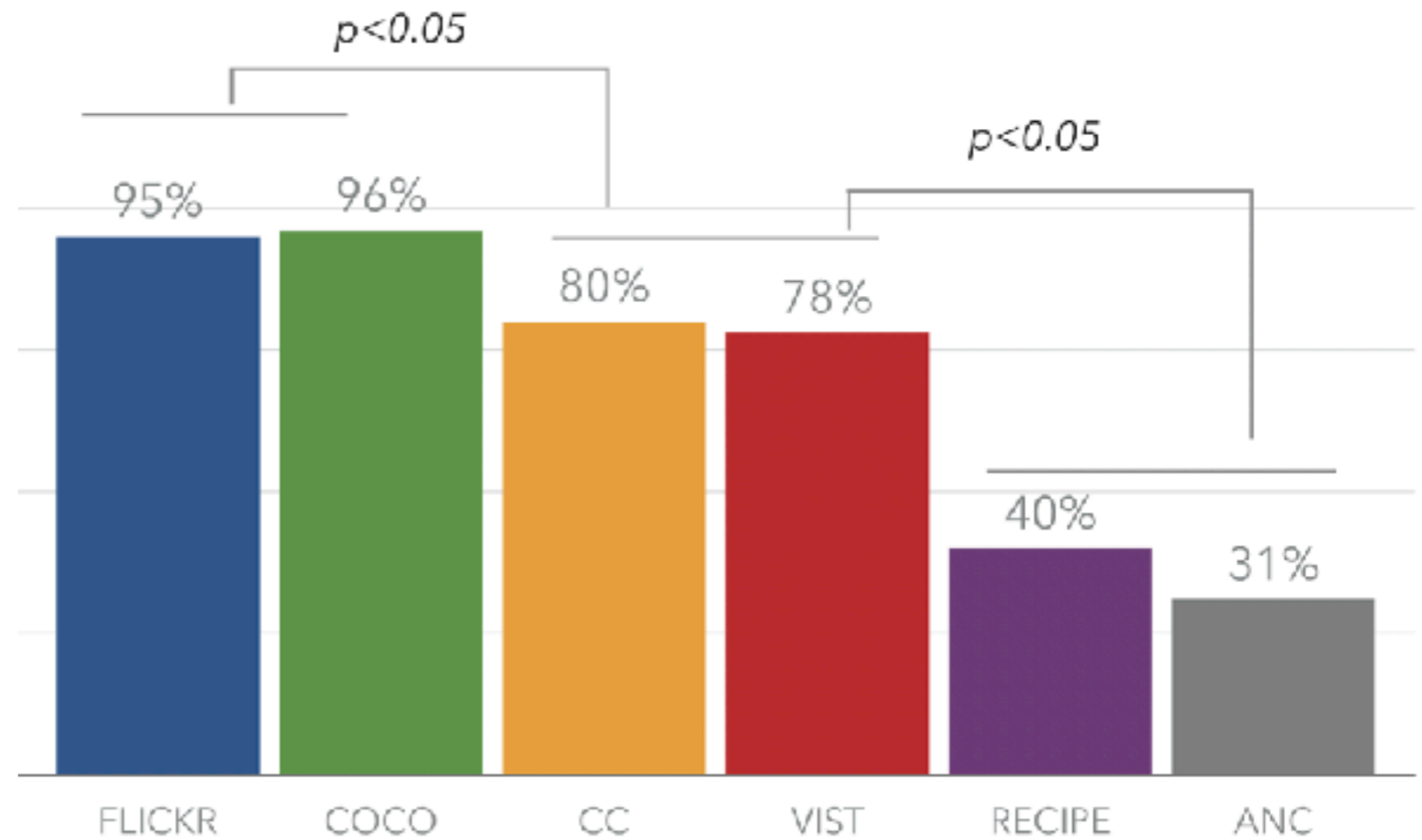
Captions show a distinctively limited distribution of verbs, with strong preferences for specific tense, aspect, and lexical aspect.



Corpora



Machine-authored



Captions describe indefinite temporal events.

Aspectuality Across Genre: A Distributional Semantics Approach

- ▶ We show that two elementary dimensions of aspectual class, states vs. events, and telic vs. atelic events, can be modelled effectively with distributional semantics.
- ▶ We contribute a dataset of human–human conversations annotated with lexical aspect



Arabic:

فتاة	تتحدث	على	الهاتف
girl	talk-PRS-FEM-IPFV-3SG	on	phone

A girl is talking on the phone.

Chinese:

她	正在	用	手机	通话
She	PRS	use	phone	talk

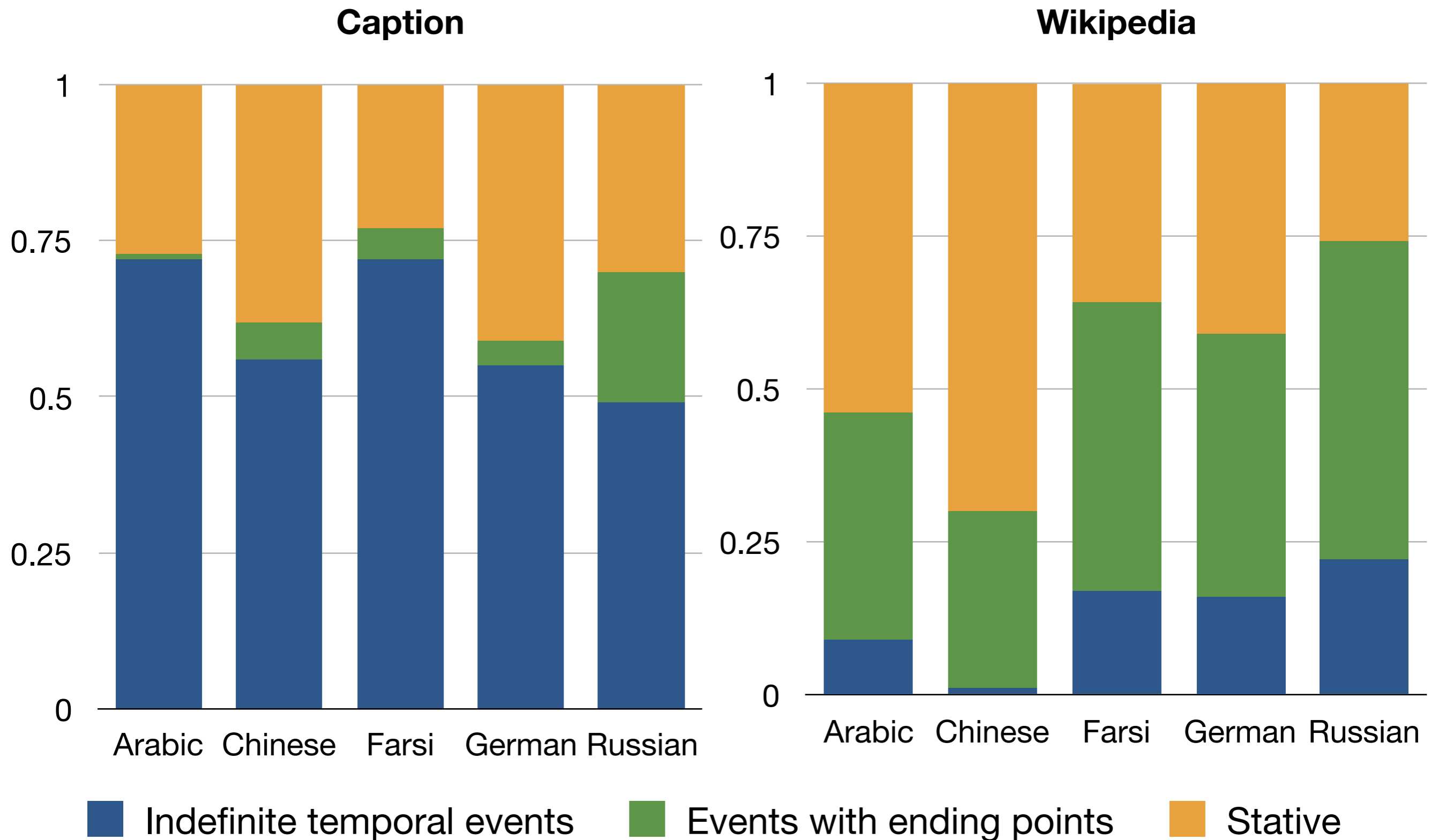
She is using a phone to talk

Farsi:

یک	نفر	با	تلفن	صحبت می کند.
One	person	with	telephone	conversation-do-PRS-IPFV-3SG

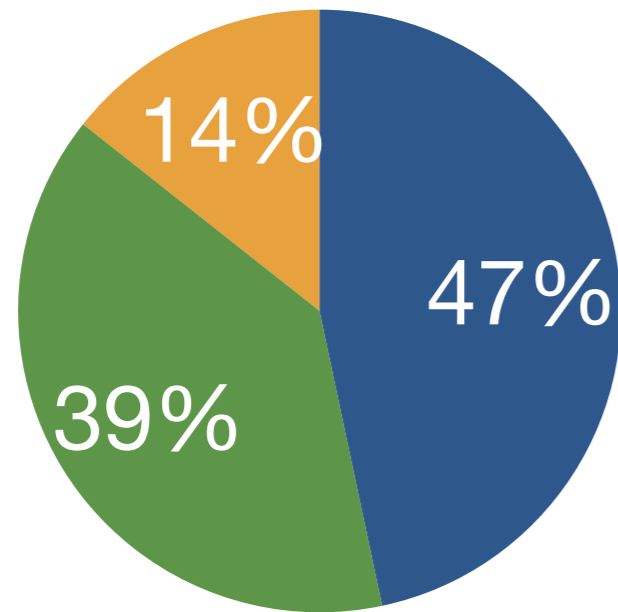
A person is talking with the phone.

Biases in Event Types–Beyond English

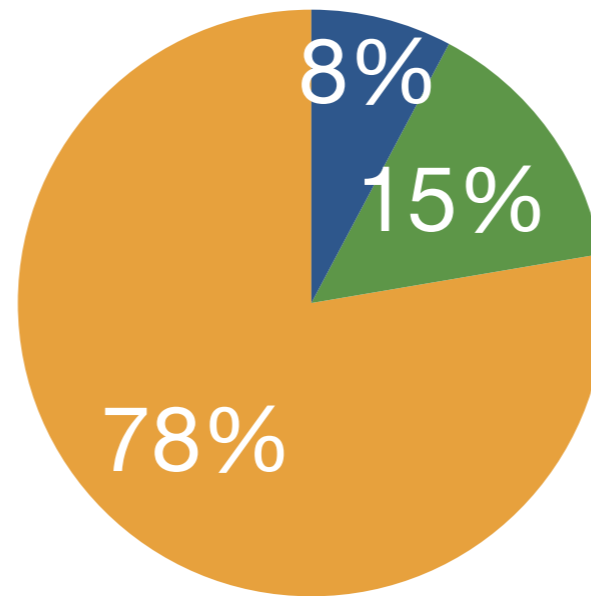


Gender across Languages

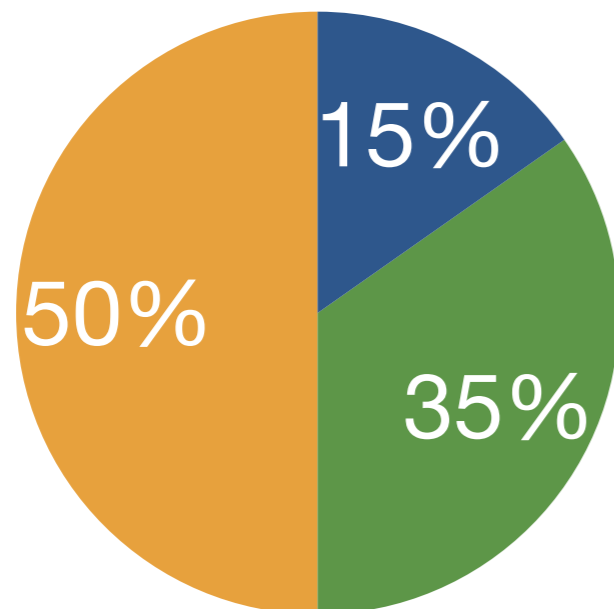
Arabic



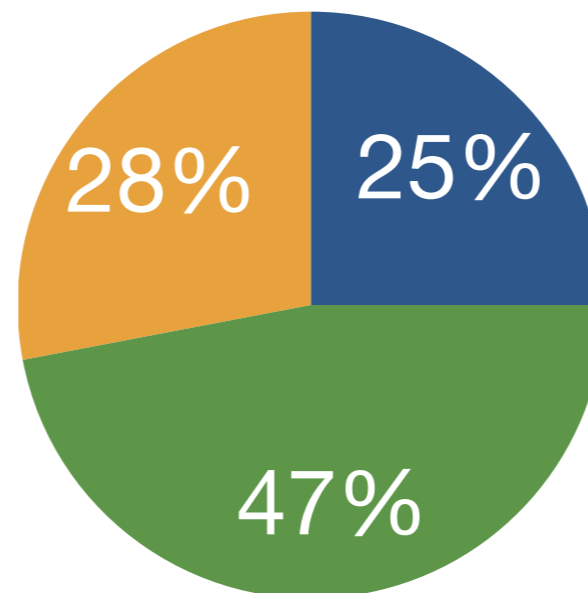
Chinese



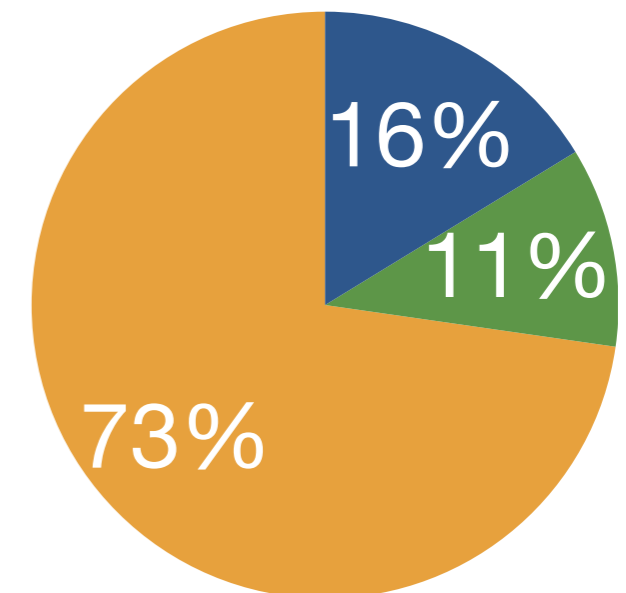
Farsi



German



Russian



Outline

- Introduction
- Connecting Text and Imagery
 - Coherence modeling in images and text
 - Temporal and logical inferences
 - Linguistic structure
 - Generating informed descriptions for images
- Multimodal Decisions in Conversational AI
- Judging the Intent of Pointing Actions with Robotic Arms
- Discussion and Conclusion

Coherence Modeling for Description Generation



Photo credit: Blue Destiny / Alamy Stock Photo

Visible: horse and rider jumping a fence.

Meta: horse and rider jumping a fence during a race.

Subjective: the most beautiful horse in the world.

Story: horse competes in the event.

Coherence Modeling for Description Generation



Photo credit: Blue Destiny / Alamy Stock Photo

Visible: horse and rider jumping a fence.

Meta: horse and rider jumping a fence during a race.

Subjective: the most beautiful horse in the world.

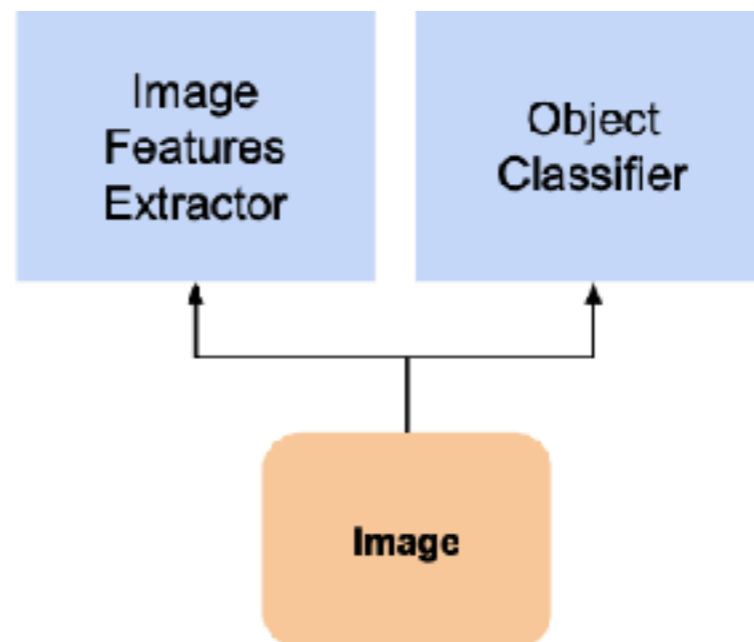
Story: horse competes in the event.

- ▶ Training data: 3.3 million image-text pairs annotated with our model.
- ▶ Features of the model:
 - ▶ Transformer-based generation model
 - ▶ Image features: Graph-RISE
 - ▶ Detected objects: Google Cloud Vision API

Proposed Architecture

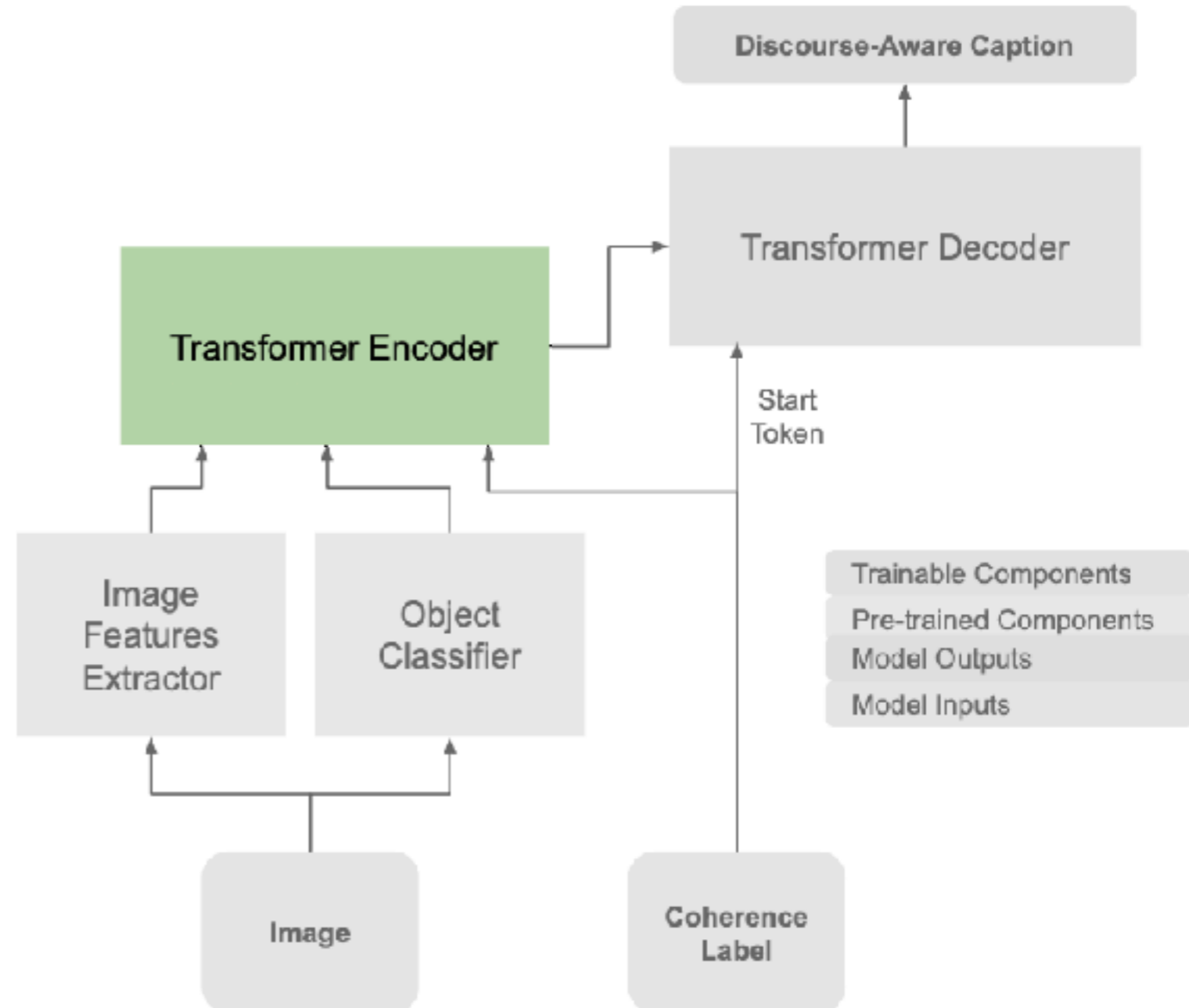
ResNet-101 network classifies images into $O(40M)$ classes.

