Natural Language Understanding and Semantic Parsing

Owen Rambow

Linguistics, Stony Brook University

Overview

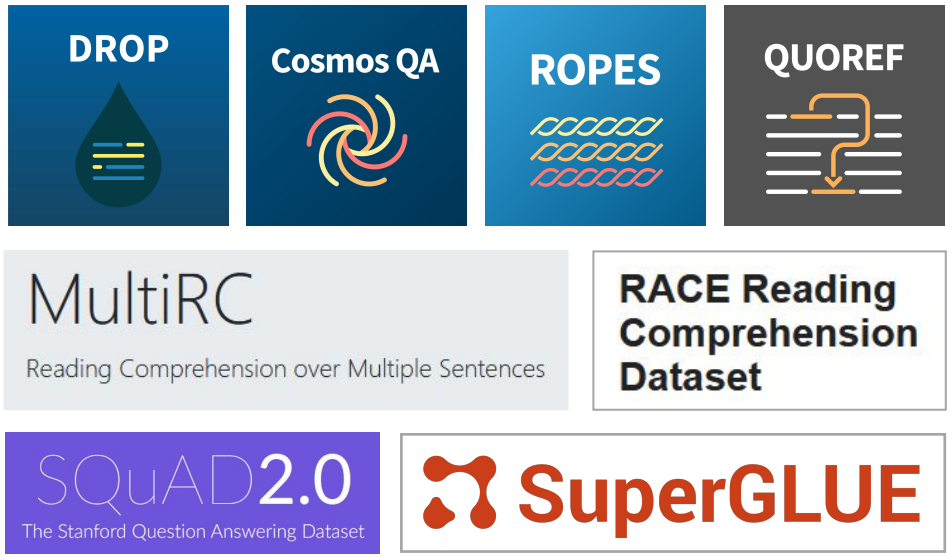
- Deep text understanding: what is it and how do we measure it
- Aspects of sentence meaning
- Semantic parsing

This Section Based on ACL2020 Paper

- Work with former colleagues from Elemental Cognition, a startup
- Jesse Dunietz, Gregory Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and David Ferrucci
  - To test machine comprehension, start by defining comprehension
- “Machine Reading Comprehension” = deep understanding of texts

Existing Benchmarks for Text Understanding

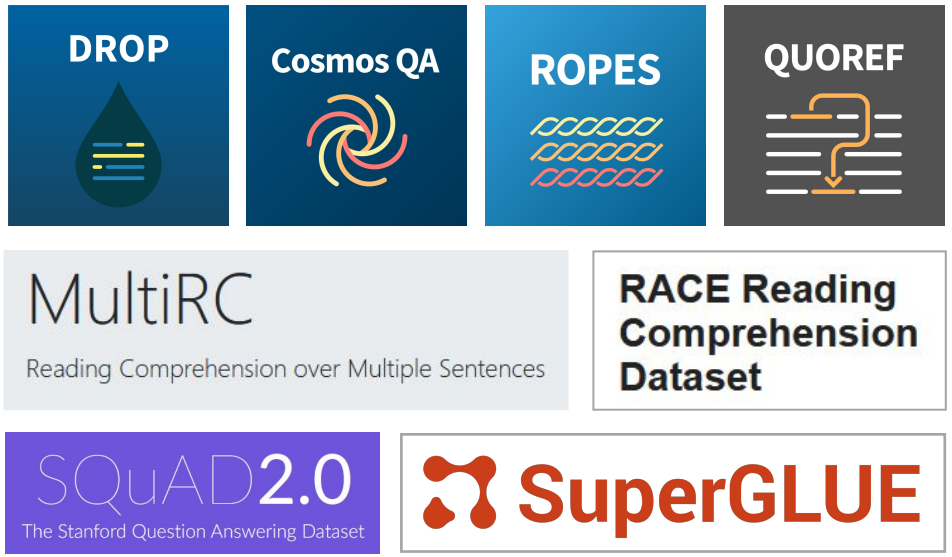
Many benchmarks for MRC



...

Existing Benchmarks for Text Understanding

But Do They Measure Performance of Systems for Specific Tasks?



?

- ✔ Answer questions about a clinical timeline
- ✔ Assess compliance with regulations or legal obligations
- ✔ Read a description of a game world and adjust a character's responses
- ✔ Read a human's directions and guide a robot to find the item requested
- ✔ ...