Pre-trained 512-dimensional vectors trained to predict objects.



Proposed Architecture

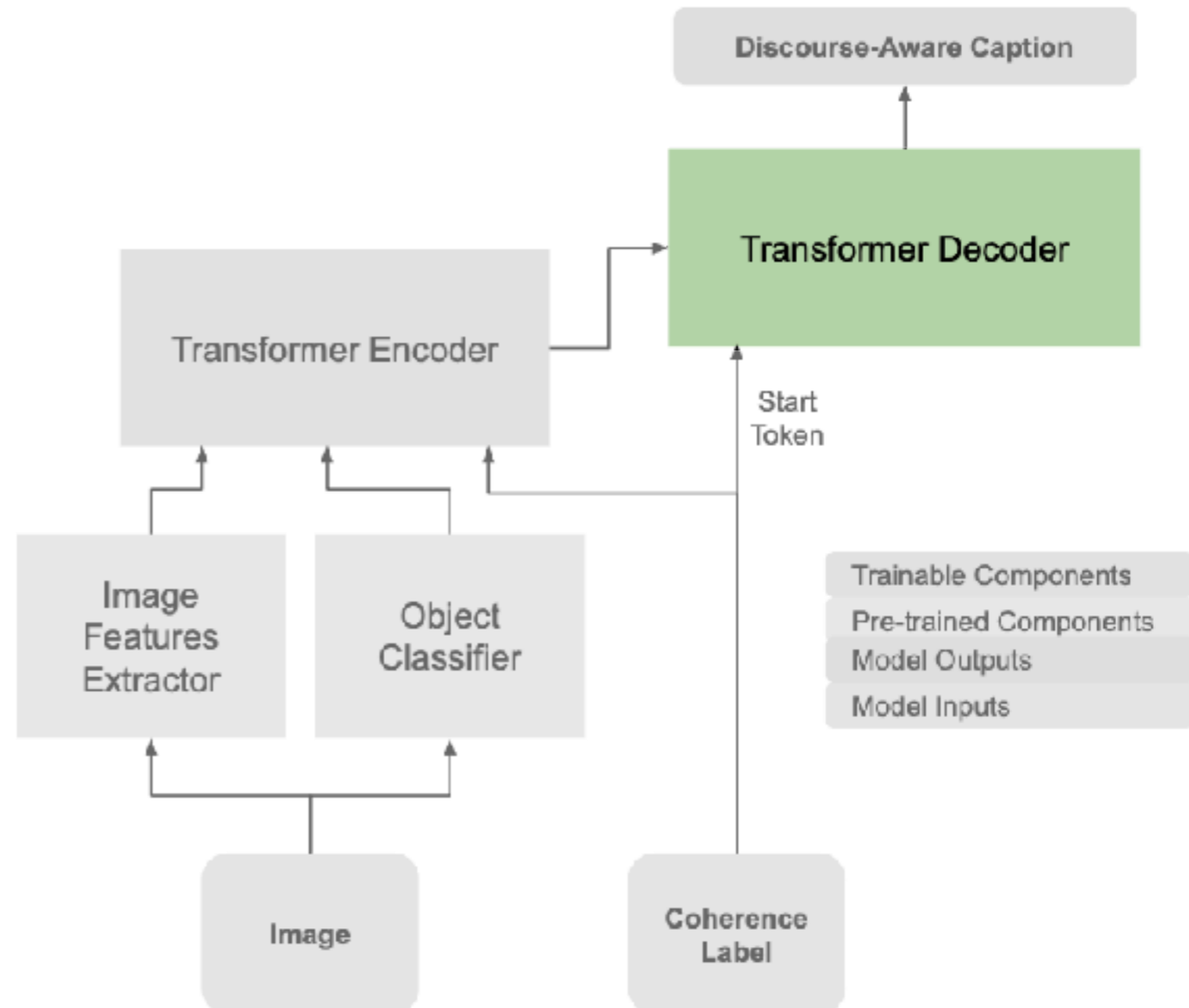
Changing context independent-embeddings to context-dependent embeddings



Proposed Architecture

Changing context independent-embeddings to context-dependent embeddings

Controlling the input of the decoder

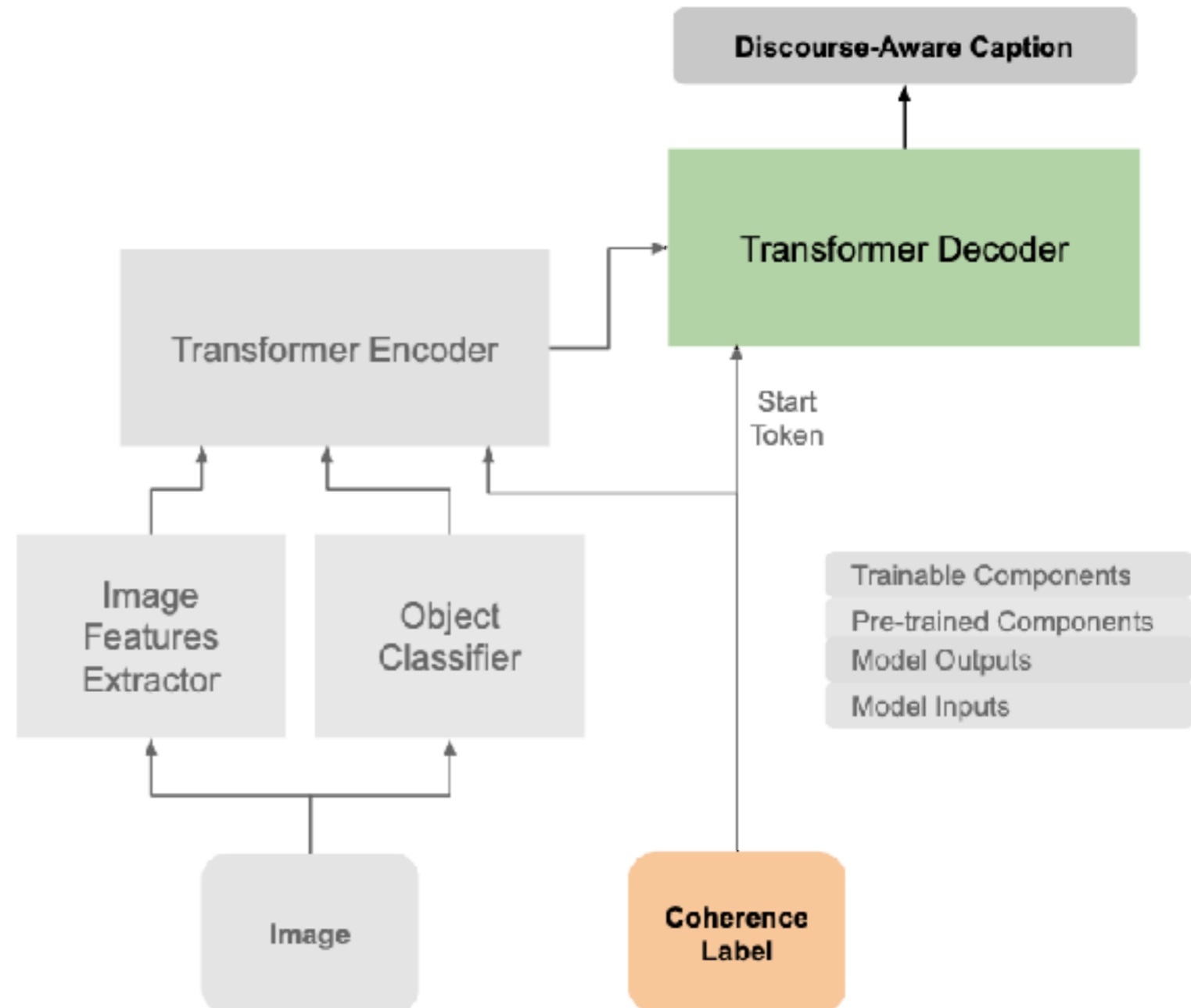


Proposed Architecture

Changing context independent-embeddings to context-dependent embeddings

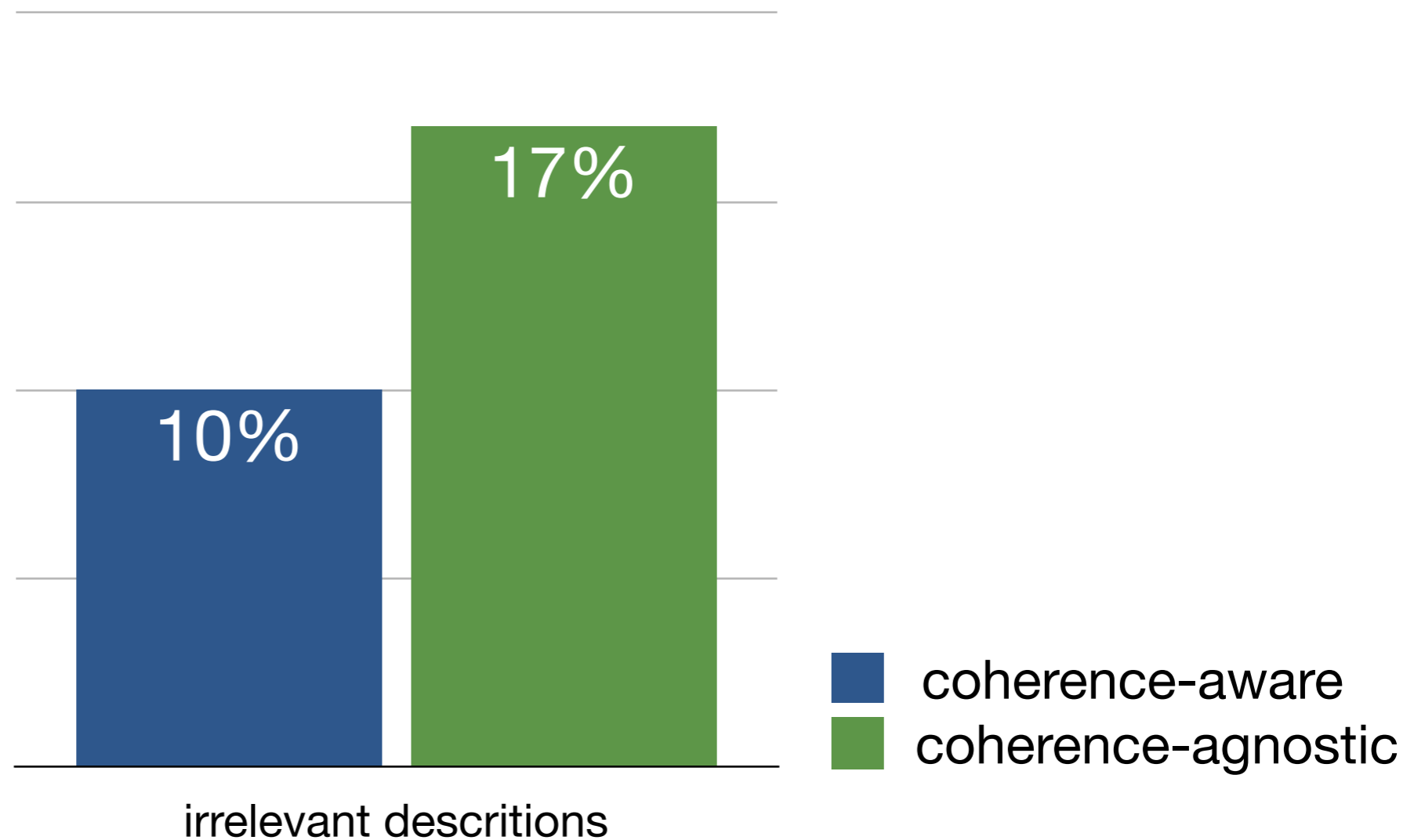
Controlling the input of the decoder

Generating controlled captions



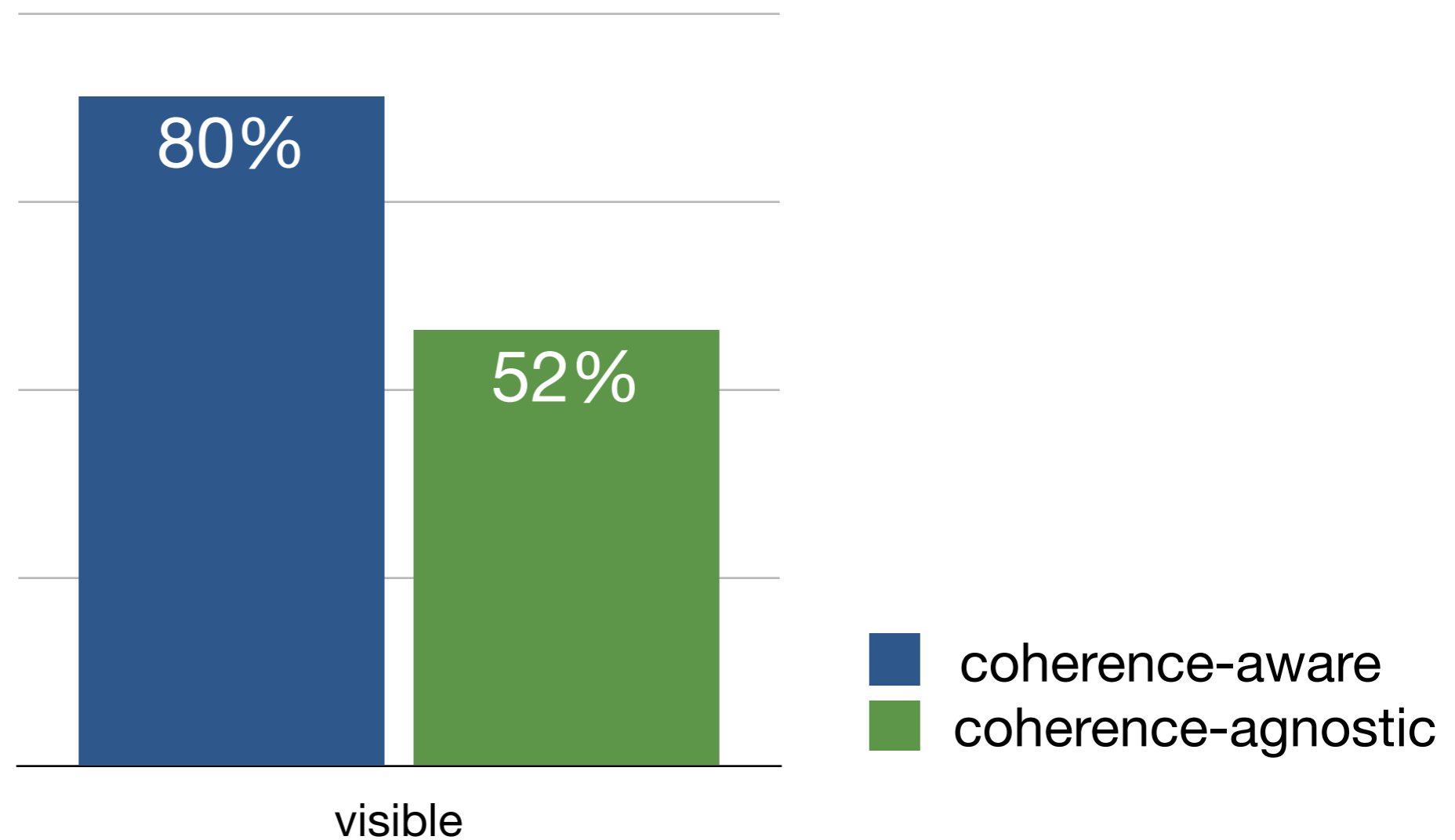
Results

Overall, the quality of the generated descriptions have improved.



Results-Visible

The rate of (non-overlapping) visible descriptions goes up.





Coherence-aware Visible: the pizza at restaurant
Coherence-agnostic: The best pizza in the world



Coherence-aware Story: How to spend a day.
Coherence-agnostic: Dogs playing on the beach.

Evaluation-Crowdsourcing

Rates of “Good” visible captions

- ▶ Coherence aware: 86%
- ▶ Coherence agnostic: 74%
- ▶ State of the art models in 2019: 67%

Preference

- ▶ 68.2% prefer captions generated by the coherence-aware model versus 31.8% prefer captions generated by the coherence-agnostic model

Evaluation-Crowdsourcing

The average scores of the “Quality” of the visible captions on a scale of 0 to 5

- ▶ Coherence aware: 3.44
- ▶ Coherence agnostic: 2.83

The average scores of the “Relevance” of the visible captions on a scale of 0 to 5

- ▶ Coherence aware: 4.43
- ▶ Coherence agnostic: 4.40

Evaluation-Automatic Metrics

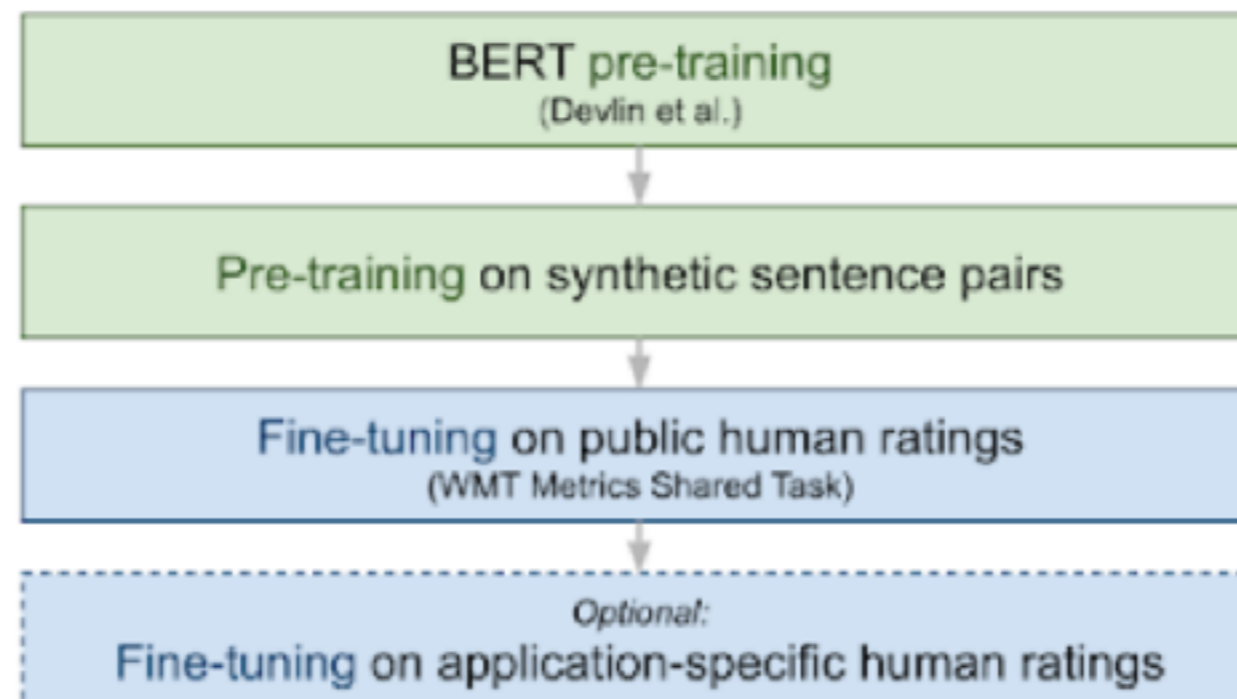
CIDEr scores

- ▶ Coherence aware: 0.958
- ▶ Coherence agnostic: 0.964

Reference-driven automatic generation metrics do not reflect these differences.

Discourse-aware BEURT

- ▶ We introduce a discourse-aware learned generation metric.
- ▶ Our proposed metric calculates different scores with respect to the goal of the task.



Related work

None of previous approaches attempt to characterize **information-level inferences between images and text**. They have focused on

- ▶ Th contrasts across style and genre (Guo et al., 2019).
- ▶ The content of text and imagery as complement (Vempala and Preotiuc-Pietro, 2019).
- ▶ Directing attention and engaging perceptual reasoning (Kruk et al. 2019).
- ▶ Eliciting emotion (Shuster et al. 2019).

Part 2: Grounding

Robot Learning Through Language Communication



Robots Need to Say No!



Challenges of effective multimodal dialogue

Complex domains, goals and tasks

- Substantial uncertainty (and possibility for error) moment by moment

Rich ways to interact with users

- Systems can interact more effectively by doing reasoning and planning

Natural conversation is the model for the interaction

- Systems need to build on how people interact with one another

Key question: how can you build systems that reason under uncertainty to interact naturally

Answer: enable systems to achieve common ground

Grounding in the sense of Herb Clark

Common ground: defined as mutual information that underwrites coordination

- Family of related notions throughout cognitive science of language use,
Including mutual knowledge and mutual belief
Key Figures: Herb Clark, Barbara Grosz, David Lewis, Craige Roberts, Robert Stalnaker

P is common ground if

- There is evidence E that shows (sufficient for current purposes) that:
P is true
All participants in the conversation have evidence E

How people achieve common ground

A dynamic, interactive process

- Seeking and providing evidence of mutual understanding
Clark & Marshall (1981), Clark & Wilkes-Gibbs (1986), Clark & Schaefer (1989), Clark (1996).

Process includes

- Making contributions
Acknowledging and accepting others' contributions
Asking for and providing clarification
Demonstrating understanding
Detecting and repairing misunderstanding and other errors

Results in information becoming common ground, so interlocutors can rely on it going forward

How people achieve common ground

Using all the communicative resources available in face-to-face conversation

- Taking a turn and offering a follow-up verbal utterance
- Lightweight spoken feedback (back-channel utterances)
- Head nods and other facial displays
- Hand gestures and other communicative body movements
- Eye gaze and attention
- Actions in pursuit of task goals

Achieving common ground in dialogue systems

Three complementary approaches:

- Shaping and coordinating conversational interaction, so interlocutors naturally get evidence that makes conversational state common ground.
- Exhibiting and tracking the kinds of behavior people use in grounding, so users' natural grounding behavior helps make the system more robust.
- Making moment-by-moment decisions about sources of uncertainty in conversation, in order to assess what is common ground and what is not, and react accordingly.

Each of these approaches can help a system create common ground

- Approaches are best used in combination, but this is still rare in multimodal dialogue.

Grounding – An overloaded term

Not symbol grounding (e.g., Harnad 1990)

- Capabilities (such as linking symbols to perceptual classifiers) that ensure that computer representations have intrinsic meaning.
- Symbol grounding for linguistic meanings is an important research area (“grounding words perception and action” Deb Roy TICS 2005)

Grounding – An overloaded term

Can lead to confusion

- Multimodal conversational systems typically need to do both:
 - Ground word meaning in perception, to understand and generate situated utterances
 - Make meanings common ground, to achieve good outcomes with human users
- Processes are not independent
 - Learn word meanings by making instances of unfamiliar concepts common ground
 - Give symbols intrinsic meanings through strategies that ensure understanding is shared

Remember there are two separate but related kinds of grounding.

Grounding – Not limited to dialogue

Grounding is a useful perspective for analyzing all HCI artifacts

- Do systems show how they interpret user input?
- Do systems make sure users recognize and understand system contributions?
- In short, do systems display sufficient information and feedback to users
So that system meets its obligations with respect to collaboration and common ground?

Brennan (1998). Grounding with and through computers.

- Remains significant design perspective (e.g., Coactive design, Johnson and colleagues, Journal of Human Robot Interaction, 2014)
- Won't consider it further here

This Tutorial

Reviews research related to grounding in dialogue systems

Three key takeaways

- Grounding is key to understanding human-human conversation
 - Helps to organize and explain diverse and frequent phenomena
- Grounding is key to effective system design and implementation
 - Learning to monitor user understanding
 - Using human-like behaviors to demonstrate understanding
 - Using human-like behaviors to demonstrate non-understanding and obtain clarification
 - Eliciting natural feedback from users
 - Bootstrapping improved system capabilities
- Understanding grounding gives key insights into dialogue tasks and architectures
 - What research advances grounding capabilities, and why?

For ACL audience

Grounding separates true conversational agents from command-and-control and chat systems

- Grounding is a key mechanism for robust and flexible interaction
- Grounding is a key test of systems' competence in dialogue as an interaction modality
- Grounding demands deep engagement with human-centered approaches to communication and collaboration

For ACL audience

Review of landmark techniques, systems and results from leaders in NLP,
But often published outside NLP, at IJCAI, AAAI, ICMI, AAMAS, IVA, HRI, etc.

- Dan Bohus, Microsoft
- Justine Cassell, CMU
- Joyce Chai, Michigan
- David DeVault, USC
- Raquel Fernandez, Amsterdam
- Jonathan Ginzburg, Paris
- Oliver Lemon, Heriot-Watt
- Verena Rieser, Heriot-Watt
- David Schlangen, Potsdam
- Candy Sidner, WPI
- David Traum, USC

For ACL audience

Contextualized with fundamentals and recent advances from across disciplines

- Cognitive science
Empirical accounts and computational models of human-human conversation
- Human-Computer Interaction
Design principles and evaluation methods for creating usable systems
- AI
Techniques for perception, diagnosis, planning and learning

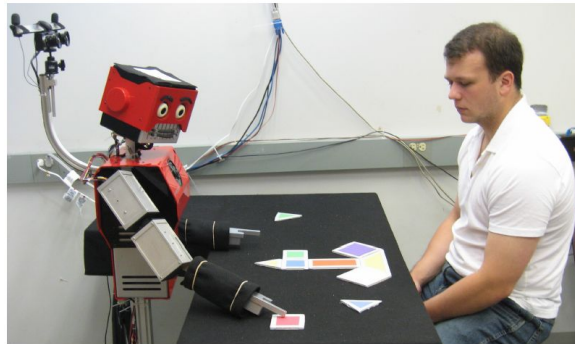
Many exciting NLP research opportunities on the horizon thanks to progress in these areas

Achieving common ground in dialogue systems

Shaping and coordinating conversational interaction: **Orchestrating engagement.**

- Sidner, Lee, Kidd, Lesh and Rich, AIJ 2005. Holroyd, Rich, Sidner and Ponsler, ROMAN 2011. Bohus and Horvitz, Sigdial 2009. Bohus, Saw and Horvitz, AAMAS 2014.

Make things common ground **directly**, by ensuring that system and users maintain **joint attention**



Engagement with Melvin, from Holroyd et al (2011).

Achieving common ground in dialogue systems

Exhibiting and tracking natural behavior in grounding: **Information-state update approaches.**

- Rickel and Traum, AAMAS 2002.
Nakano, Reinstein, Stocky and Cassell, ACL 2003.
Mehlmann, Janowski, Häring, Baur, Gebhard and André. ICMI 2014.

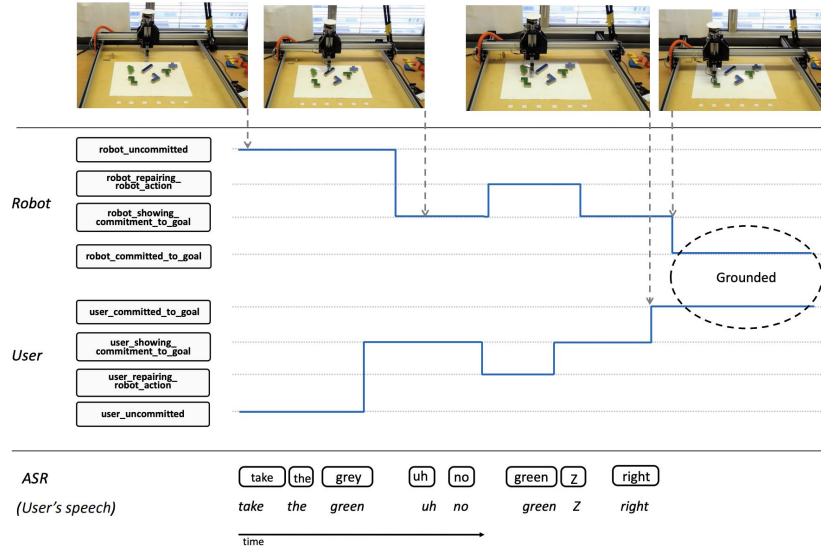


MACK updates content
as grounded based on
user gaze behavior
(Nakano et al 2003).

Achieving common ground in dialogue systems

Making **moment-by-moment decisions** about sources of uncertainty in conversation.

- Hough and Schlangen, HRI 2017



Hough and Schlangen's (2017) Robot uses incremental processing to decide on actions early and recognize that this makes its commitments manifest.

A distinctive multi-modal challenge: Joint attention

The ability to intentionally coordinate attention with another agent

- Dyadic joint attention - interlocutors coordinate attention towards one another
- Triadic joint attention - interlocutors coordinate attention towards a common object

Carpenter, Nagell, Tomasello, Butterworth and Moore (1998). *Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age*.

Joint attention as a complex skill

Direct others' attention

- Pointing, manipulation and other embodied actions
- Referring expressions and other linguistic actions

Follow others' attention

- Track others' gaze and pointing
- Seek out and attend to referents of referring expressions

Maintain split attention

- Attend to object and register visual information about it
- Attend to interlocutor and track their gaze

Joint attention and common ground

Joint attention intrinsically gives access to mutual information

- If interlocutors jointly attend to visual information P, this situation makes P common ground.

This is the canonical real-world basis for common ground in conversation

- Compare the physical copresence heuristic of Clark and Marshall (1981)
- Note that even this is the result of active coordination among interlocutors

Orchestrating Engagement

Orchestrating joint attention

Need to distinguish **function (joint attention)** from **behavior (gaze)**

Human user may use gaze for

- Turn taking
- Hand-eye coordination
- Scanning the visual environment
- Acknowledging those who are not in the conversation

See Cassell (2000), Human conversation as a system framework.

Orchestrating joint attention

Need to distinguish **function (joint attention)** from **behavior (gaze)**

In generating appropriate system gaze behavior

- System may have multiple layers of control proposing gaze behavior
- What behavior should take precedence?
- What will user infer about system attention and grounding from selected behavior?

Engagement for orchestrating joint attention

Definition of engagement

- The process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume or terminate an interaction.
Definition from Bohus and Horvitz (2009)
- Typically conceived as a longer-term state that's robust to temporary shifts in attention (for turn-taking, multi-party interaction, and visual search)

See also Glas and Pelachaud ACHI 2015

- Note that some other definitions and accounts of engagement are one-sided
- Note that some other definitions of engagement also capture interest and positive affect

Process based implementations of engagement

Sidner, Lee, Kidd, Lesh and Rich, AIJ 2005. Holroyd, Rich, Sidner and Ponsler, ROMAN 2011.

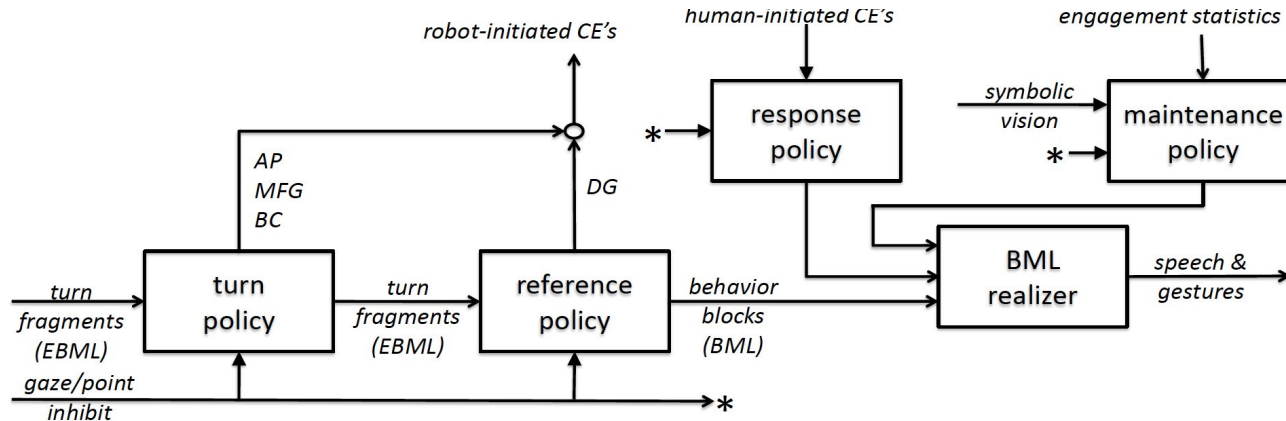
- Models dynamics of engagement in terms of connection events
 - Directed gaze (triadic joint attention)
 - Mutual facial gaze (dyadic joint attention)
 - Adjacency pair (contingent action across parties, across turns)
 - Backchannel (contingent action across parties, within turns)
- Recognizes connection events from user
- Generates connection events on robot

Engagement achieves grounding because system signals and adapts to achieve joint attention

Process based models of engagement

Generator needs to initiate connections, respond to them, and maintain contact

- Multiple functions for the same behaviors
- Realized in a typical robotic layered control architecture



Integrating connection behaviors in Holroyd et al (2011)

Contribution of engagement to successful dialogue

Evaluations suggest that engagement succeeds in making interaction state common ground

Note common methodology: interactions with system, assessed for

- Subjective measures (user judgment of understanding, believability, fluency)
- Overall dialogue features (task success, number of turns)
- Distribution of target behaviors (task actions, gaze, utterances, nonverbal feedback)

Contribution of engagement to successful dialogue

Findings

- Nonverbal signals of conversation state make dialogue more efficient, make users judge conversation more smooth and lifelike (e.g., Cassell and Thorisson, Applied AI 1999).
- System elicits more nonverbal back-channels (nods) if system recognizes and reciprocates (e.g., Sidner, Lee, Morency and Forlines, HRI 2006).
- Revealing uncertainty about engagement (through hesitation and pauses) makes disengagement smoother (e.g., Bohus and Horvitz, ICMI 2014)
- Nonverbal cues to turn-taking elicit appropriate responses from users, Provided system produces and tracks them incrementally (e.g., Skantze, Hjalmarsson, Oertel, Speech Communication 2014).

Limitations of engagement strategies

It's easy to build a system that keeps the user talking

- Give verbal and nonverbal feedback a human would use to show they have understood and are ready for the next installment.
- Such systems can be quite successful and effective (e.g., DeVault et al, AAMAS 2014).

It's harder to give and understand feedback as meaningful

- Use it as evidence about understanding or non-understanding

This requires a different approach, which we turn to next.

Information-state update for multi-modal grounding

Multi-modal grounding in a directions kiosk



Attention as evidence of understanding

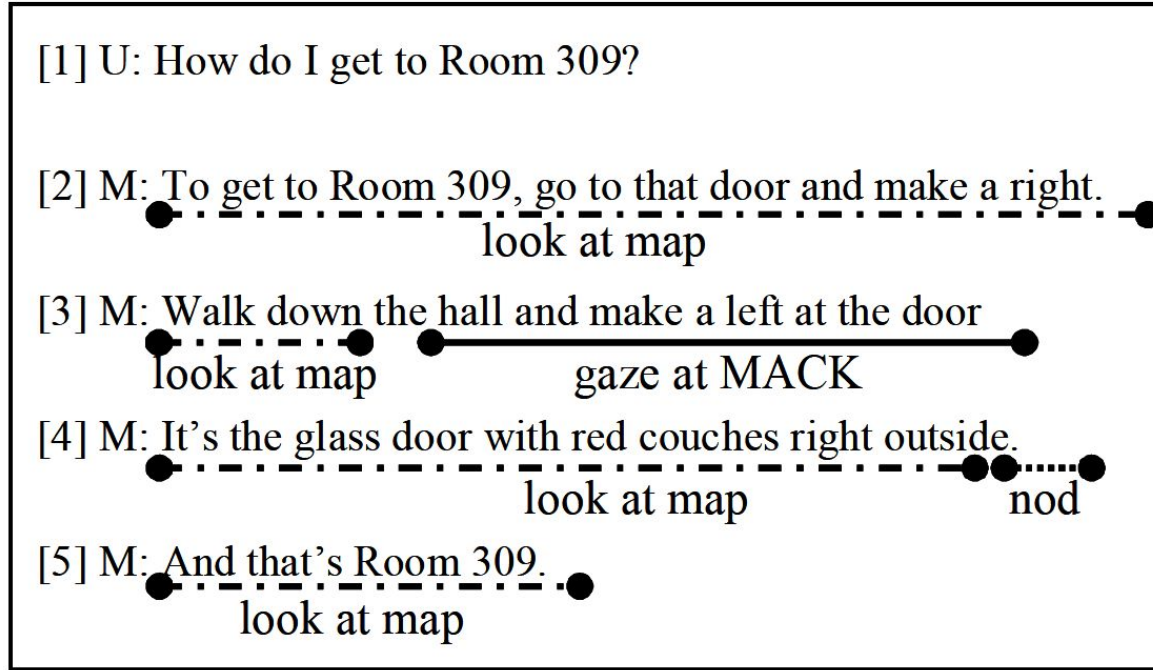


Figure 4: Example of user (U) interacting with MACK (M). User gives negative evidence of grounding in [3], so MACK elaborates [4].

Attention as evidence of grounding

Task-specific meaning for joint attention based on human-human data

- When direction-follower gazes at end of turn in directions,
Human direction-giver's next turn is typically an elaboration (73% of the time).
- When direction-follower gazes at the map at end of turn in directions,
Human direction-giver's next turn is typically next instruction step (52% of the time)

System implements this interpretation for user gaze:

- If user gazes at map at end of direction turn,
System interprets the user as acknowledging and confirming system instructions:
Instructions are grounded and system moves on
- If user gazes at system at end of direction turn
System interprets user as signaling a problem:
System elaborates, with goal that user can confirm after additional information.

Implementation and evaluation

Grounding is tracked using information state update (ISU) approach

- Reviewed in detail in Session 3 of this tutorial
- Same model handles verbal confirmations (yes, ok) and nonverbal confirmations (gaze, nod)
- Model handles system utterances and behavior

Evaluation: WoZ study of grounding with human-like multimodal cues vs no multimodal cues

- Human-like condition elicits human-like behavior from users, baseline does not

Result has now been broadly replicated with end-to-end systems

- Skantze, Hjalmarsson, Oertel, Speech Communication 2014,
Mehlmann, Janowski, Häring, Baur, Gebhard and André ICMI 2014.