Outline of this Section

1. What is missing from the tests
2. Analyzing stories with a “template of understanding” (ToU)
3. SOTA systems fall short on our story ToU

Problems with Existing Text Understanding Benchmark Datasets

Sourcing method	Examples	Artificially easy	Focus on difficulty rather than content	Omit obvious questions	Retrospective passage selection
Manually written questions	TREC-8, SQuAD, SNLI, MNLI, NewsQA	✔		✔	
…+ tricky twists	DROP, ROPES, MultiRC, HotpotQA, CosmosQA…		✔	✔	
Naturally occurring questions	BoolQ, MS MARCO, ELI5			✔	✔
Tests designed for humans	TriviaQA, SearchQA, ARC, RACE	✘		✔	✔
Algorithmically generated questions	CNN/DM, ReCoRD, ComplexWebQuestions, WikiHop, bAbI	✘	✔		

Possible exceptions: ProPara, some bAbI tasks, TACIT project

Proposal: Establish what content systems will need to grasp

- Content = information expressed, implied, or relied on by the passage

Outline of this Section

1. What is missing from the tests
2. Analyzing stories with a “template of understanding” (ToU)
3. SOTA systems fall short on our story ToU

We propose defining a “template of understanding” for some genre of application-relevant texts

- Choose Genre
- Characterize what type of info needs to be extracted = “Template of Understanding” (ToU)
- Select texts
- Determine relevant passage content
- Design thorough tests for content

Stories: a promising genre of application-relevant texts for defining a ToU

✔ **Useful for many applications**

✔ **Strong evidence for what content applications will need**

Cognitive science research indicates that human readers attend to (Graesser et al., 1994; Zwaan et al., 1995):

- Locations
- Timeline
- Causes
- Motivations

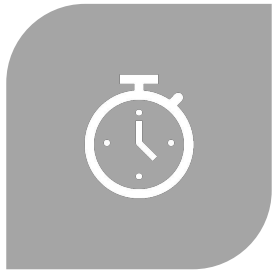
Spy icon by Adrien Coquet from the Noun Project

Our Template of Understanding for Stories

4 clusters of questions reflecting the content human readers attend to

SPATIAL

Where are entities positioned and how are they oriented throughout the story?



TEMPORAL

What events/states occur and with what timing?

CAUSAL

How do events/states mechanistically lead to other events/states described or implied by the text?

MOTIVATIONAL

How do agents' beliefs, desires, goals, and emotions lead to their actions?

Evaluation approach: explicitly annotate answers

"Records of Understanding" (RoUs)

Sample story fragment:
One day, it was raining. When Allie arrived, Rover ran out the door. He barked when he felt the rain. He ran right back inside.

Spatial *(sample entries)*:

- Rover is in the yard from when he runs out the door until he runs inside.
- Rover is in the house from when he runs inside until the end of the story.

Temporal *(sample entries)*:

- Allie arrives just before Rover runs outside.
- It is still raining at the end of the story.

Motivational *(sample entry)*:

- Rover runs inside, rather than staying put, because:
  - Rover has the goal of not getting rained on, because:
    - Rover is getting rained on.
      - It is raining.
      - When it is raining, things that are outside tend to get rained on, whereas things inside do not.
    - Rover does not like getting rained on.
      - …
  - He forms a plan to achieve his goal:
    - If he runs inside, he will be inside.
    - If Rover is inside, he will not get rained on.
      - When it is raining, things that are inside tend to not get rained on (?).

How to Use a Record of Understanding in an Evaluation?

- Internal use at Elemental Cognition: system produces a specific KR which we also use to encode the RoU
- Proposed cross-system evaluation: System produces natural language output for the questions
  - Human judges compare against gold RoU
    - Trained judges
    - Crowdsourcing
  - Use Pyramid method (Nenkova & Passonneau 2004) to account for valid variation in correct answers (multiple RoUs)
  - Use some technology to automate evaluation
    - Simple, such as BLEU
    - More complex using (semantic) parsing
  - Issue very similar to evaluation for other NLP problems (MT, summarization, …)
- Also, other options (see ACL paper)

Outline of this Section

1. What is missing from the tests
2. Analyzing stories with a “template of understanding” (ToU)
3. SOTA systems fall short on our story ToU

Case Study: RACE Stories (Lai et al., 2017)

- English exams for middle and high school Chinese students, ages 12 to 18
- Texts and questions
- Best system performance in 2017 paper:
  44% against 95% for humans
- We use first two stories from dev set

The content-first approach is rarely followed by existing MRC benchmark datasets

Sourcing method	Examples	Artificially easy	Focus on difficulty rather than content	Omit obvious questions	Retrospective passage selection
Manually written questions	TREC-8, SQuAD, SNLI, MNLI, NewsQA	✔		✔	
…+ tricky twists	DROP, ROPES, MultiRC, HotpotQA, CosmosQA…		✔	✔	
Naturally occurring questions	BoolQ, MS MARCO, ELI5			✔	✔
Tests designed for humans	TriviaQA, SearchQA, ARC, RACE	✘		✔	✔
Algorithmically generated questions	CNN/DM, ReCoRD, ComplexWebQuestions, WikiHop, bAbI	✘	✔		

Possible exceptions: ProPara, some bAbI tasks, TACIT project

Small multiple-choice dataset based on our Template of Understanding

RoU (answers to ToU questions)

Spatial:

- …

Temporal:

- …

Causal:

- …

Motivational:

- …

Q1a) What actually happened when Mr. Green and the man drove together?

A. They came to a small house.
B. They came to a hotel.
C. They traveled around the country.
D. They stopped several times at the side of the road.

Q1b) How did the man's directions actually turn out?

A. The directions the man gave led to where Mr. Green wanted to go.
B. The directions the man gave led to where the man wanted to go.
C. The directions Mr. Green gave led to where the man wanted to go.
D. The directions Mr. Green gave led to where Mr. Green wanted to go.