Moment-by-moment decisions about uncertainty in conversation

Significance of behaviors in time

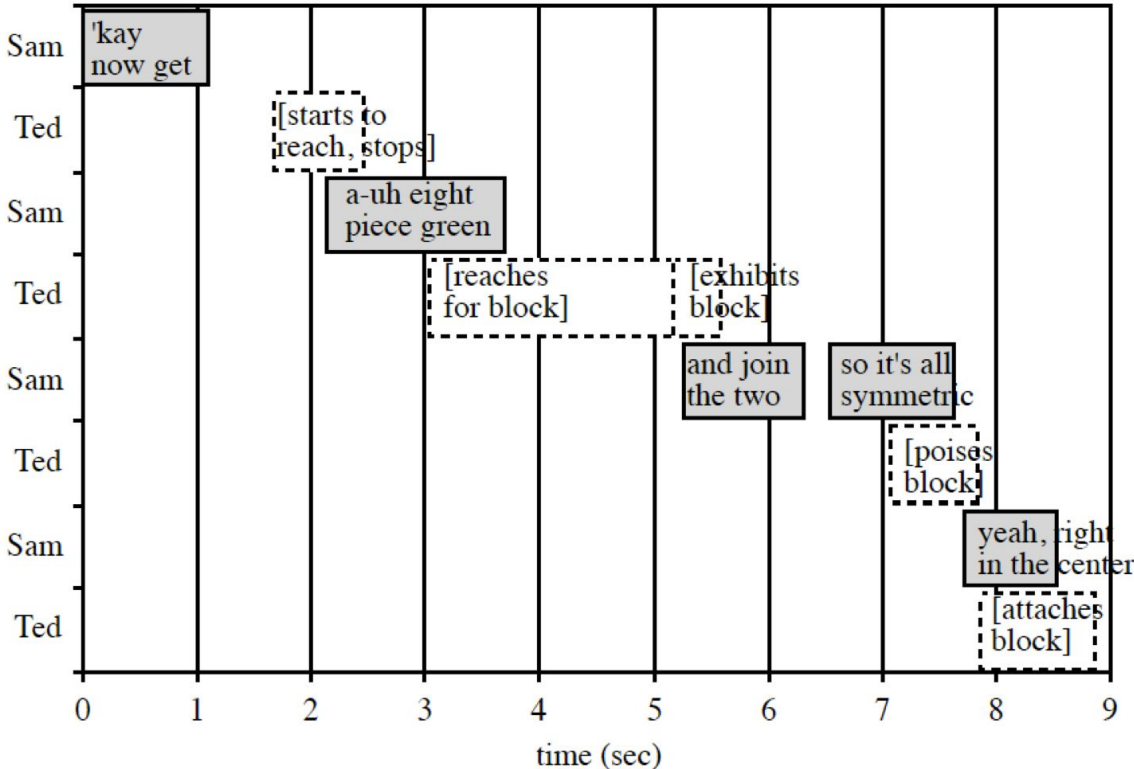
Human interlocutors monitor and display understanding in real time.

- Speakers produce utterances in short installments
- Audiences give feedback about understanding or non-understanding incrementally
- Speakers adapt their communication based on audience feedback

Demonstrated elegantly by Clark & Krych (2005)

- Asked pairs to instruct and follow lego assembly
- Tracked coordinated behaviors from logs moment by moment

Incremental grounding in human conversation



Director Sam instructs follower Ted in lego assembly from Clark & Krych (2005). Sam breaks up instructions, paces them based on Ted's response, and confirm Ted's display of understanding.

Requires incremental architecture

Recognize and interpret fragments

Understand whether interpretations are provisional

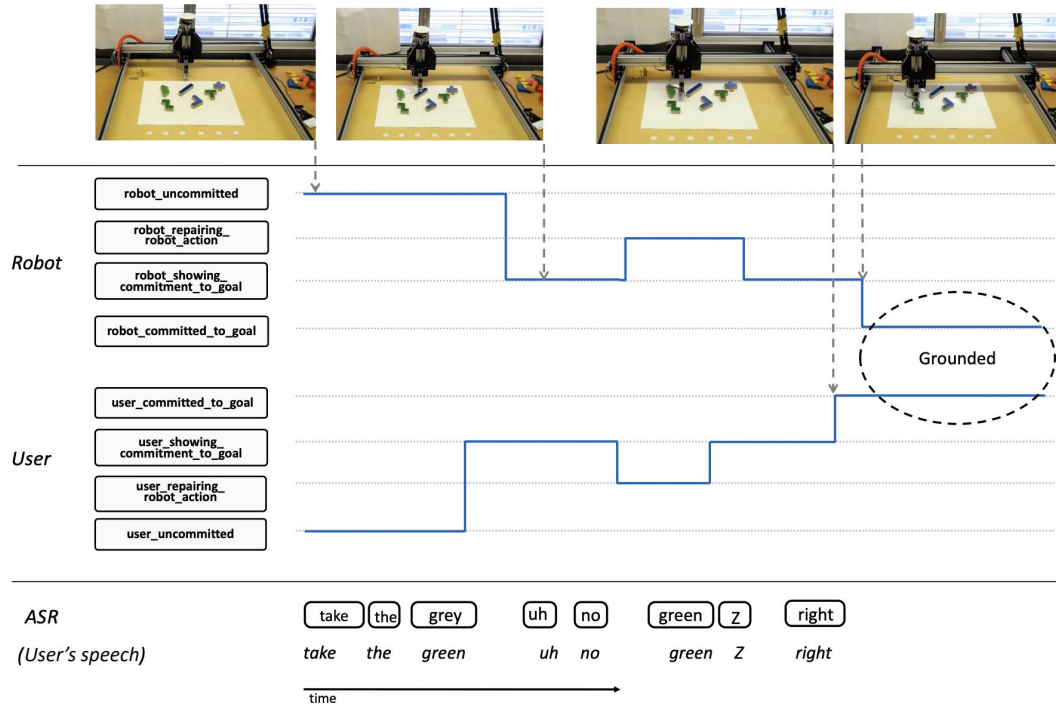
Track context incrementally

Act on likely interpretations

Advantageous in spoken dialogue systems but crucial in multi-modal interaction

Grounding Uncertainty for Simple Robots

Uncertainty can be communicated to users by principles of grounding in dialogue interaction even without natural language generation.



Implementation and evaluation

Grounding is tracked via **incremental** ISU dialogue architecture

- Reviewed in detail in Session 3 of this tutorial

Key evaluation: robot that moves faster when it's confident of its interpretation

- Users in interaction accurately recognize confidence robot designed to display

Interim summary

Multi-modal systems with diverse grounding behaviors

- Speech, gaze, gesture, physical action

Consistent emphasis on characterizing, modeling and replicating human grounding strategies

Diverse architectures, emphasizing

- Reactive control, collaborative agency and on-line inference
- Sometimes closely modeled on spoken dialogue architecture, sometimes not

Need to take more careful stock of standard grounding models

- To understand when and how they can (and can't) extend to multi-modal dialogue

How people achieve common ground

A dynamic, interactive process

- Speakers and audience collaborate
- Seek and provide evidence of mutual understanding

How people achieve common ground

Clark and Schaefer (1989)

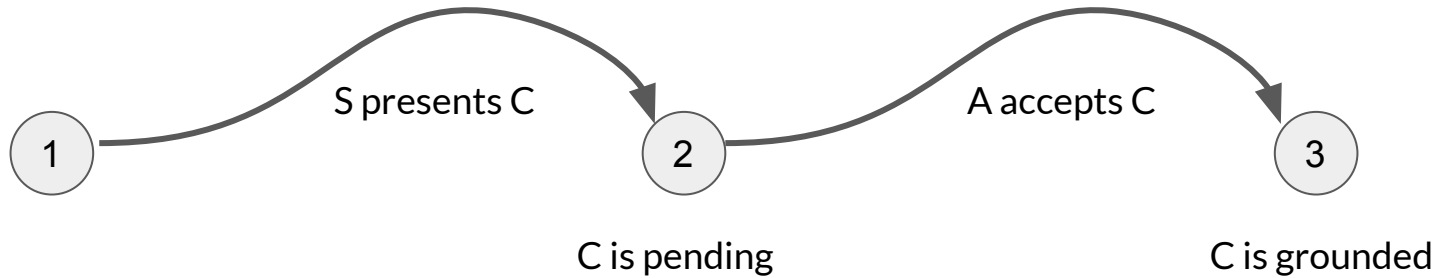
- Speaker **presents** a contribution, making speaker's attitude common ground.
- Audience **accepts** a contribution, making audience's attitude common ground.
- As a result, contribution itself is common ground.

Expert: And attach the pink thing so it covers the hole in the middle.

Apprentice: (pause) Got it. One way-valve. We're all set.

(example from Cohen & Levesque IJCAI 1992)

Information state update (ISU) model



Acceptance can be recognized from many different behaviors (or even none at all)

Traum, PhD Thesis (1994), Traum and Larsson J Language Engineering (2003)

Acceptance requires positive evidence

Clark and Schaefer (1989)

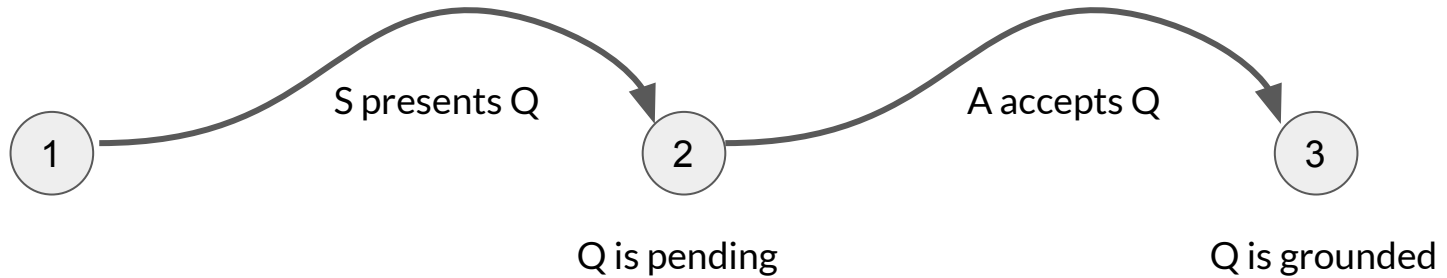
- Confirmation
- Paraphrase and repetition
- Continued attention (compare Nakano et al 2003)
- **Initiation of relevant next contribution**

S1: how far is it from Huddersston to Coventry

S2: um. about um a hundred miles

(example from Clark and Schaefer)

Information state update (ISU) model



Answer Q only in state 3

- Tacitly infer the transition from 2 to 3 via intention recognition or coherence (Thomason, Stone & Devault 2006, Lascarides and Asher, J Semantics 2009)

Audience can also give negative evidence

Clark and Schaefer (1989)

- Requests for repetition or clarification

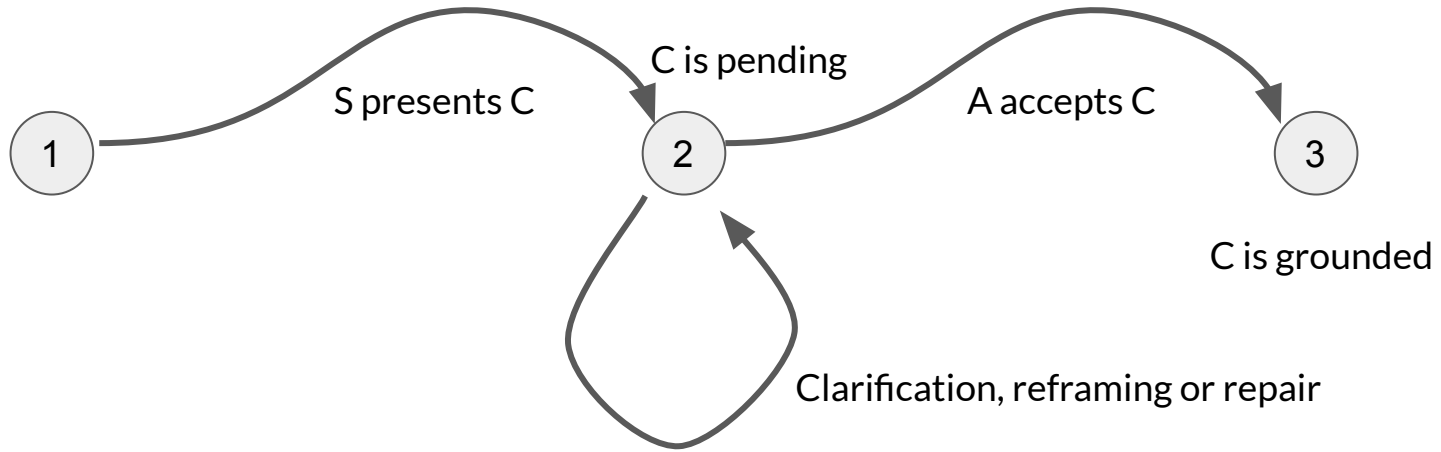
S1: we wo uh what shall we do about uh *this* boy then

S2: Duveen?

S1: m

S2: well I propose to *write*, uh saying

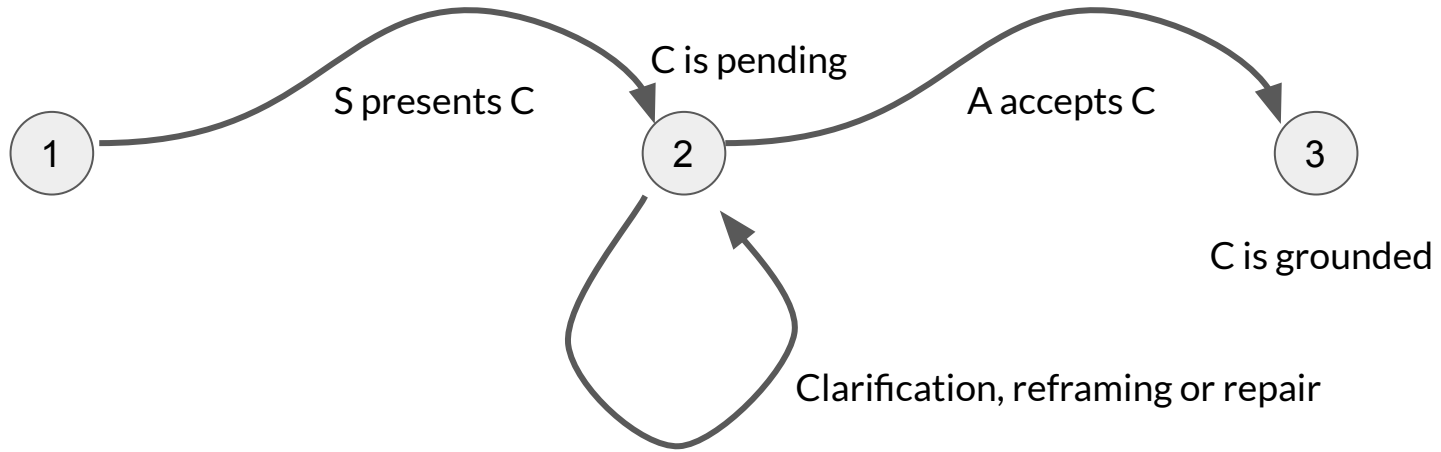
Information state update (ISU) model



State 2 allows interlocutors to initiate **side sequences**

- For example with clarification questions, other moves that give **negative evidence**
- Ensuing exchanges are subordinate
Attached by subordinating relations (Clark and Schaefer, 1989; Lascarides and Asher, 2009)
Nested subdialogues in QUD model (Ginzburg 2012)

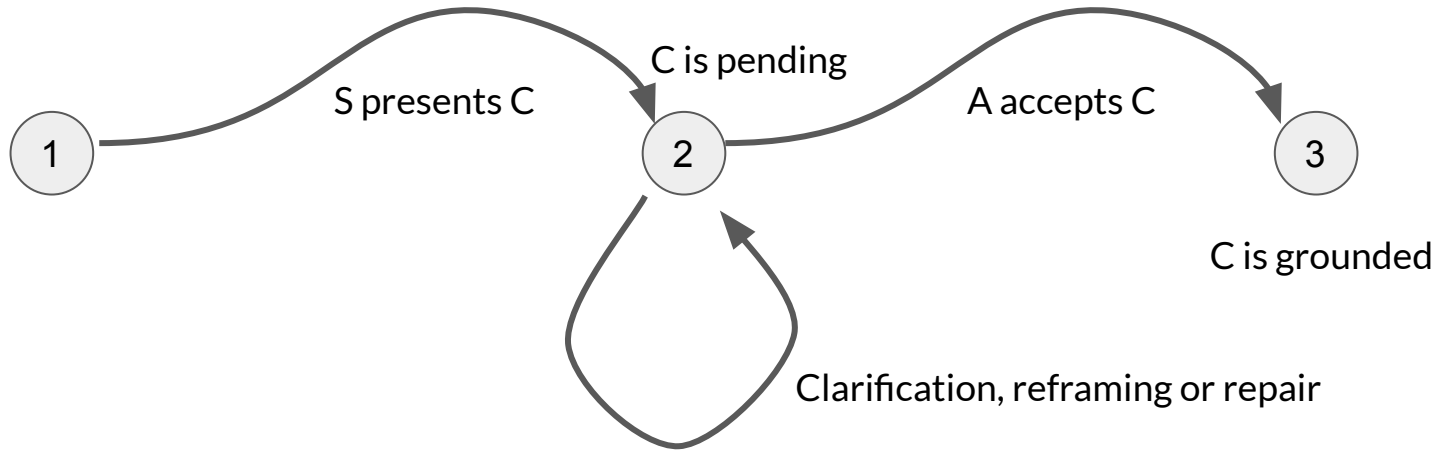
Information state update (ISU) model



Complete model also

- Lets presentation and acceptance differ, enabling 3rd turn repair
- Lets contributions be rejected in state 2 and remain ungrounded
- Lets contributions be abandoned and remain ungrounded

Information state update (ISU) model



Grounding is just one of many layers in an information-state model (Rickel and Traum 2002)

- For example, handle turn-taking in parallel
With functions that take the turn, yield the turn, etc.
Realized by and inferred from appropriate behaviors in context

How ISU systems achieve grounding

Enabling users to use natural grounding strategies

- Users judge whether they understand and accept
- Users signal their state and system tracks it
- Users are usually right
- System can respond to problems

Systems also do their part

- System must choose grounding action when user contribution C is pending
- System chooses clarification when understanding is problematic
- System provides positive evidence of understanding otherwise

ISU approaches: summary

Pluses:

- Approach handles common cases naturally and robustly
- Treats different roles in dialogue symmetrically
- Easy to extend to multimodal dialogue

Minuses:

- Not good at handling ambiguity and partial information
Especially when resolved across multiple turns
- Not good at making quantitative tradeoffs
For example when understanding is likely but not known

Modeling grounding probabilistically

For engineering simple spoken dialogue systems, POMDPs are the go-to technique

- See survey in Young, Gašić, Thomson and Williams Proc IEEE 2013

Decision-theoretic model

- Actions update state stochastically
 $P(s_t | s_{t-1}, a_{t-1})$
- States are hidden, so agent must reason with noisy observations
 $P(o_t | s_t, a_{t-1})$
- Rewards quantify the agent's task outcomes
 $r(s_t, a_t)$

Learn model parameters and find strategy to optimize expected reward

The logic of POMDPs

POMDPs select actions as a function of **belief** – distribution $b_t(s_t)$ over states

Decision-theory lets choice reflect uncertainty

- value of information (sensing to improve later decisions)
- long-term consequences (establishing preconditions to improve later actions)

Bayesian filtering for belief update – maintains evidence optimally

- $b_{t+1}(s_{t+1}) \propto P(o_{t+1} | s_{t+1}, a_t) \sum P(s_{t+1} | s_t, a_t) b_t(s_t)$

POMDPs and probabilistic grounding

Capture many grounding phenomena

- Knows how to pursue an extended line of questioning
(e.g., incremental slot-by-slot confirmation vs single final summary confirmation)
As a function of system's current uncertainty and likelihood strategy will resolve it
- Knows when to stop
value of information finds point of diminishing returns
- Integrates uncertain information across utterances
Including multiple noisy ASR results

POMDPs – key limitations

Practical limits on data and computation have so far been simple

- Simple dialogue representations
Discrete flat, attributes of user intent; limited dialogue history; simple actions
- One-sided reasoning
No tracking user uncertainty or anticipating user coordination
- Restricted models of interaction
Discrete time whole-turn choices

Modeling grounding probabilistically

Alternative approach – focus on tracking expressive dialogue state representation

- Illustrated by Paek and Horvitz UAI (2000)
Seminal work linking cognitive science of grounding and AI methods

They describe dialogue in terms of a Bayesian belief network

- Formalize hidden variables leading to observed user behavior
- Use probabilistic inference to reason about levels of system understanding
- Make predictions about immediate effects of system strategies
- Choose strategy with best immediate effect

Modeling grounding probabilistically

DeVault PhD 2008, “Contribution Tracking”

- Hierarchical task-based model of dialogue state, as in ISU models
- Simple particle filter for state tracking with data-driven probabilities
- Plan utterances using “conformant planning”,
symbolic AI techniques for complex planning under uncertainty
- Symmetric grounding facilities regardless of system role in interaction

(See DeVault & Stone EACL 2009, Stone & Lascarides Semdial 2010, McMahan & Stone Sigdial 2013)

Looking ahead

Increasing abilities to learn probabilistic cognitive models of utterance choice in dialogue

- For example:
Andreas and Klein, EMNLP 2016.
Monroe, Hawkins, Goodman and Potts, TACL 2017.
McDowell and Goodman, ACL 2019.
McMahan and Stone, SIGIDIAL 2020.

These models can be reused in Bayesian belief network models of dialogue

Looking ahead

Increasing range of probabilistic models of nonverbal behavior

- For example:

Wang, et al, Probabilistic models of physical actions in HRI, RSS 2012.

Sheikhi and Odobez, Gaze and attention in human-robot conversation, PRL 2015.

Lee and Marsella, Head nods and eyebrow raises in multiparty conversation, IVA 2012.

These models can also be reused in Bayesian belief network models of dialogue

Modeling grounding incrementally

Two keys to incrementality:

- Getting information in real time
- Deciding real-time responses

Hough and Schlangen HRI 2017

- Report and interpret **partial** recognition results
- Follow policy set out in information-state grounding model

Paetzel, Manuvinakurike and David DeVault SIGDIAL 2015

- Predict and anticipate **complete** recognition results, before user finishes
- Follow policy based on data-driven optimization

Incremental ISU

(Hough and Schlangen HRI 2017)

Parallel states for user and robot

- Pending contribution
- Grounded contribution
- No contribution: incremental information not yet available
- Repair in progress: pending contributions from interlocutors did not match

Incremental classification of user utterances as parts of contributions or repairs

Optimizing for incremental response

Paetzel, Manuvinakurike and David DeVault SIGDIAL 2015

Set up a space of possible policies for responses

- Threshold for confidence in correct interpretation to act early
- Threshold for timeout to act late

Use grid search and simulation of real algorithms on in-domain data to tune parameters

Interim summary

Grounding models differ in scope:

- Track uncertainty to reason about meanings of system and user utterances
- Track commitments to reason about alignment of system and user belief
- Track in real time to select real-time responses

They differ in how they lead to system behavior

- Provide guardrails for well-designed interactions
- Make explicit but heuristic choices
- Make optimal, long-term plans

Some models extend naturally to multi-modal data, others do not

Handling diverse actions in dialogue

For modeling grounding

- Some approaches extend easily to multi-modal communication
- Others do not

Similarly, for understanding and generation

- Some NLP concepts and approaches fit multi-modal communication
- Others do not

Handling diverse actions in dialogue

A brief tour of understanding across modalities

- What do different modalities have in common?
- What makes each modality unique?
- What do you need to infer about a communicative action
To predict its implications for grounding?

Looking at a few key case studies

- An invitation to explore lots of great descriptive work
- Ekman 1969, Argyle 1976, Goodwin 1981, McNeill 1992, Kendon 2004, etc.

Some important cases

Physical actions

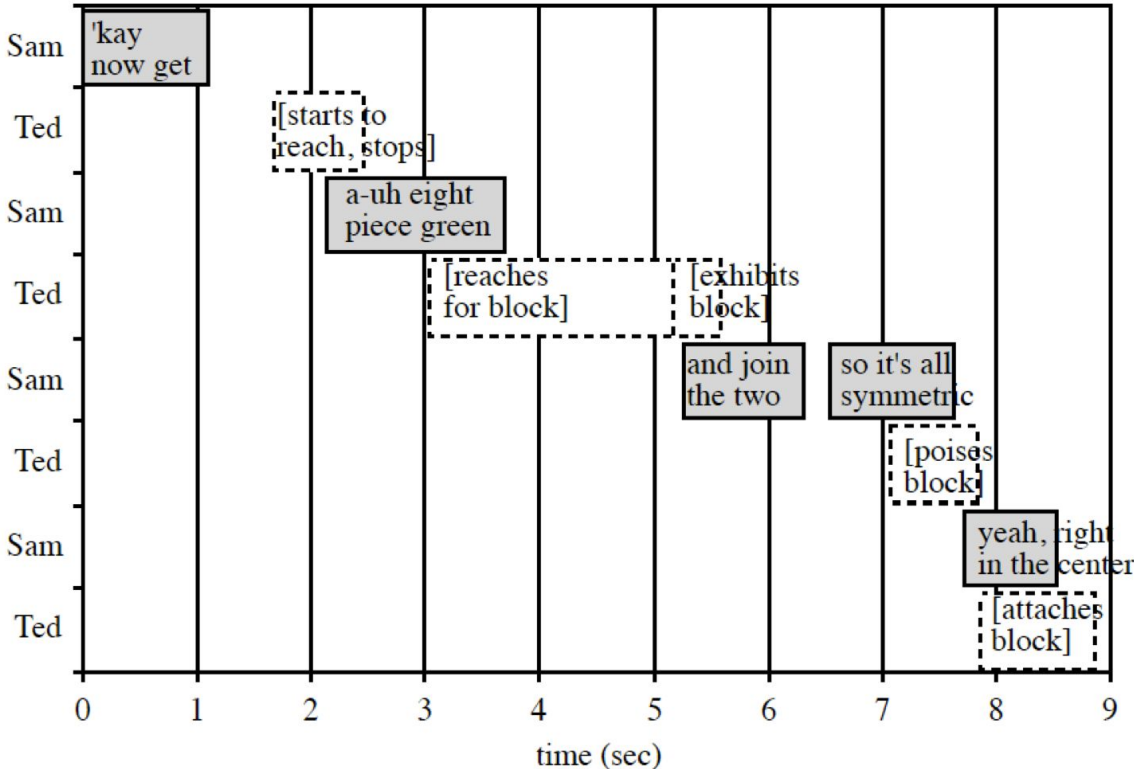
Coverbal gestures

Facial displays

Back-channel vocalizations

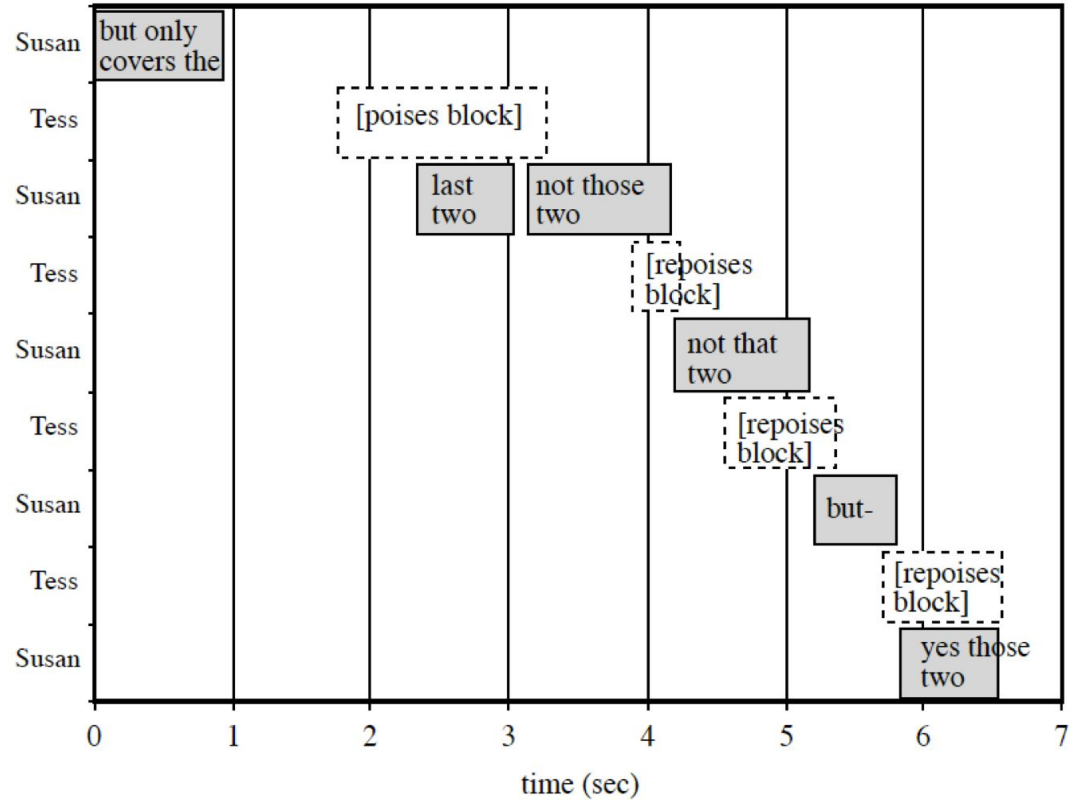
Demonstrations

Physical actions



Director Sam instructs follower Ted in lego assembly from Clark & Krych (2005). Sam breaks up instructions, paces them based on Ted's response, and confirm Ted's display of understanding.

Actions, inference and grounding



Director Susan instructs follower Tess in lego assembly from Clark & Krych (2005). Susan recognizes that Tess's poise is intended to match her instruction but does not.

Actions, inference and grounding

To track the grounding effect of others' actions, systems need to do **intention recognition**

- In the sense of Pollack (1992)
- Explain why the agent thought the action was good
- Link actions to inferred background beliefs and goals of the agent
Even if those beliefs are false or the goals aren't shared

To track the grounding effect of others' actions, systems need to assume **coherence**

- In the sense of Hobbs (1979)
- Why attribute a false belief vs an unexplained goal?
- Assume that contributions to interaction favor specific kinds of relationships

Actions, inference and grounding

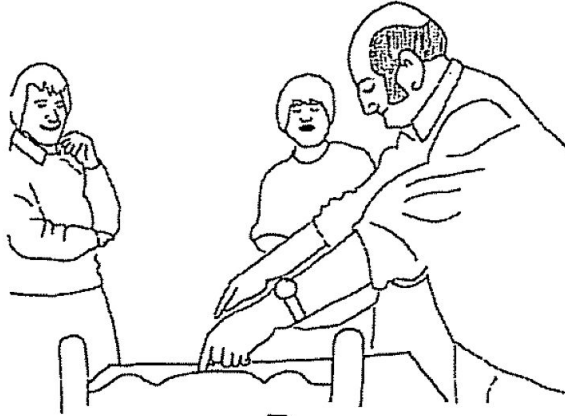
In sum:

- We can infer Tess's reposing at time 4 is evidence of non-understanding **only if**
We recognize that Tess intends the reposing to show understanding
We recognize she falsely believes the block's new poise is the instruction target

Grounding depends on action understanding, not action recognition

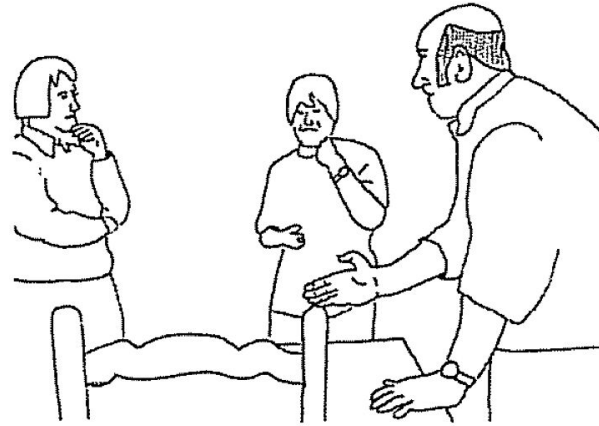
Coverbal gesture

Christmas cake example from Kendon (2004)



B

and it was (1.02) this sort of (0.4) size



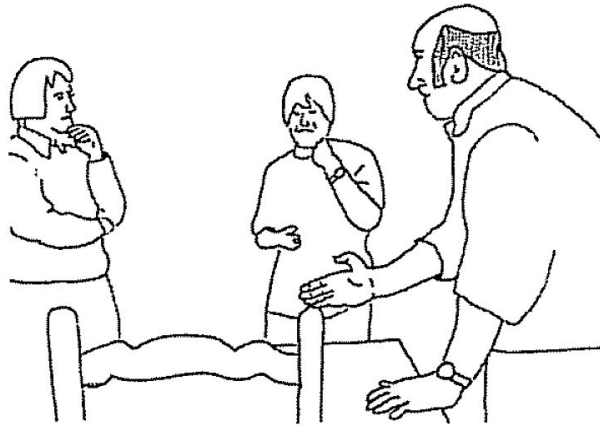
C

and he'd cut it off in bits

Inspired analyses in Lascarides & Stone, *Gesture* 2009 and *Journal of Semantics* 2009

Gesture and depiction

Christmas cake example from Kendon 2004



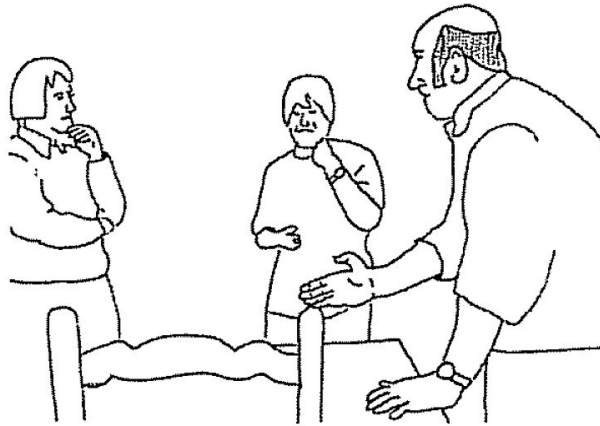
C

and he'd cut it off in bits

Speaker's hand **depicts** knife by shape, and **locates** knife in cake by relation to previous gesture. Speaker's **arm and body** depict action of agent.

Gesture and depiction

Christmas cake example from Kendon 2004



C
and he'd cut it off in bits

Speaker's hand **depicts** knife by shape, and **locates** knife in cake by relation to previous gesture. Speaker's **arm and body** depict action of agent.

Gesture **combines** multiple independent elements, with **iconic** and **deictic** interpretations.

Gesture needs to be **understood** not just recognized (like language and action).

Facial displays

Q: What year was Rutgers founded?



(silence)



seventeen



forty-two



maybe?

Example from Stone & Oh, 2008

Conversational facial displays

Facial displays can express emotion

But in conversation they more often **depict** emotion to comment on synchronous speech

- To give information about what speaker is doing and how they appraise it
- Combining elements of multiple emotional displays in creative ways

Emphasized in research by Bavelas and Chovil (Gesture 2018), among others.

Back-channel vocalizations

Sometimes more like gesture and facial expressions – not lexical

- “h-nmm, hh-aaaah, hn-hn, unokay, nyeah, ummum, uuh, um-hm-uhhm, um and uh-huh”
- Productively generated rather than finite in number
- Sound-meaning mapping is rule-governed rather than arbitrary

Ward, Pragmatics and Cognition 2006 – see also Prosodic Patterns 2019.

Demonstration

Vi Hart <http://www.youtube.com/watch?v=z6lL83wl31E>

- You don't need numbers or fancy equations to prove the Pythagorean theorem. All you need is a piece of paper. There's a ton of ways to prove it, and people are inventing new ones all the time, but I'm going to show you my favorite—except instead of looking at diagrams we're gonna fold it.

Video analyzed in detail by Stone and Stojnić RPP (2015) with similar conclusions

- Track real world effects of actions
- But connect them to speaker intentions by belief-desire reasoning
- And principles of discourse coherence

Interim summary

Dealing with multi-modal dialogue means accepting the differences across modalities

- Not just communicative content, but also practical action and natural reactions
- Contributions have different basis from language and require different reasoning
- Coherence and intention recognition are the right levels to bridge the differences

Received wisdom – folk theories, scientific descriptions, engineering benchmarks – often too simple

- Frameworks like emotional expression or gesture recognition just scratch the surface
- Idea of multi-modal fusion or fission is inherently limited

Actions create common ground because they reveal and coordinate agents' mental states

- Enduring problem for dialogue systems

Grounding in Dialogue with Animated Agents and Robots

This session

We will discuss

- Why HRI needs to bridge representations?
- How to do it?
 - The role of multimodality in grounding.
 - Learning actions and word meanings in human-machine dialogue.
 - Ways that we can signal and represent uncertainty in dialogue.
- Bottlenecks of generating co-ordinated multimodal presentations.

Common Ground in Human-Robot Dialogue

Agents can engage in joint tasks if

- humans and agents both make extra efforts to bridge the gap and strive for a common ground of the shared world.
- computational models for language grounding take collaboration into consideration.
- they incorporate collaborative effort from human partners to better connect human language to its own representation and make extra collaborative effort to communicate its representation.

Multimodal Collaboration in Referential Communication

How conversation partners collaborate and mediate shared basis when they have mismatched visual perceptual capabilities?