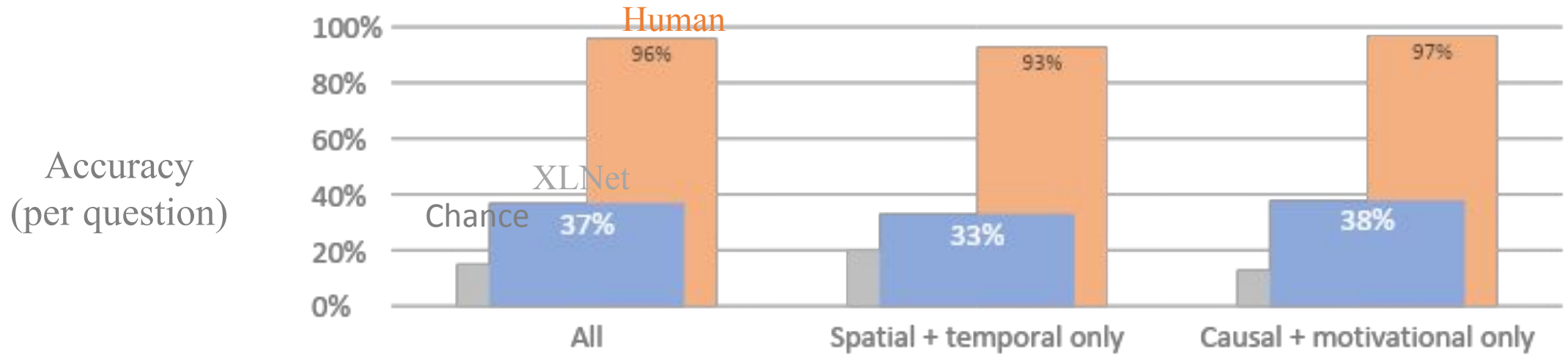
201 questions (91 "variant groups")

System We Use: XLNet

- Yang et al., 2019
- Transformer-based language model
  - Trained on BooksCorpus and English Wikipedia
- 82% on RACE task (within 5% at publication time of top performer)
  - Up from 44% in two years!

XLNet performs poorly on ToU-based questions

That's even within the guardrails of multiple-choice.



Summary of this Section

- We can define deep understanding based on “Templates of Understanding”
- ToUs represent knowledge relevant for a text genre
  - Stories are ubiquitous across applications
- Current machine learning technology may produce good results on existing text understanding tests, but bad results on deep understanding tests
- ToUs can serve as a basis for a cross-system evaluation

Overview

- Deep text understanding: what is it and how do we measure it
- Aspects of sentence meaning
- Semantic parsing

Aspects of Sentence Meaning

- Idea: need to process each sentence to build up meaning of text
- Deep understanding (previous section) =
  - Meaning of all sentences in text
    + common sense/background knowledge
    + inference
- Goal of this section: highlight complexity of task of extracting sentence meaning

Word Sense

- Same word can have different meanings:
  - John bought a car
  - John bought the story
- Can use context to do word sense disambiguation (WSD)
- For NLP: Need an inventory of word senses and annotated corpora
- Active area in NLP

Semantic Predicate-Argument Structure: Who Did What to Whom

- Given a meaning for a predicate (usually, a verb), can we determine how other elements in the sentence relate to it?
- Example:
  - John bought a car for $3,000 from his sister
  - Buyer = John
  - Seller = his sister
  - Price = $3,000
- WSD + Semantic Predicate-Argument Structure = Semantic Parsing
  - Will discuss in more detail

Coreference

- Need to resolve co-reference between NPs
  - Paul was poor. John bought him/himself a car.
  - I took my dog to the vet yesterday. He bit him in the hand. (Sidner 1979)
  - I took my dog to the vet yesterday. He gave him an injection.
- Active area in NLP, many annotated corpora in many languages
- Event co-reference much more complex
  - Liam left for Limerick. His departure devastated us.

Quantifier Scope

- Scope of quantifiers with respect to each other constrained by syntax but not determined
  - Every student learned a song from a local resident
  - $\exists$ x, x a song; $\exists$ y, y a local resident; $\forall$ z, z a student: learn(z,x,z)
  - $\exists$ x, x a song; $\forall$ z, z a student; $\exists$ y, y a local resident: learn(z,x,z)
  - $\exists$ y, y a local resident; $\forall$ z, z a student; $\exists$ x, x a song: learn(z,x,z)
  - $\forall$ z, z a student; $\exists$ y, y a local resident; $\exists$ x, x a song: learn(z,x,z)
- Not a focus in NLP currently but important for inference
  - Three students learned a song from a local resident – how many songs were learned?
- Gets more complex with event variable
  - Three men carried four pianos – how many carrying events? 1, 3, 4, 12?
- Not clear that always fully disambiguated in human communication

Implicatures

- An implicature is not a logical inference, but a conclusion the audience can draw from what was NOT said
  - I bought two pencils
    - => I did not buy three pencils, because if I had, I would have said so
  - Sandy is a friend
    - => Sandy is not my lover or my spouse because if so, I would have said so
  - Contrast: Sandy is a bore/a psychologist/Senegalese
    - Sandy may well be a friend, my lover or my spouse, or not
- Based on assumptions about rational conversation ("Grice's maxims")
- Contribute important aspects of meaning
- Not an active area in NLP, but some formal work on modeling

Modeling Cognitive State

- Language does not only convey propositional content, but also the attitude of the author towards the propositional content
- Attitudes: belief and sentiment
- Examples

Sentence	Belief	Sentiment
John will leave tomorrow	1	?
John may leave tomorrow	.5	?
I hope John will leave tomorrow	.5	1
Mary claims John will leave tomorrow	-.5	?
Mary says John will leave tomorrow	?	?