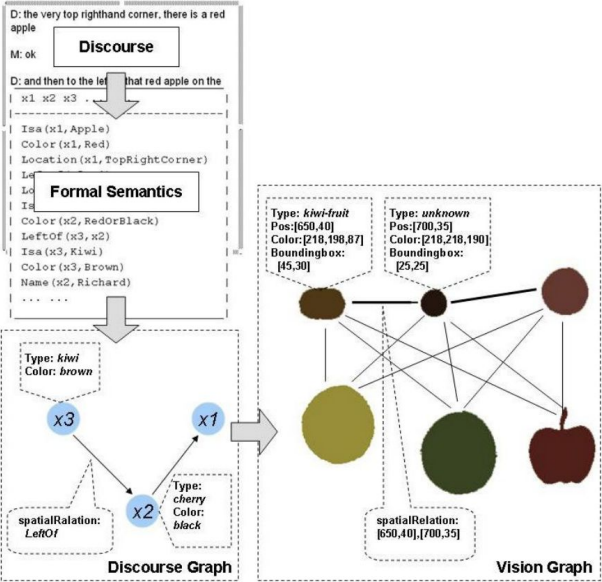
Director



Matcher

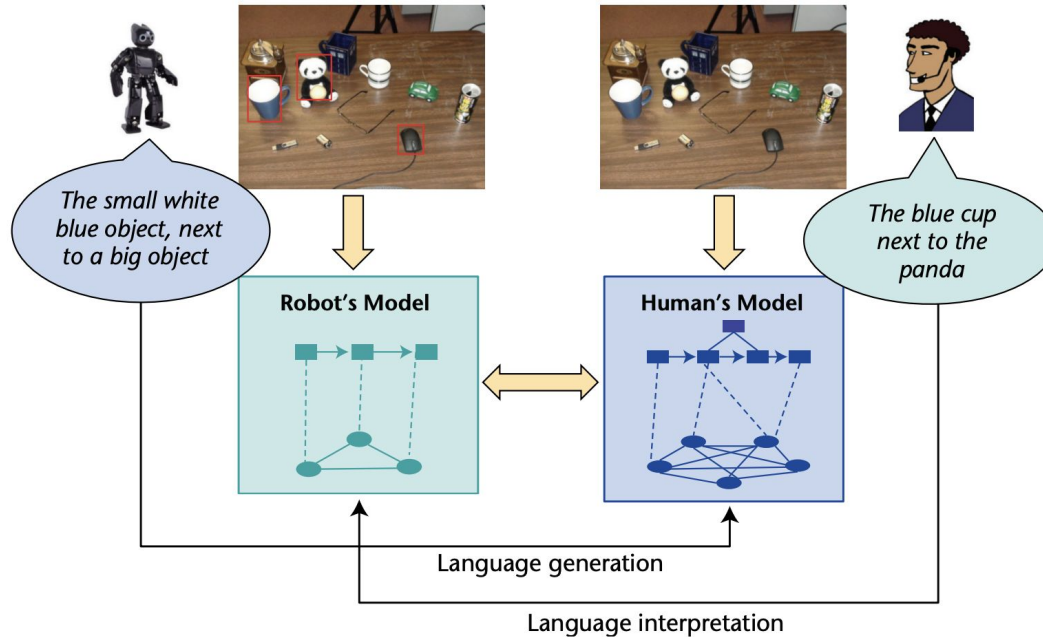


- D1: there is basically a cluster of four objects in the upper left, do you see that?
- M1: yes
- D2: ok, so the one in the corner is a blue cup
- M2: not a cup, I see there is a square, it is blue
- D3: alright, I will go with that, right under that is a yellow pepper
- M3: ok, I see apple but orangish yellow
- D4: ok, so that yellow pepper is named Brittany
- M4: uh, the bottom left of those four? Because I do see a yellow pepper in the upper right
- D5: the upper right of the four of them?



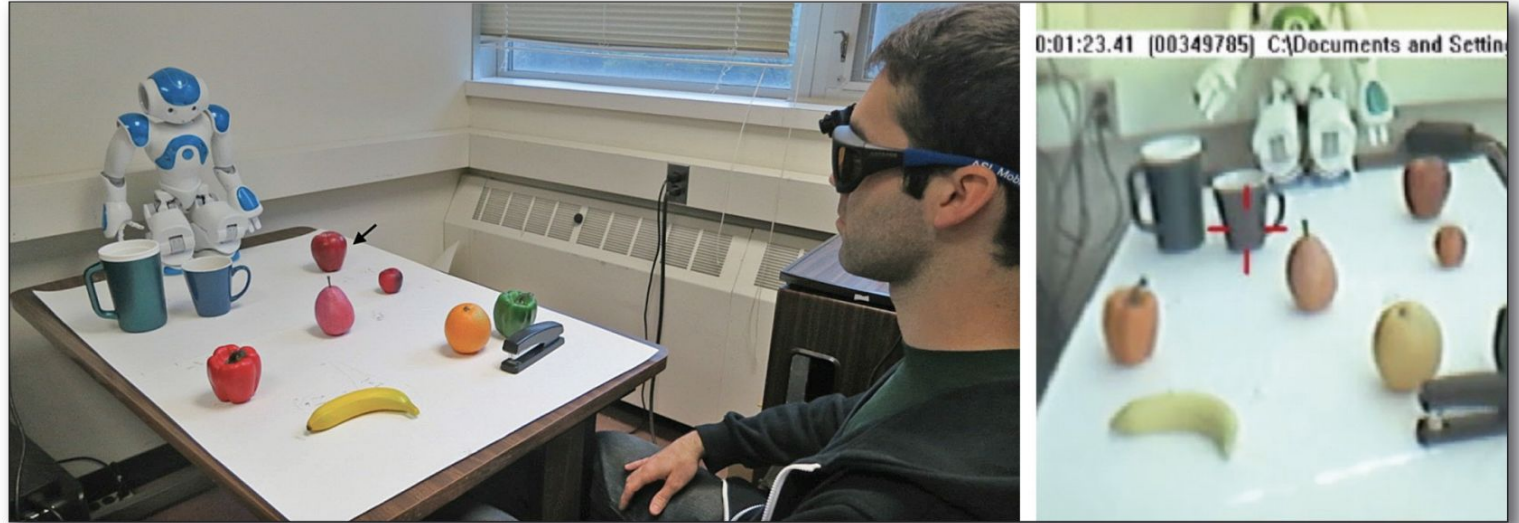
Situated Human-Robot Communication

Better shared representations are possible by employing optimization approaches based on linear programming to automatically learn the weights to match the referent during dialogue.



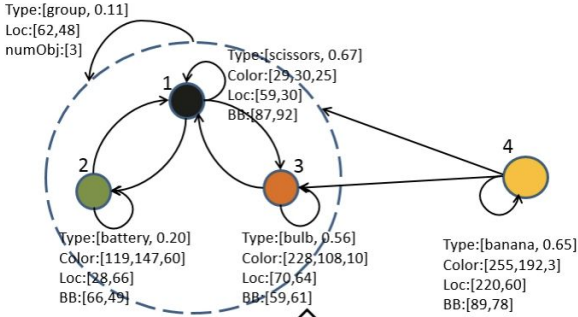
Multimodal Collaboration in Referential Communication

Instead of a single minimum description to describe a target object, episodes of expressions are generated based on human feedback during human-robot interaction.



Multimodal Collaboration in Referential Communication

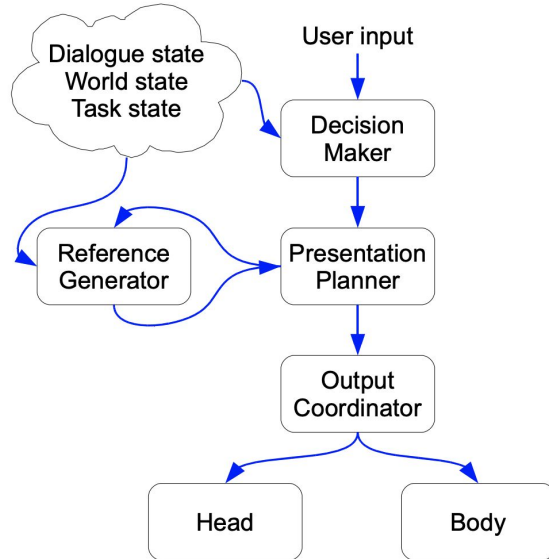
An example of the robot's internal representation of the shared environment. Incorporating the robot pointing gesture consistently outperformed the models that don't incorporate the pointing gesture.



R1:	“Do you see a group of two on the left?” <i>(Robot points to the group)</i>
H1:	“Yes” <i>(The group becomes the landmark)</i>
R2:	“That group I was just talking about, do you see an object in the back that is on the right within that group?” <i>(Describe the intended object in relation to the landmark)</i>
H2:	“No” <i>(Human rejects the description)</i>
R3:	“That group I was just talking about, do you see an object on the right that is to the right of that group?” <i>(Robot chooses another description strategy)</i>
H3:	“Yes”

Multimodal Collaboration in Referential Communication

When referring to objects in the context of a task-based interaction involving jointly manipulating objects, a much richer notion of context is available, which permits a wider range of referring options e.g. manipulating objects as part of the task.



JAST: [picks up and holds out green cube] First, fasten this green cube to the end of a short slat using the red bolt.

USER: Should I use this slat? [pointing to a five-hole slat]

JAST: No, use a three-hole slat. [picking up and holding out a three-hole slat]

USER: [takes slat; picks up bolt]

Collaborative Effort Toward Common Ground

- Collaborative effort
 - Low: acceptance
 - High: acceptance + description
- Task incompleteness: If the robot fails to acquire new knowledge it will stop.



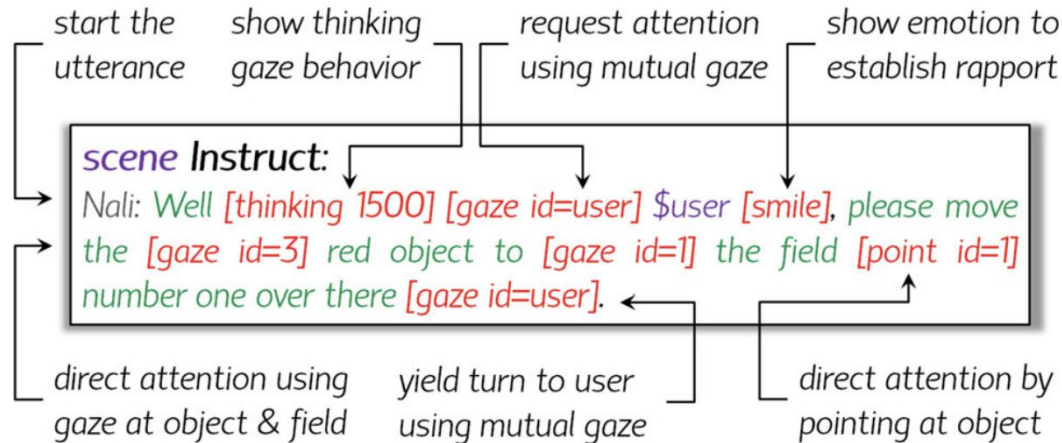
A: The green cup is called Bill.

B: Ok, the green cup is Bill.
[point to the inferred object]

Shared basis	Low mismatch		High mismatch	
	low	high	low	high
Collaborative effort	low	high	low	high
Perceived Common Ground	0.71	0.25	0.58	0.25

Modeling Grounding for Interactive Social Companions

- Natural grounding behavior requires the precise synchronization of numerous parallel and bidirectional behavioral aspects.
- Maintaining the common ground requires domain knowledge but has also numerous social aspects, such as attention, engagement and empathy.



Collaborative Step-by-Step Instructions

- For robots to follow human language instructions and perform actions in the physical world, grounding language to perception alone is not sufficient.
- How can we connect language commands with the corresponding sequence of primitive robotic operations?
 - The human operator can teach the robot high-level actions (for example, stack) in a step-by-step manner.
 - Given this teaching and learning instance, how should the robot internally represent knowledge or grounded semantics for the verb frame stack?
 - A more desirable representation for grounded semantics of the verb frame stack(A, B) should capture the desired **goal state**.

Collaborative Step-by-Step Instructions

Human: Stack the green block on the left to the green block on the right

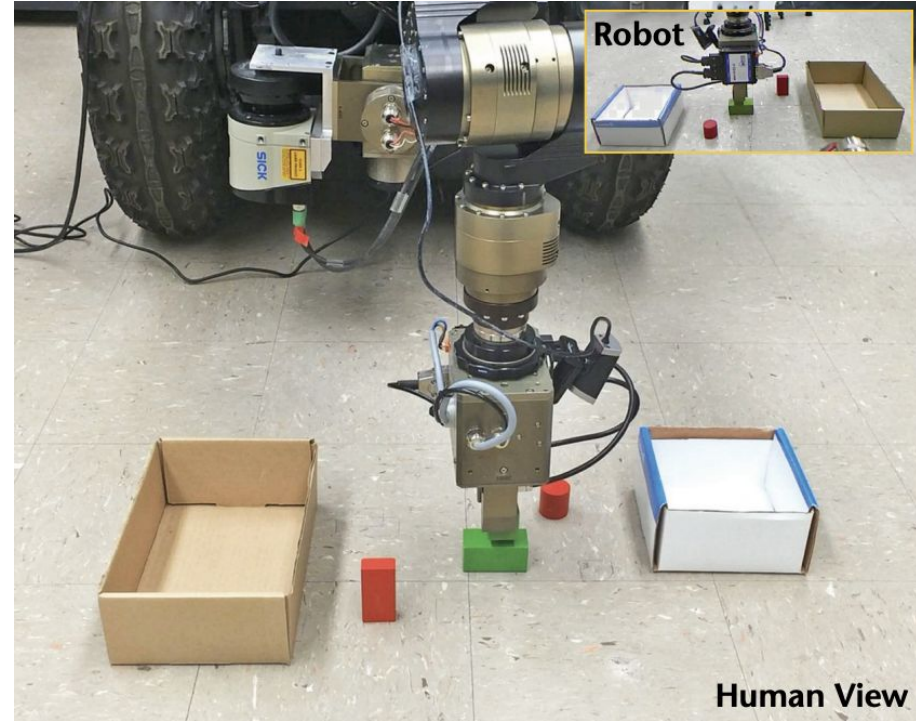
Robot: I don't know how to perform this stack, please give step-by- step instructions.

Human: Move the green block on the left to the top of the green block on the right.

Robot: [performing the move action]

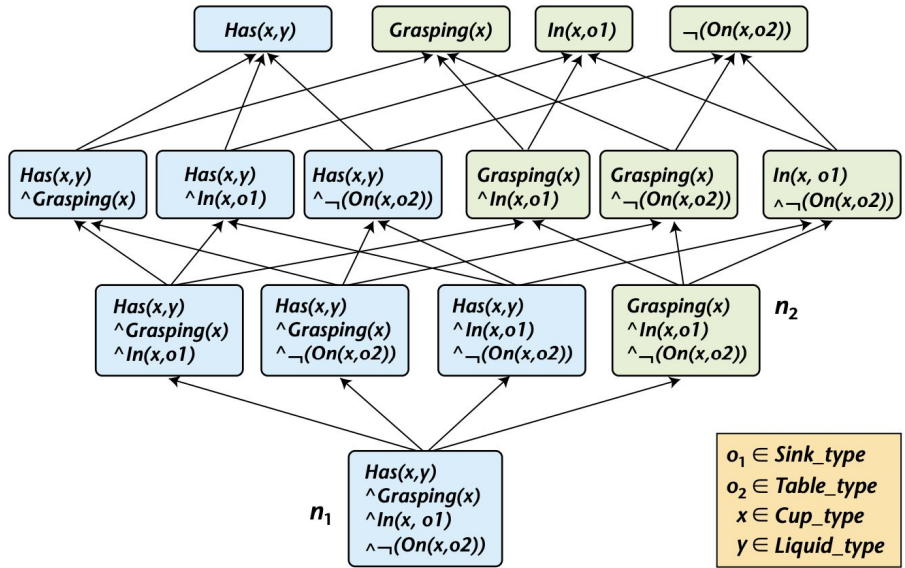
Human: Open gripper.

Robot: OK.



Learning Grounded Verb Hypothesis Space

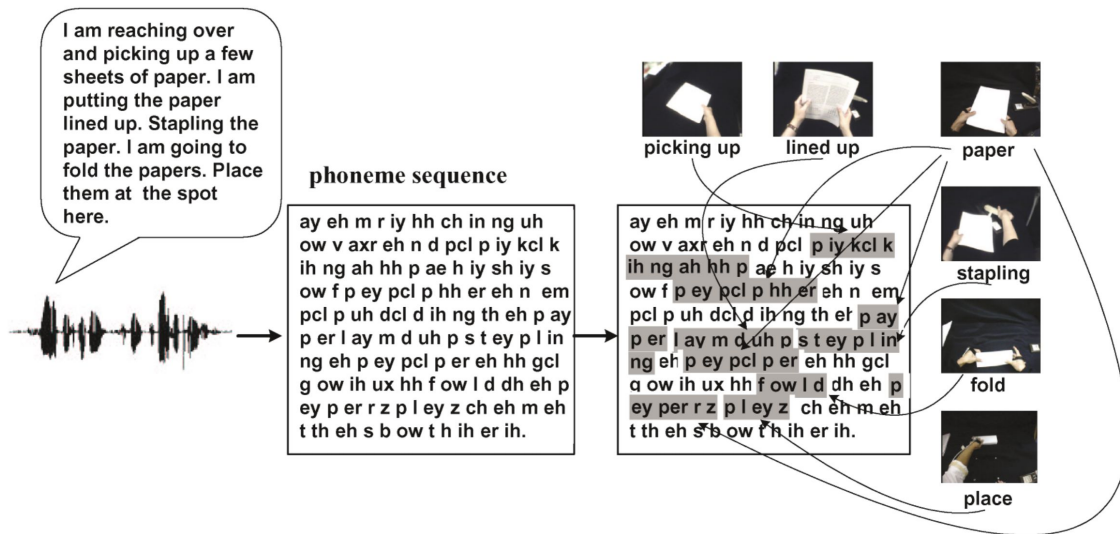
The goal state, which is represented by a conjunction of logical predicates can be acquired by the robot after performing the low level operations.



An example of grounded hypothesis space for verb frame $fill(x, y)$.

A Multimodal Learning Interface for Word Acquisition

Solely statistical learning of co-occurring data is less likely to explain the whole story of language acquisition. The inference of speaker's referential intentions from their body movements provides constraints to avoid the large amount of irrelevant computations.

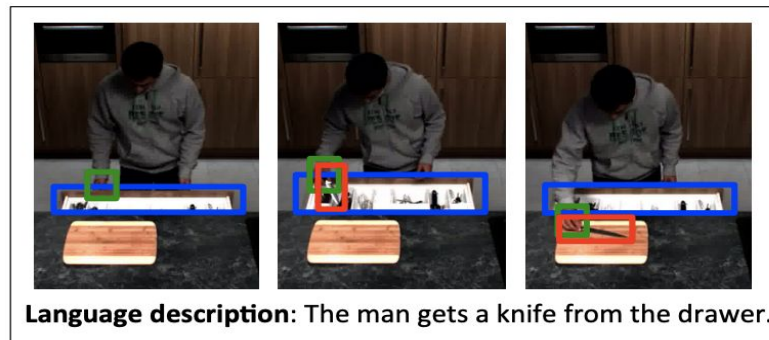


Gaze policies in HRI

- Robot makes saccades contingent on human looking
- Policies track target object, look at partner
- Looks at partner elicit mutual gaze
- Mutual gaze improves multimodal coordination and synchrony

Modeling Physical Causality of Verbs for Grounded Language Understanding

- Crowdsourcing and automatic segmentation methods can successfully scale up the previous attempts for studying grounded verb meanings and creating multimodal corpora.
- This work applies causality modeling to the task of grounding semantic roles to the environment. using two approaches:
 - a knowledge-based approach and
 - a learning-based approach.



Verb: "get"

Agent: ground to the hand in the green box

Patient: "knife", ground to the object in the red box

Source: "drawer", ground to the object in the blue box

The man gets a knife from drawer.