- Sentiment analysis is a massive area in NLP, Belief less so
- Active area of interest for me

Overview

- Deep text understanding: what is it and how do we measure it
- Aspects of sentence meaning
- Semantic parsing
  - Ontologies
  - Experiments

Semantic Parsing: Steps

- Choose **ontology** which has argument structure
- Determine content words
  - Complexity about multi-word expressions
- For each content word (“trigger”):
  - Do word sense disambiguation (WSD) with respect to the ontology
  - Identify semantic arguments as text spans or by syntactic head word
  - Identify semantic role label (semantic role labeling = SRL)
- In theory, semantic parses are assembled into a single structure
  - Not always done; Abstract Meaning Representation most successful in this respect (based on PropBank)

Ontologies

- Ontology = inventory of what there is
  (ὄντος, *ontos*, 'being' or 'that which is')
- Consists of:
  - Terms designating classes of things/situations/events
  - Relations between terms
    - is-a
    - part-of
    - …
  - Additional information
    - Argument structure: *someone gives something to someone* Important for Semantic Parsing
    - Attributes: objects have a color
    - Typical attribute values: an elephant is usually grey

PropBank

- Palmer et al. 2005
- Traditional lexicographic approach to word sense:
  - For each word, list senses (= predicates, or “role sets”)
- Add semantic roles
  - Argument roles are predicate-specific
    - Except ARG0 and ARG1 always correspond to agent and patient
  - Adjunct roles (time, place, manner) are identical across predicates

PropBank Example: *buy*

- Roleset id: buy.01 , *purchase*
- Roles:
  - Arg0-PAG: *buyer*
  - Arg1-PPT: *thing bought*
  - Arg2-DIR: *seller*
  - Arg3-VSP: *price paid*
  - Arg4-GOL: *benefactive*
- Roleset id: buy.05 , *accept as truth*
- Roles:
  - Arg0-PAG: *believer*
  - Arg1-PPT: *thing believed*

- Roleset id: bring.01 , *carry along with*
- Roles
  - Arg0-PAG: *bringer*
  - Arg1-PPT: *thing brought*
  - Arg2-GOL: *benefactive or destination brought-for, brought-to*
  - Arg3-PRD: *attribute, state after bringing, secondary action*
  - Arg4-DIR: *ablative, brought-from*

Meaning of roles ARG2 and greater predicate-specific!

PropBank Example: *buy*

- Roleset id: buy.01 , *purchase*
- Roles:
  - Arg0-PAG: *buyer*
  - Arg1-PPT: *thing bought*
  - Arg2-DIR: *seller*
  - Arg3-VSP: *price paid*
  - Arg4-GOL: *benefactive*
- Roleset id: buy.05 , *accept as truth*
- Roles:
  - Arg0-PAG: *believer*
  - Arg1-PPT: *thing believed*

- Roleset id: scare.01 , *(cause to) become afraid, afraid*
- Roles
  - Arg0-PAG: *scary entity*
  - Arg1-PPT: *scared entity*
  - Arg2-MNR: *instrument (if separate from arg0)*
  - Arg3-EXT: *intensifier, extent*

Precise meaning of roles ARG0 and ARG1 can also be predicate-specific!

FrameNet

- Chuck Fillmore; Baker et al. 1998
- Starts with a situation, such as a commercial transaction, and asks how language can express it:
  - John **bought** a car for $3,000 from his sister
  - John **paid** his sister $3,000 for a car
  - John's sister **sold** him a car for $3,000
  - $3,000 **bought** John a car
  - A car **cost** John $3,000

FrameNet: Meaning 1 of *buy*

Commercial_transaction

Lexical Unit Index

Definition:

These are words that describe basic commercial transactions involving a Buyer and a Seller who exchange Money and Goods. The individual words vary in the frame element realization patterns. For example, the typical patterns for the verbs buy and sell are: BUYER buys GOODS from the SELLER for MONEY. SELLER sells GOODS to the BUYER for MONEY.

His $20 TRANSACTION with Amazon.com for a new TV had been very smooth.

FEs:

Core:

Buyer [Byr]	The Buyer wants the Goods and offers Money to a Seller in exchange for them.
Goods [Gds]	The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
Money [Mny]	Money is the thing given in exchange for Goods in a transaction.
Seller [Slr]	The Seller has possession of the Goods and exchanges them for Money from a Buyer.

FrameNet: Meaning 2 of *buy*

Fall_for

Lexical Unit Index

Definition:

A Victim comes to have incorrect beliefs as a result of exposure to a Deception. The Deception is necessarily orchestrated by one or more agents, though these are not expressible by the LUs in this frame.

The public was too clever to FALL FOR the type of propaganda prevalent in London at the time.

FEs:

Core:

Deception [dec]	The Deception is perceived by the Victim, leading to false beliefs. The Deception, orchestrated by an agent or agents with the intention of deception, may be any misinformation, event, or situation perceived by the Victim. Even the smartest person can fall for a cleverly-designed scam.
Victim [vic]	The Victim has false beliefs as a result of exposure to a Deception. Seventeen respected scientists FELL FOR the lie before it was revealed as a hoax.

Comparison

PropBank

- Word-specific meanings
- Small label inventory for semantic relations
- Meaning of labels not uniform across predicates (even if many people think they are!)
- Good coverage of verbs and event nouns in several languages
- Lots of annotated texts

FrameNet

- Meaning defined by situations
- Labels of semantic relations specific to frames
- Meaning of labels specific to frames
- Ok coverage of verbs and event nouns but gaps
- Most annotations are example sentences not in context

NLP has favored Propbank

Hector

- Ontology created at Elemental Cognition (Ariel Diertani)
- We like the situation-focus of FrameNet
- FrameNet has terrible coverage of entities
  - Use FrameNet only for events and states
    - But we expanded its is-a hierarchy
  - Use NOAD (New Oxford American Dictionary) for entities, which provides an excellent is-a hierarchy
- Work in progress

Overview

- Deep text understanding: what is it and how do we measure it
- Aspects of sentence meaning
- Semantic parsing
  - Ontologies
  - Experiments

Joint Work with Former Colleagues from Elemental Cognition

- Joint work with Aditya Kalyanpur (lead), Or Biran, Tom Breloff, Jennifer Chu-Carroll, Ariel Diertani, and Mark Sammons
- Warning: Work in Progress
  - Some methodological infelicities: missing experiments that mean we can't always understand what contributes to improvements

Experiment 1: Genrative

- Comparison of GPT-2 language model and T5 encode-decoder model
  - Bigger is better
  - T5 performs far better
- Ask me if interested in details