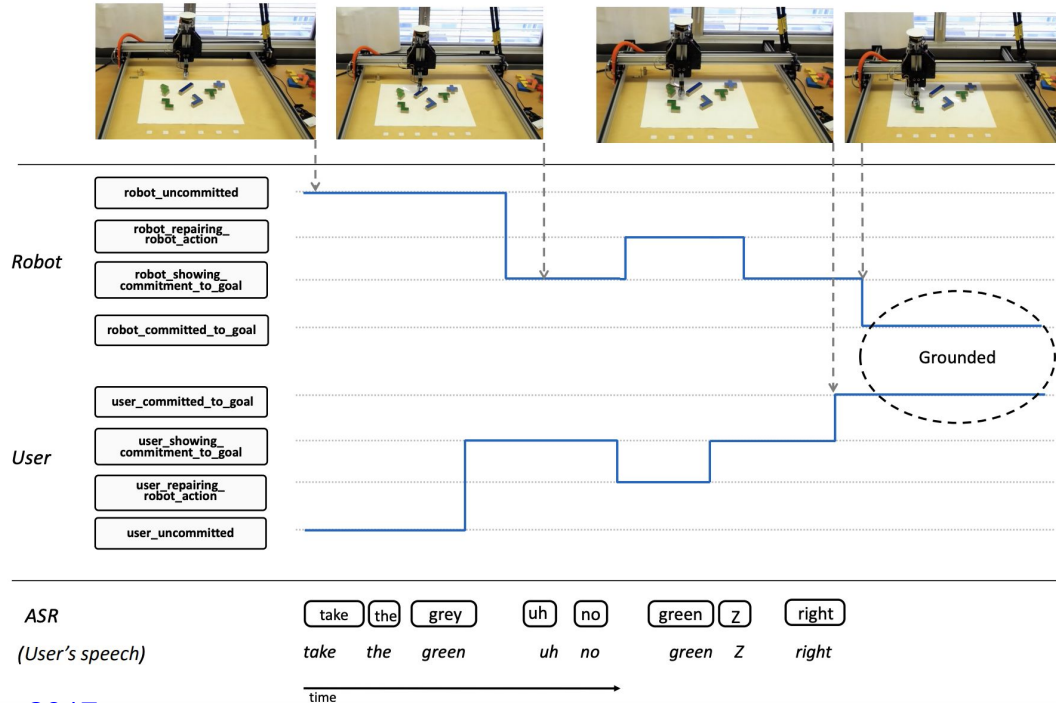
Grounding Uncertainty for Simple Robots

For effective HRI:

- Robots must go beyond having good legibility of their intentions shown by their actions.
- Robots should ground the degree of uncertainty they have.
- A robot not only needs to monitor when its internal goal is becoming legible, but the robot should also be able to ground the degree of commitment to its goal.
- It's important to achieve the tradeoff between 'safety' and speed of movement similar to system.

Grounding Uncertainty for Simple Robots

Uncertainty can be communicated to users by principles of grounding in dialogue interaction even without natural language generation.



Grounding Uncertainty for Simple Robots

Inferring uncertainty :

- **Uncertainty through repair only:** only grounds uncertainty by allowing repairs to change its goal and therefore change its action.
- **Uncertainty through movement:** also allows repairs to change its goal but also exhibits its own level of confidence about its goal through its speed of movement and waiting time before acting.

Evaluation:

- Understanding: to what degree did you feel the robot understood what it had to do? (1-7)
- Confidence: to what degree did you feel the robot had confidence in its decisions to act?(1-7)

Communicating Uncertainties in Situated Interactions

- This approach harnesses a representation that captures both the magnitude and the sources of uncertainty, and a set of policies that select and coordinate the production of nonverbal and verbal behaviors.
- The methods are designed to enlist participants' help in a natural manner to resolve uncertainties arising during interactions.



Figure 1. Nonverbal expressions of uncertainty.

Identifying Opportunities for Empathetic Responses

- Multimodal empathy analysis for informed turn-taking strategies.
- Automatic recognition of opportunities for providing empathetic responses.
- Future directions: learning to choose an adaptive threshold for providing personalized empathetic responses.



Figure 1: A participant and the virtual agent, Ellie.

A: How have you been feeling lately?

H: Um kind of uh I guess sorta sorta depressed generally

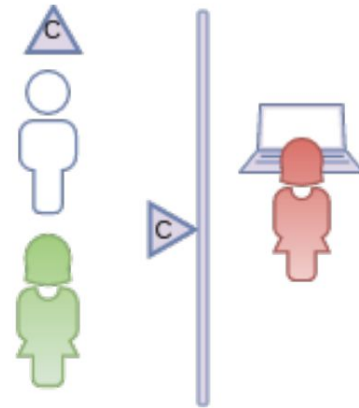
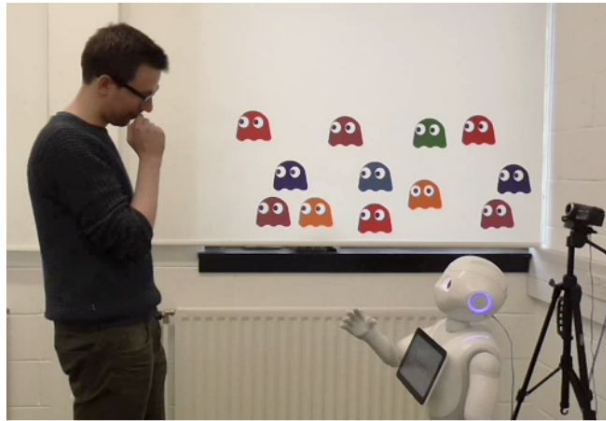
A: Tell me more about that

H: Uh just uh feeling tired and sluggish and um less less motivated and less interested in things

A: I'm sorry to hear that

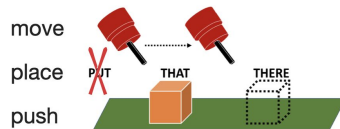
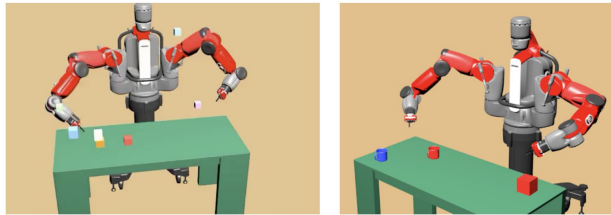
Use of Multimodal Features

- Dialogue features correlate with the user's perception of a robot, as well as correlations between emotional features and robot likeability.
- These characteristics may in future be used as an online reward signal for in-situ Reinforcement Learning based adaptive human-robot dialogue systems.



Bottleneck: Multimodal Generation

- Generating multimodal communicative actions even when working with humanoid robots is challenging: Accuracy, Commonsense knowledge, Biases of human collaborators
- How the content of presentations in different modes related to each other?
 - System building approach: keeping track of the coherence and synchrony is a challenge.
 - Data-driven approaches fail to take into account the information goal.



Alikhani et al, 2020



Bollini et al, 2012



Zhao et al, 2019

Adapting Grounding to Modality and Culture

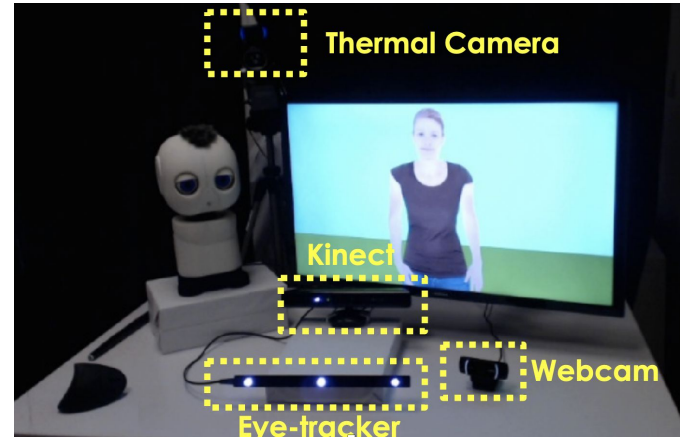
This session

- Choice of modality and grounding
- Non-verbal grounding in languages beyond English
 - German
 - Japanese
 - French
 - Sign languages

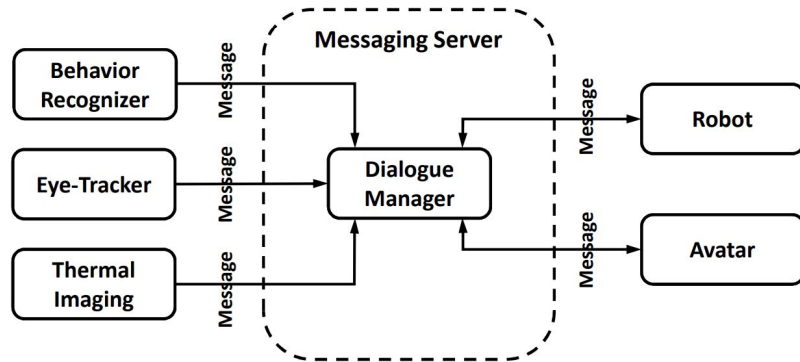
Adapting to User's Choice of Communication

Dialogue policy selects individual actions and planned multiparty sequences based on perceptual inputs about the baby's internal changing states of emotional engagement.

Tools: Eye-tracker, Webcam, Thermal Camera, ...



The robot and the avatar teaching infants sign language.



Adapting to User's Engagement

- Grounding strategy to user engagement
- For engaging multimodal interactions with students, Crystal Island combines
 - commercial game technologies
 - intelligent tutoring systems
 - rich narrative structures



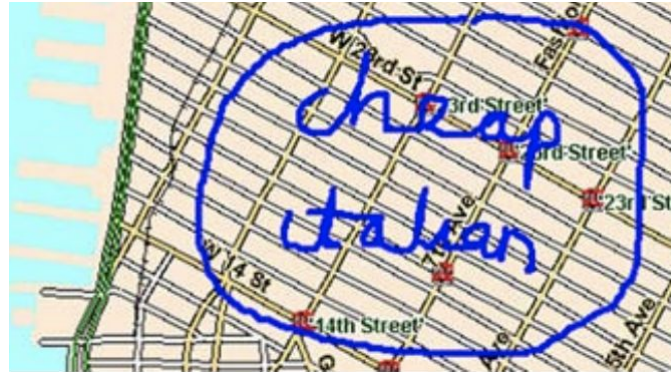
A snapshot of the Crystal Island interface.

Adapting to user's preference of communication mode

- Mobile interfaces can allow the user and system to adapt their choice of communication modes.
- MATCH combines finite-state multimodal language processing, a speech-act based multimodal dialogue manager, dynamic multimodal output generation, and user-tailored text planning and provides a mobile multimodal speech-pen interface to restaurant and subway information for New York City.



MATCH running on Fujitsu PDA



Unimodal pen command.

Information Across Modalities

What information is in text vs images?

Alikhani et al. show the potential of crowdsourcing and machine learning models for learning inferences in text and imagery.



Lower peaches into the boiling water and simmer until skin loosen, 30 to 60 seconds.

information in text	information in images
---------------------	-----------------------

do it clearly	put as much
---------------	-------------

let cool for	blend and blend
--------------	-----------------

season lightly	cut side towards
----------------	------------------

favorite toppings	cover with
-------------------	------------

Top bigram and trigram Naive Bayes SVM features.

Crowd-sourcing NLG Data: Pictures Elicit Better Data

- This work presents a framework for crowdsourcing NLG data.
- Utterances elicited by pictorial meaning representations are judged as significantly more natural, more informative, and better phrased, with a significant increase in average quality ratings.



A family-friendly, Sushi/Japanese restaurant, cheap, neither near the centre of town nor near the river.



A restaurant by the river, serving pasta/Italian food, highly rated and expensive, not child-friendly, located near Cafe Adriatic.

Examples of pictorial meaning representations.

Non-verbal grounding in languages beyond English

- Cultural differences
 - Showing understanding: different nodding gestures
 - Gestures that can be considered rude:
 - Pointing with the index finger
 - Thumbs up 👍
 - OK! 🤞
 - Thank you!
 - “stop” or “talk to the hand”



Non-verbal grounding in languages beyond English

- Corpora and systems designed for studying grounding in other languages
 - German:
 - A corpus of natural multimodal spatial scene descriptions
 - They study shape, size, distance and language-dependent properties of pointing action.

Non-verbal grounding in languages beyond English

- Corpora and systems designed for studying grounding in other languages
 - Japanese:
 - A robot that can acquire new words and their meanings while engaging in multi-domain dialogues.
 - Spanish:
 - Physically grounded language acquisition system to spanish

Non-verbal grounding in languages beyond English

- Corpora and systems designed for studying grounding in other languages
 - French:
 - A toolkit for language grouping for dialogue processing
 - A corpus of language games that both integrate the various activities required for dialogue and ground unknown words or phrases in a specific context, which helps constrain possible meanings.
 - Sign language:
 - Studies in differences of spatial language use in american sign language and English. E.g. Emmorey and Casey

Interim summary

- The choice of modality exploits grounding.
- Systems need to adapt their choice of communication mode to the user's
 - Preferences
 - Language
 - Culture

Exploiting Language Grounding for Grounding in Multimodal Contexts

Grounding in Dialogue vs Language Grounding

- [Grounding](#) in the dialogue community: the collection of "mutual knowledge, mutual beliefs, and mutual assumptions" that is essential for communication between two people.
- Language grounding: connecting linguistic symbols to perceptual experiences and actions.
Examples:



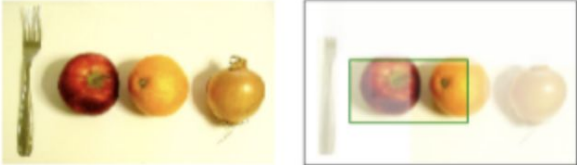
Dog reading newspaper (NP)



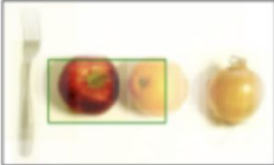
Swim (V)

Does this always work?

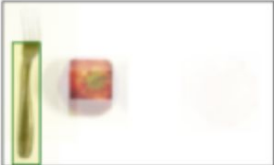
Many popular models exhibit poor visual grounding.



fork



apple



orange



What color are the bananas?

Yellow



Examples of related training data

Generating Grounded Descriptions

Natural language for visual reasoning



The left image contains twice the number of dogs at the right image, and at least two dogs in total are standing.

Visual question answering



What is the dog carrying?

Generating Grounded Descriptions

Visual semantic role labeling



feeding				
agent	food	source	eater	place
man	fish	hand	dolphin	pool

Visual commonsense graph



Auxiliary Text

Event: [Person1] is holding onto a bronze statue while waves of water crash around him.

Place: Inside a sinking ship

Before, [Person1] needed to...

- Realize the boat is sinking.
- See the water coming.
- Swim towards the statue.

Because, [Person1] wanted to...

- Save himself.
- Keep his head above water.
- Wait for help to arrive.

After, [Person1] will most likely...

- Scream for help.
- Regret boarding the ship.
- Get washed away.

Grounded Situated Natural Language Understanding



“...Walk straight, right before you reach the bed.”

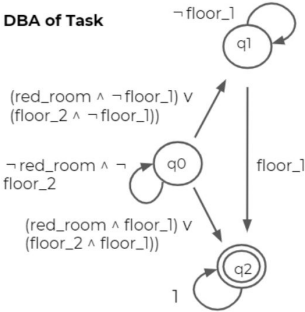


Touchdown instructions:
“Orient yourself so that the umbrellas are to the right. Go straight....”

Natural Language Command

First either go to the second floor or the red room and then go to the first floor

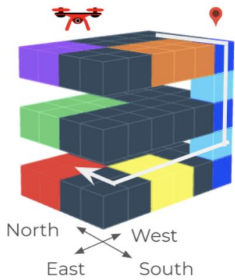
DBA of Task



LTL Task Specification

TASK: $F((floor_2 \vee red_room) \wedge F(floor_1))$

Executed Plan in Environment



Leader's view

...
 \bar{x}_3 : turn left and head toward the yellow hearts, but don't pick them up yet. I'll get the next card first.
 \bar{x}_4 : Okay, pick up yellow hearts and run past me toward the bush sticking out, on the opposite side is 3 green stars
 [Set made. New score: 4]
 ...

Related previous ACL tutorial:
http://multimodalnlp.github.io/mlr_tutorial.pdf

Generating coherence-aware descriptions for images

This work describes

- A coherent approach for characterizing **information-level inferences between images and text.**
- A controllable generation model that can provide descriptions for images with respect to the information goal.



Visible: horse and rider jumping a fence.

Meta: horse and rider jumping a fence during a race.

Subjective: the most beautiful horse in the world.

Story: horse competes in the event.

Multimodal Search-Based Dialogue

Multimodal search-based dialogue extends visually grounded question answering systems into multi-turn conversations with access to an external database.

- Dataset: Multimodal Dialogue ([Sasha et al., 2017](#))
- Generating responses in such systems requires data-driven models grounded in conversational text history, vision and knowledge base inputs.

U1 SHOPPER: Hi there

A1 AGENT: Hi, anything i can help you with today?

U2 SHOPPER: I am here to shop for tapered type casual trousers that my brother would like .

U3 SHOPPER: he is 26 years of age.

A2 AGENT: Oh. Good

A3 AGENT: let me just make a quick search through my catalogue

A4 AGENT:



U4 SHOPPER: Are the products in the 5th, 1st and 2nd images suited for multicoloured pocketed?

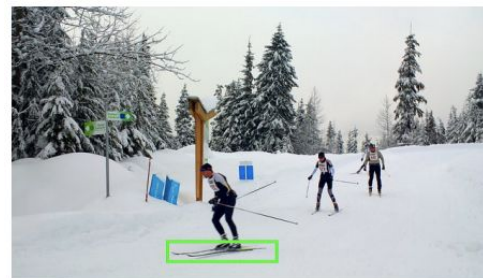
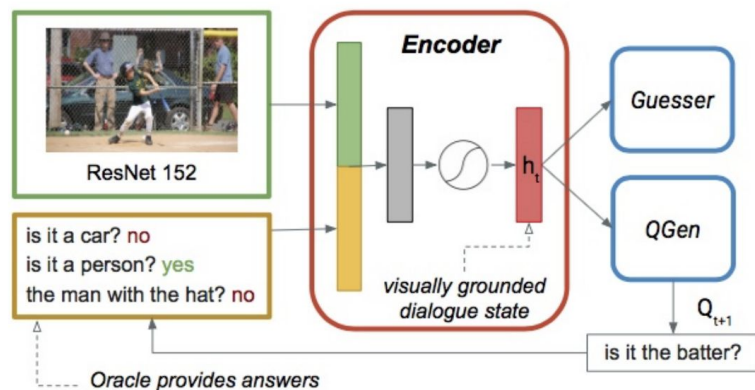
A5 AGENT: Yes

Example chatlog depicting multimodal user-agent interaction

Interactively Identifying an Object

How to integrate visual grounding with dialogue system components ?

- This work proposes a grounded dialogue state encoder.
 - A decision-making module decides which action needs to be performed next given the current dialogue state, i.e. whether to ask a follow-up question or stop the dialogue.



GDSE-CL





[*success*]

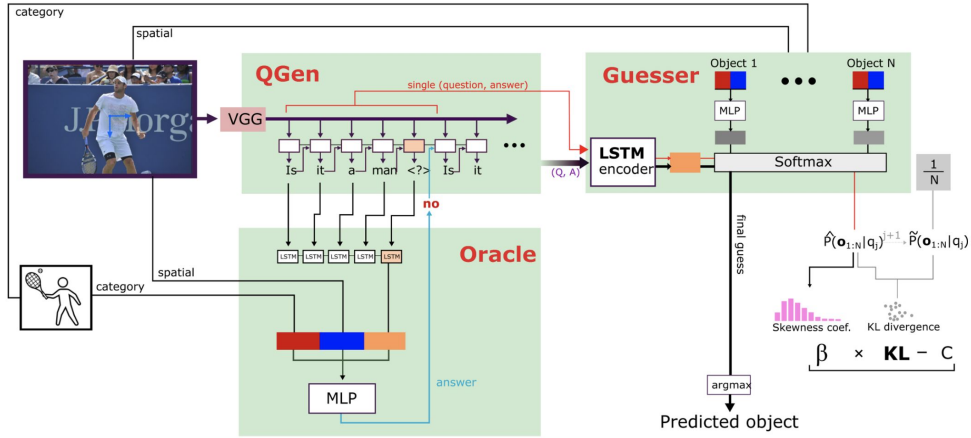
1. is it a person? no
2. is it a tree? no
3. is it a ski? yes
4. is it both skis of the person in front? yes

DM stops asking to guess

Interactively Identifying an Object

While challenging for machines, humans are good at asking effective questions and providing context dependent answers. This work's focus is on automatically generating a series of questions about an image with a single objective while reducing the search space for the agent.