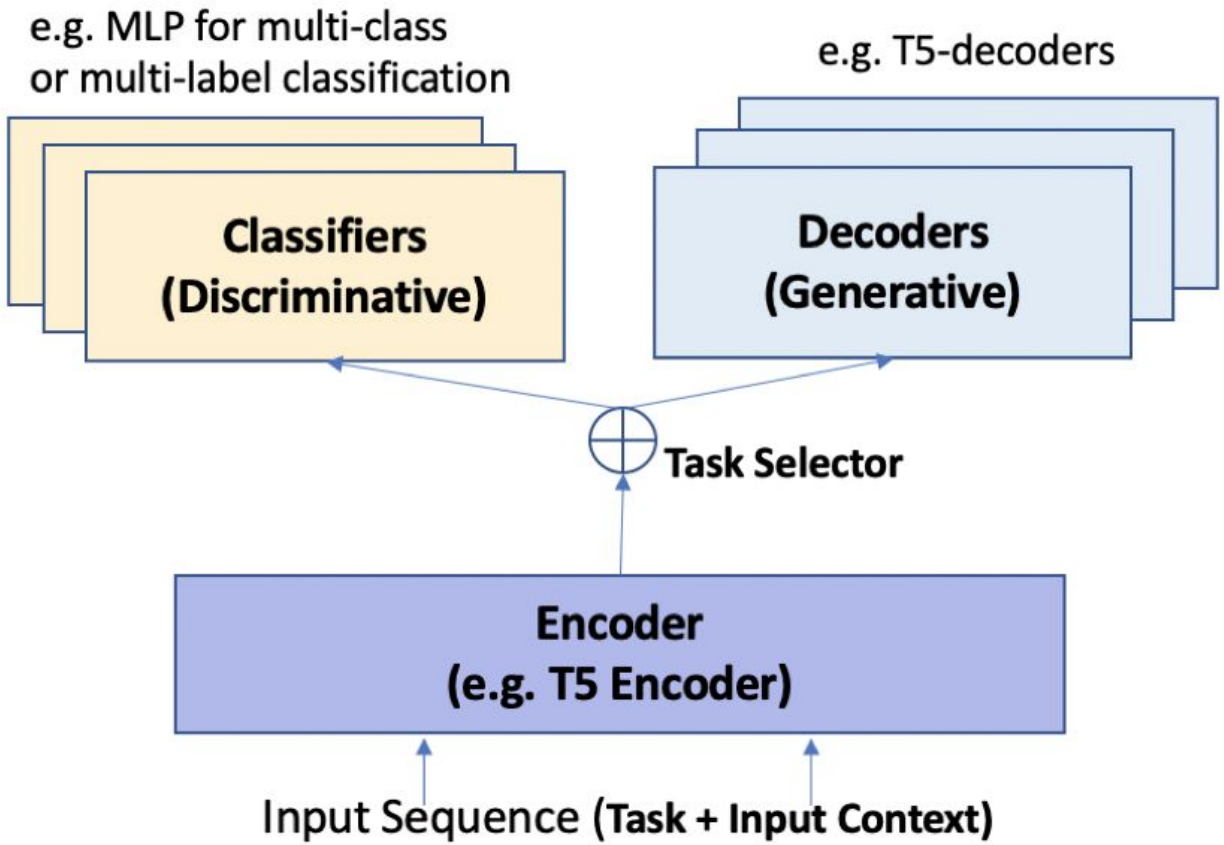
Experiment 2: Multi-Task Learning

- 2 models, trained together
- Input 1: sentence and trigger marked in sentence, and with position numbers
- Output 1: frame name (= word sense)
- Input 2: sentence and trigger marked in sentence, and with position numbers and with frame name
- Output 2: roles and spans for roles
- Ontology used: Hector
- Architectures used:
  - T5 encoder-decoder model

Data Format

Input	Output
FRAME: 0 Two 1 of 2 the 3 cast4 fainted 5 and 6 most 7 of 8 the 9 rest 10 * repaired * 11 to 12 the 13 nearest 14 bar 15 .	Self motion
ARGS for Self motion: 0 Two 1 of 2 the 3 cast 4 fainted 5 and 6 most 7 of 8 the 9 rest 10 * repaired * 11 to 12 the 13 nearest 14 bar 15 .	Self mover = 6- 9| Goal=11-14 |
FRAME: 0 He 1 blinked 2 , 3 taken 4 aback 5 by 6 the 7 * vigour * 8 of 9 her 10 outburst.	Dynamism
ARGS for Dynamism: 0 He 1 blinked 2 , 3 taken 4 aback 5 by 6 the 7 * vigour * 8 of 9 her 10 outburst.	Action = 8-10 |
FRAME: 0 The rain 1 * dripped * 2 down 3 his 4 neck.	Fluidic motion
ARGS for Fluidic motion: 0 The rain 1 * dripped * 2 down 3 his 4 neck.	Fluid = 0-1 | Path=2-4|

Architecture



Results (All with T5)

Generative Model	Frame Accuracy	Role Precision	Role Recall	Role F1
Generative	87%	81%	83%	82%
Multi-Task	90%	85%	83%	84%

Experiment 3: Multi-Task on Standard Ontologies and Datasets

- Input 1: sentence and trigger marked in sentence, and with position numbers
- Output 1: frame name (= word sense)
- Input 2: sentence and trigger marked in sentence, and with position numbers and with frame name
- Output 2: roles and spans for roles
- Ontologies used: **PropBank and FrameNet**
- Architectures used:
  - T5 encoder-decoder model

PropBank Experiments

- Systems
  - He et al. 2018
  - Li et al. 2019
  - Full-Gen: our one-step generative system
  - Multi-Task: our two-step multi-task system
- Data: CoNLL 2012 data
- Metrics
  - CoNLL eval script: evaluates (predicate, role) pairs with exact match on role span

Results: PropBank Frame Prediction (Test Set)

System	CoNLL Metric
He at al. 2018	82.9%
Li et al. 2019	83.1%
Full-Gen	82.3%
Multi-Task	83.7%

FrameNet Experiment

- Systems
  - Sesame: state of the art (Swayamdipta et al. 2017)
  - Full-Gen: our one-step generative system
  - Multi-Task: our two-step multi-task system
- Data: FN1.7 evaluation data
- Conditions:
  - Gold: the gold frame is given
  - Pred: the frame is also predicted
- Metric:
  - Exact match: roles must match on span AND role name

Results: FrameNet Frame Prediction (Test Set)

System	Frame Accuracy
Sesame	86.5%
Full-Gen	87.0%
Multi-Task	87.5%

Results: FrameNet Role Prediction (Test Set)

System	Gold-Test			Pred-Test		
	P	R	F	P	R	F
Sesame	62%	55%	58%	57%	49%	52%
Full-Gen	71%	73%	72%	63%	65%	64%
Multi-Task	75%	76%	76%	66%	67%	66%

Discussion

- Our T5-based Multi-Task approach beats state of the art
- Much simpler, as no ad-hoc neural architecture
  - Little parameter tuning
- Error analysis: 25% (FrameNet) to 31% (PropBank) of errors are gold errors, or plausible responses

Conclusion to Talk

- A long way to deep text understanding
  - But maybe we can start thinking about defining the problem, and evaluating systems
- Even at the sentence level, we are only starting to cover the full range of meaning we need to access
- For semantic parsing, we have good resources and pretty good results
  - More work needed
- In general, we need to understand what the linguistic resources are actually encoding
- My personal goals at Stony Brook:
  - Continue working on Hector ontology
  - Continue working on semantic parsing
  - Expand work on cognitive state detection

Additional Slides

Results: FrameNet Role Prediction (Test Set)