Human	Robot
 	 
Is it an aircraft? no Is it on the lower part? yes Is it a vehicle? yes Is it the yellow vehicle? yes	Is it an aircraft? no Is it an aircraft? no Is it a vehicle? no Is it a wing? no Is it a person? no Is it a vehicle? yes
Predicted Object Yellow Vehicle	Predicted Object White Vehicle
Ground Truth Yellow Vehicle	Ground Truth Yellow Vehicle



Accuracy: 67.19%
 Human-level performance: 84.4%

Interactively Learning Visually Grounded Word Meaning

- Dataset: [BURCHAK](#)
- The best performing model integrates an incremental, semantic parsing/generation framework, Dynamic Syntax and Type Theory with Records, with a set of visual classifiers that are learned throughout the interaction and which ground the meaning representations that it produces.

(a) Multiple Dialogue Actions in one turn

L: so this shape is wakaki?

T: yes, well done. let's move to the color.

So what color is this?

(b) Self-Correction

L: what is this object?

T: this is a sako ... no no ... a suzuli burchak.

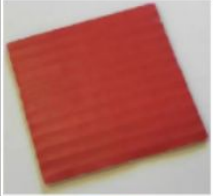
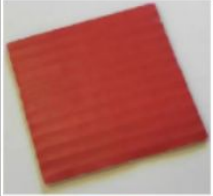

(c) Overlapping

T: this color [[is]] ... [[sa]]ko.

L: [[su]]zul[[i?]]

T: no, it's sako.

L: okay.

Dialogue Example		Final Semantics in TTR
T: what is this? S: a red circle? T: no, a red square. S: oh, okay.		$\left(\begin{array}{l} X_{=o1} : e \\ p2 : \text{red}(X) \\ p3 : \text{square}(X) \end{array} \right)$
T: what can you see? S: something orange. T: what shape is it S: a square. T: no, it's a circle. S: uhu		$\left(\begin{array}{l} X_{1=o2} : e \\ S=s : \text{per} \\ p : \text{circle}(X1) \\ p1 : \text{orange}(X1) \\ p2 : \text{see}(S,x1) \end{array} \right)$

Interim summary

We have discussed

- recent advances in language grounding that can potentially help in designing better multimodal dialogue systems.
 - Grounded natural language understanding
 - Controllable and informed generation
- Successful examples of systems that can learn perceptually grounded word meaning from human.

Learning Multimodal Grounding: Visual Dialogue and Other Datasets

From VQA to Visual Dialogue

Visual Dialogue requires an AI agent to hold a meaningful dialogue with humans in natural, conversational language about visual content. We will concentrate on identifying different grounding phenomena as identified in the first part of this tutorial.

1. What can current end-to-end systems do?
2. How can we extend these works to visual dialogue?
3. Why visual dialogue? How can it contribute to studying grounding in conversation?

History for Visual Dialog: Do we really need it?

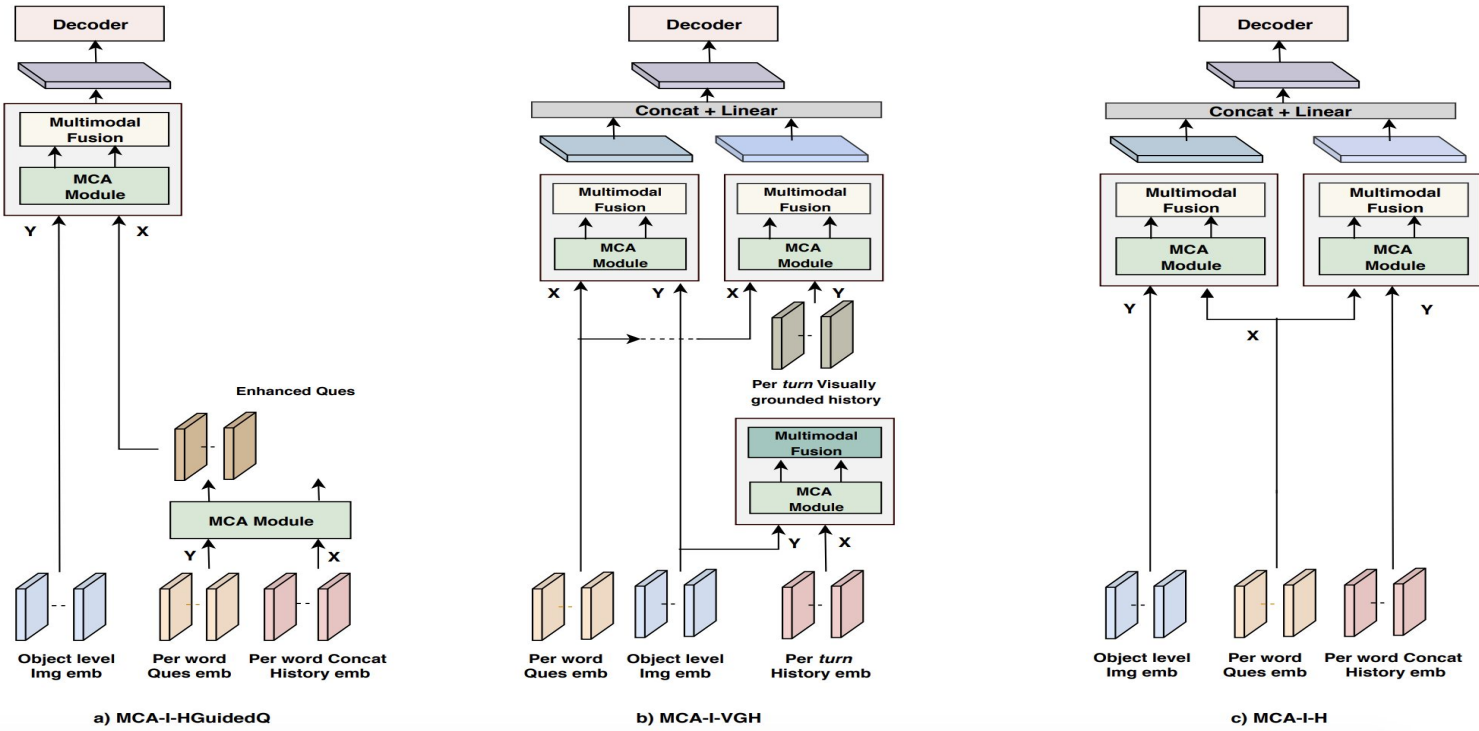
Competitive results on visual dialogue can be achieved by replicating the top performing model for VQA and effectively treating visual dialog as multiple rounds of question-answering, without taking history into account.



Caption	Current Question																
A group of skiers racing up a mountain	About how many?																
Conversational History / Context	Answer options																
Q1 Is 1 winning? A1 no. Q2 Do they have numbers? A2 yes.	<table border="1"><thead><tr><th>Answer options</th><th>Relevance</th></tr></thead><tbody><tr><td>• not really</td><td>0.0</td></tr><tr><td>• maybe 5 or 6, hard to see all of him</td><td>0.6</td></tr><tr><td>• 0 of those either</td><td>0.0</td></tr><tr><td>• few of them</td><td>0.4</td></tr><tr><td>• looks about 7</td><td>0.8</td></tr><tr><td>• 7 (GT answer)</td><td>0.4</td></tr><tr><td>.....</td><td>....</td></tr></tbody></table>	Answer options	Relevance	• not really	0.0	• maybe 5 or 6, hard to see all of him	0.6	• 0 of those either	0.0	• few of them	0.4	• looks about 7	0.8	• 7 (GT answer)	0.4
Answer options	Relevance																
• not really	0.0																
• maybe 5 or 6, hard to see all of him	0.6																
• 0 of those either	0.0																
• few of them	0.4																
• looks about 7	0.8																
• 7 (GT answer)	0.4																
.....																

Incorporating Dialogue History

A comparison of different architectures.



Modality-Balanced Models for Visual Dialogue

- Kim et al., propose to maintain two models and combine their complementary abilities for a more balanced multimodal model.
 - A large number of conversational questions can be answered by only looking at the image without any access to the context history.
 - Previous joint-modality models are more prone to memorizing the dialogue history.
 - Image-only models are more generalizable and perform substantially better.



Cap: down on a busy street, an oversized bus takes up half of a lane of traffic as cars zoom by on the other side

...
Q8: can you see a building
A8: yes 2 buildings
Q9: are they big
A9: yes numerous levels
Q10: can you see a pole
A10: yes a street pole

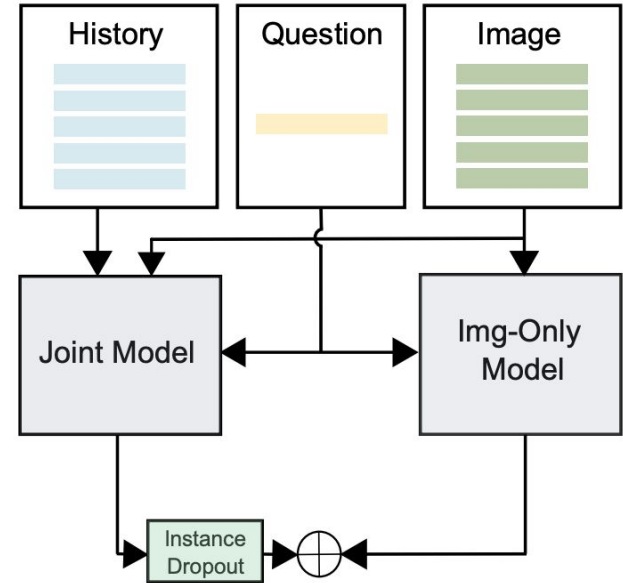


Cap: a decoration that looks like a traffic light next to plants

...
Q3: is there a lot of plants
A3: i only see 2
Q4: are they in pots
A4: yes
Q5: what color are they
A5: green
...

Modality-Balanced Models for Visual Dialogue

- Since each of the proposed models has different abilities, their complementary abilities are exploited together.
- Human Evaluation: Is image alone enough?
 - 100 images ~ 1000 questions
 - around 80% of the questions can be answered only from images.
 - using only history is not enough(only 1% of the questions can be answered).



Consensus Dropout Fusion.

Optimization Between Dialog Policy and Language Generation

- To improve dialogue generation, this work proposes to alternatively train an RL policy for image guessing and a supervised seq2seq model to improve dialog generation quality.
- The evaluation is on the GuessWhich task.

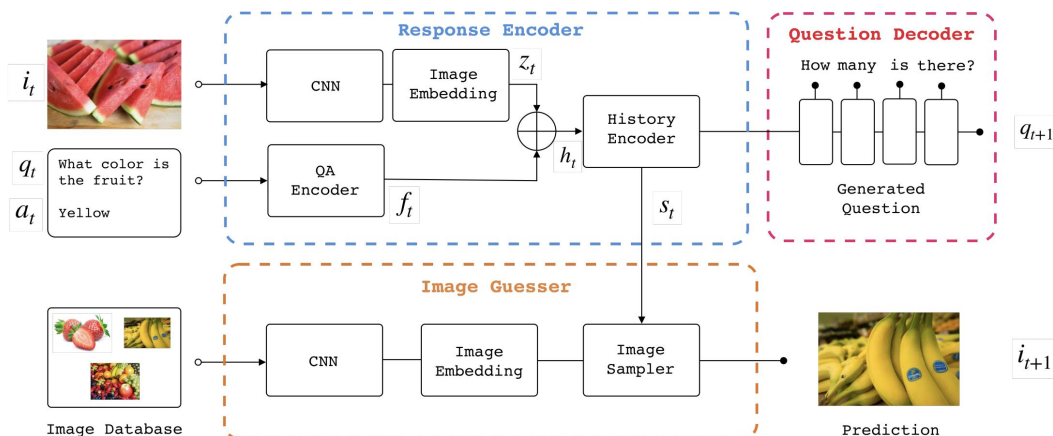


Figure 1: The proposed end-to-end framework of the conversation agent for GuessWhich task-oriented visual dialog task

Neural Multimodal Belief Tracker

Dataset: Multimodal Dialogue : Multimodal evidence can facilitate semantic understanding and dialogue state tracking. This work describes a belief tracker estimates the user’s goal at each step of the dialogue and provides a direct way to validate the ability of dialogue understanding.

1st system image attributes

Color	dark
Taxonomy	nightdress
Length	short
Material	cotton
Type	casual

user image attributes

Color	beige
Taxonomy	nightdress
Length	mini
Material	silk
Type	patchwork

Color	-
Taxonomy	nightdress
Length	-
Material	-
Type	-

$State_{t-1}$

Color	dark
Taxonomy	nightdress
Length	short
Material	cotton
Type	patchwork

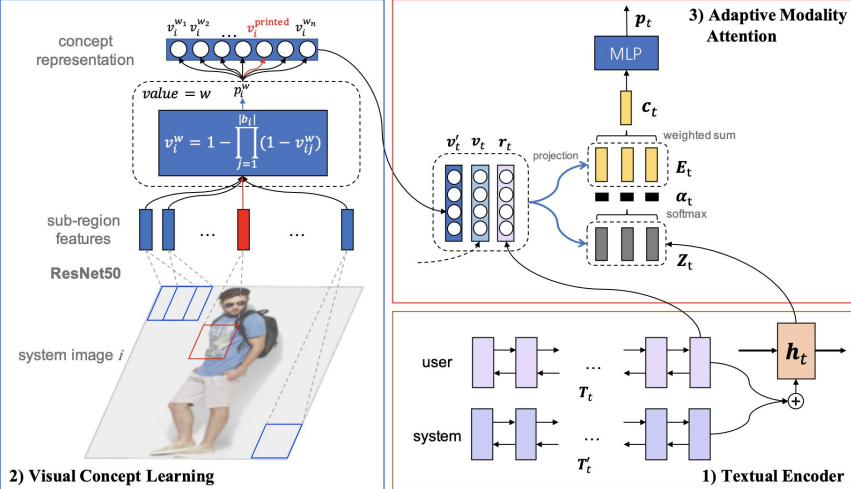
$State_t$

System:



User: I like the 1st image. Show me

something like it but in type as in this image



Limitations of Datasets

- VisDial dataset only contains very limited examples which require dialog history.
- Other goal-oriented visual dialog tasks, such as [GuessWhich?](#) and [GuessWhat?!](#) include more conversations that replicate natural dialog phenomena.
- However, there is very limited evidence that dialog history indeed matters for these tasks.
- Most of the current visual dialogue settings lack room for the occurrence of coordination phenomena prevalent in natural dialogue.



Human: Are these birds?

Bot: I don't see any

Human: Are there pigeons?

Bot: Yes

Human: But pigeons are birds!

Bot: Yes

Human: Are there birds?

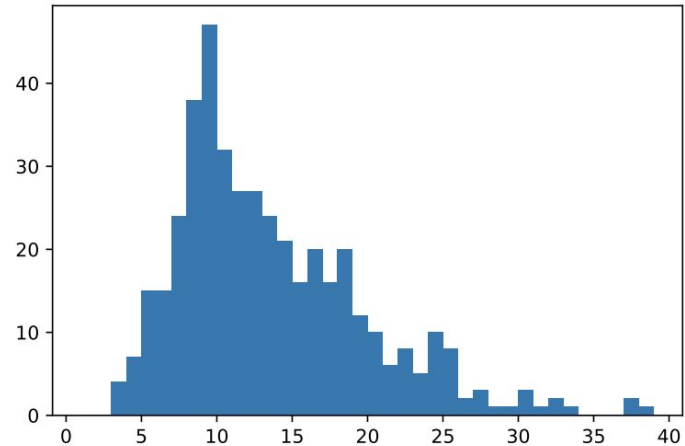
Bot: No

MeetUp! A Corpus of Joint Activity Dialogues

- MeetUp requires both visual and conversational grounding, and that makes stronger demands on representations of the discourse.
- It is a two-player coordination game where players move in a visual environment, with the objective of finding each other. They must talk about what they see, and achieve mutual understanding



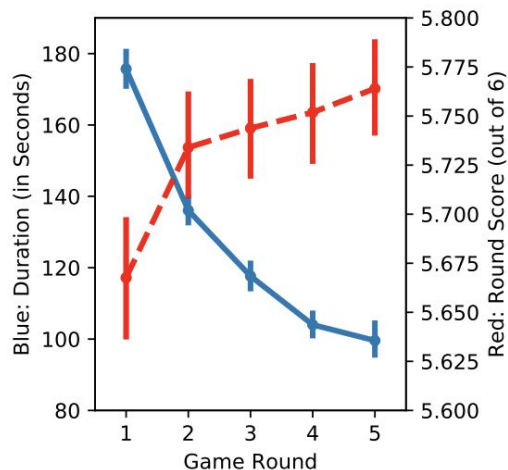
Interface



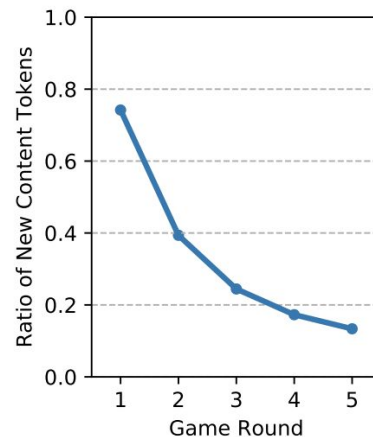
Histogram of number of turns per dialogue

The Photo Book Dataset

- Photo book is a large-scale collection of visually-grounded, task-oriented dialogues in English.
- It is designed to investigate shared dialogue history accumulating during conversation.
 - Participants repeatedly refer to a controlled set of target images.
 - This allows them to improve task efficiency if they utilise their developing common ground and establish conceptual pacts on referring expressions.



Average completion times (solid blue) and scores (dashed red) per game round.



Ratio of new content tokens over total content token count per round.

The Photo Book Dataset

- Utterances throughout a game, with final referring expressions are starkly different from both standard image captions and initial descriptions.
- More sophisticated models are needed to fully exploit shared linguistic history.



Reference chain with two segments:

- (1) A: *a woman sitting in front of a monitor with a dog wallpaper while holding a plastic carrot*
- (2) B: *carrot eating girl*
A: *no carrot eating girl on my end*

Segment to be resolved:

- (4) B: *I see the carrot lady again*



Reference chain with three segments:

- (1) A: *I have a strange bike with two visible wheels in the back*
- (2) B: *strange one*
- (3) A: *strange bike again yes*

Segment to be resolved:

- (4) B: *strange*

Summary: Adapting grounding to modality and culture

- Choice of the mode of communications exploits grounding.
 - Gestures
 - Mobile interfaces
 - Images and text
- Systems need to adapt to user's goals, abilities and preferences.
 - Tutoring systems
 - Systems that can work with infants
- We need to consider cross-cultural differences and language-dependent properties.

Summary: Language Grounding and Visual Dialogue

- State of the art models that can potentially improve dialogue systems.
 - Grounded situated natural language understanding.
 - Generating informed and contextually grounded descriptions.
- Current datasets and models are limited
 - They don't make use of context.
 - They are not linguistically rich.
 - They lack examples for modeling coordination in dialogue.

Future Directions

- Studying ways that we can incorporate uncertainty in systems decision making.
- Scaling up existing approaches with machine learning models.
- Improving turn-taking strategies in multimodal dialogue.
- Advancing multimodal generation techniques.
- Learning efficient discourse approaches for disambiguation and clarification in multimodal dialogue.
- Designing learning models for mediating common ground in conversation.