Metric	System	Gold-Test			Pred-Test		
		P	R	F	P	R	F
Exact	Sesame	62%	55%	58%	57%	49%	52%
	Full-Gen	71%	73%	72%	63%	65%	64%
	Multi-Task	75%	76%	76%	66%	67%	66%
Soft	Sesame	71%	64%	67%	63%	56%	59%
	Full-Gen	78%	80%	79%	69%	71%	70%
	Multi-Task	80%	82%	81%	71%	72%	71%

Metrics

- Exact match: roles must match on span AND role name
- Soft match: we calculate recall-precision on token basis

Results: FrameNet

Metric	System	Gold-Dev			Gold-Test			Pred-Dev			Pred-Test		
		P	R	F	P	R	F	P	R	F	P	R	F
Exact	Sesame	60%	51%	55%	62%	55%	58%	55%	47%	51%	57%	49%	52%
	Full-Gen	71%	73%	72%	71%	73%	72%	65%	66%	66%	63%	65%	64%
	Multi-Task	77%	77%	77%	75%	76%	76%	71%	70%	71%	66%	67%	66%
Soft	Sesame	71%	61%	66%	71%	64%	67%	64%	56%	60%	63%	56%	59%
	Full-Gen	80%	81%	80%	78%	80%	79%	73%	74%	73%	69%	71%	70%
	Multi-Task	83%	82%	82%	80%	82%	81%	75%	75%	75%	71%	72%	71%

“Variant Groups”

Q1a) What actually happened when Mr. Green and the man drove together?

A. They came to a small house.
B. They came to a hotel.
C. They traveled around the country.
D. They stopped several times at the side of the road.

Q1b) How did the man’s directions actually turn out?

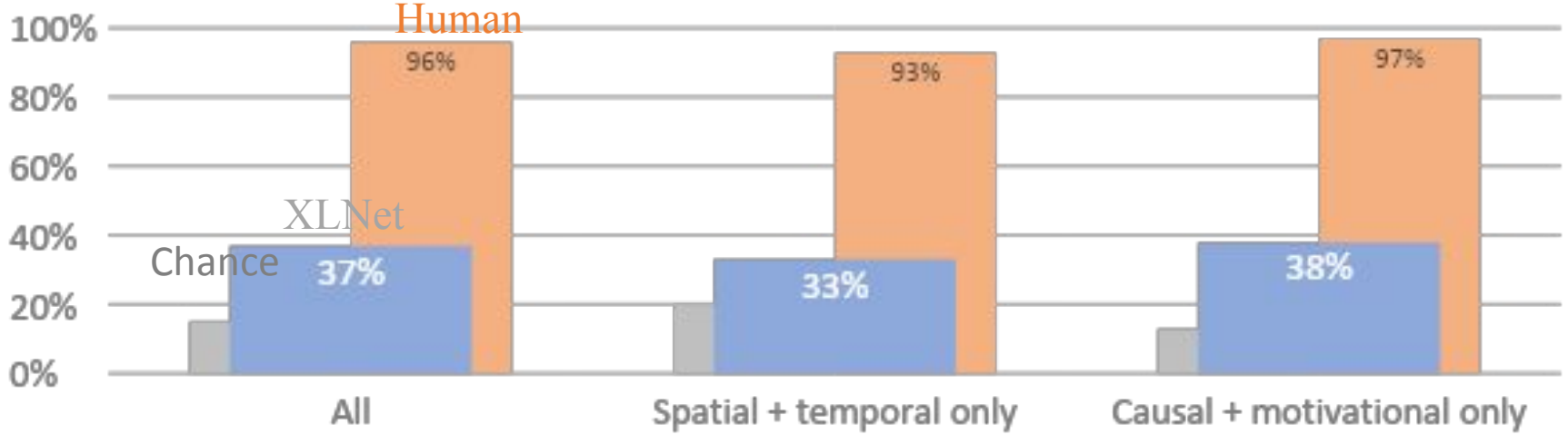
A. The directions the man gave led to where Mr. Green wanted to go.
B. The directions the man gave led to where the man wanted to go.
C. The directions Mr. Green gave led to where the man wanted to go.
D. The directions Mr. Green gave led to where Mr. Green wanted to go.

- Variant group = same question, but formulated very differently
  - More than just paraphrase
- Same content in correct answer
- Idea: if you understand this part of the story, you can answer all questions in the variant group

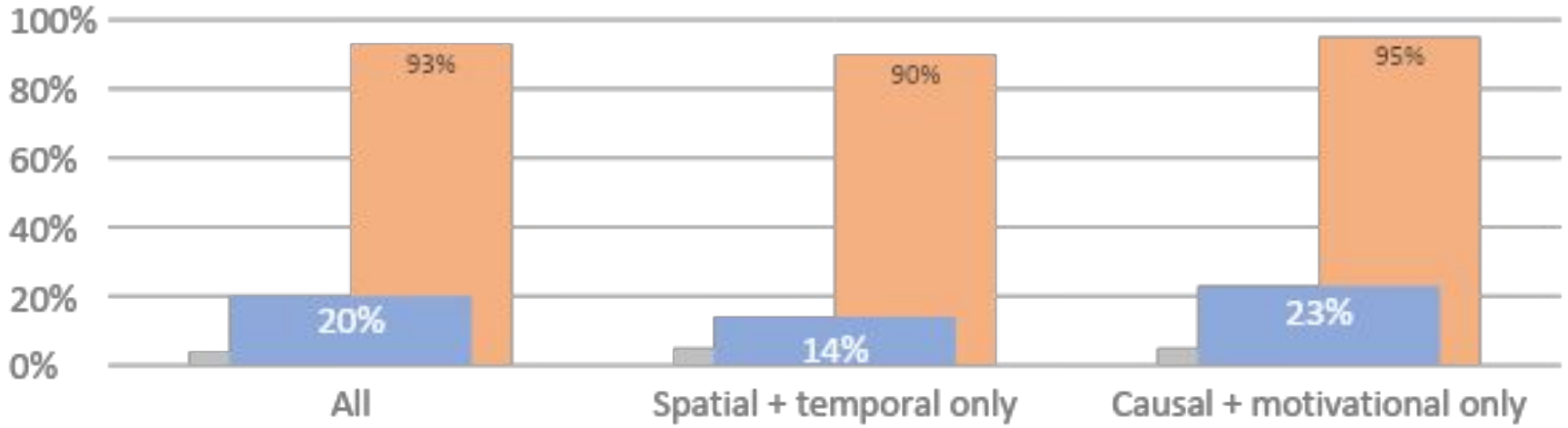
XLNet performs poorly on ToU-based questions

That's even within the guardrails of multiple-choice.

Accuracy
(per question)



Accuracy
(per "variant group")



Results

Generative Model	Parameters	Frame Accuracy	Role Precision	Role Recall	Role F1
GPT2-small	117M	77%	60%	59%	59%
GPT2-medium	345M	79%	73%	71%	72%
GPT2-large	770M	82%	77%	76%	77%
T5-small	120M	82%	77%	81%	79%
T5-base	440M	87%	81%	83%	82%

Note: no comparable results yet as this uses our own Hector!

Experiment 1: Generative

- Input: sentence and trigger marked in sentence
- Output: full description of semantic parse
- Ontology used: Hector
- Resources used:
  - GPT-2 language model
  - T5 encoder-decoder model
- Training data (for fine-tuning): 1.4 instances of word analyses

Data Format

Input	Output
Two of the cast fainted and most of the rest * repaired * to the nearest bar.	repaired = *Self motion* | most of the rest = *Self - mover* | to the nearest bar = *Goal* |
He blinked , taken aback by the * vigour * of her out- burst.	vigour = *Dynamism* | of her outburst = *Action* |
The rain * dripped * down his neck.	dripped = *Fluidic motion* | The rain = *Fluid* | down his neck = *Path* |
She * adored * shopping for bargains and street markets and would have got on well with Cherry.	adored = *Experiencer focus* | She = *Experiencer* | shopping for bargains and street markets = *Content* |
He * cleared * his throat as the young man looked up.	cleared = *Emptying* | He = *Agent* | his throat = *Source* |

Results

Generative Model	Parameters	Frame Accuracy	Role Precision	Role Recall	Role F1
GPT2-small	117M	77%			
GPT2-medium	345M	79%			
GPT2-large	770M	82%			
T5-small					
T5-base					

Results

Generative Model	Parameters	Frame Accuracy	Role Precision	Role Recall	Role F1
GPT2-small	117M	77%	60%	59%	59%
GPT2-medium	345M	79%	73%	71%	72%
GPT2-large	770M	82%	77%	76%	77%
T5-small					
T5-base					

Note: role eval is exact match on span and role label!

Results

Generative Model	Parameters	Frame Accuracy	Role Precision	Role Recall	Role F1
GPT2-small	117M	77%	60%	59%	59%
GPT2-medium	345M	79%	73%	71%	72%
GPT2-large	770M	82%	77%	76%	77%
T5-small	120M	82%			
T5-base	440M	87%			

Results

Generative Model	Parameters	Frame Accuracy	Role Precision	Role Recall	Role F1
GPT2-small	117M	77%	60%	59%	59%
GPT2-medium	345M	79%	73%	71%	72%
GPT2-large	770M	82%	77%	76%	77%
T5-small	120M	82%	77%	81%	79%
T5-base	440M	87%	81%	83%	82%

Note: no comparable results yet as this uses our own Hector!

Discussion

- Bigger model is better
- Encoder-decoder model T5 better than language model GPT-2

Discussion

- Methodological problem: we changed BOTH the input format (numbers) AND the approach (generative vs multi-task)
  - Unclear which contributed to the (small) improvement

PropBank Experiments

- Systems
  - He et al. 2018
  - Li et al. 2019
  - Full-Gen: our one-step generative system
  - Multi-Task: our two-step multi-task system
- Data: CoNLL 2012 data
- Metrics
  - CoNLL eval script: evaluates (predicate, role) pairs with exact match on role span (like our Exact Match metric for FrameNet)

Results: PropBank Frame Prediction (Test Set)

System	Frame Accuracy
He at al. 2018	82.9%
Li et al. 2019	83.1%
Full-Gen	82.3%
Multi-Task	83.7%