STRUCTURAL DATA TO TEXT GENERATION: MODELING AND EVALUATION

Yao Fu, yao.fu@columbia.edu
Columbia University
April 29th 2019

I. MODELING

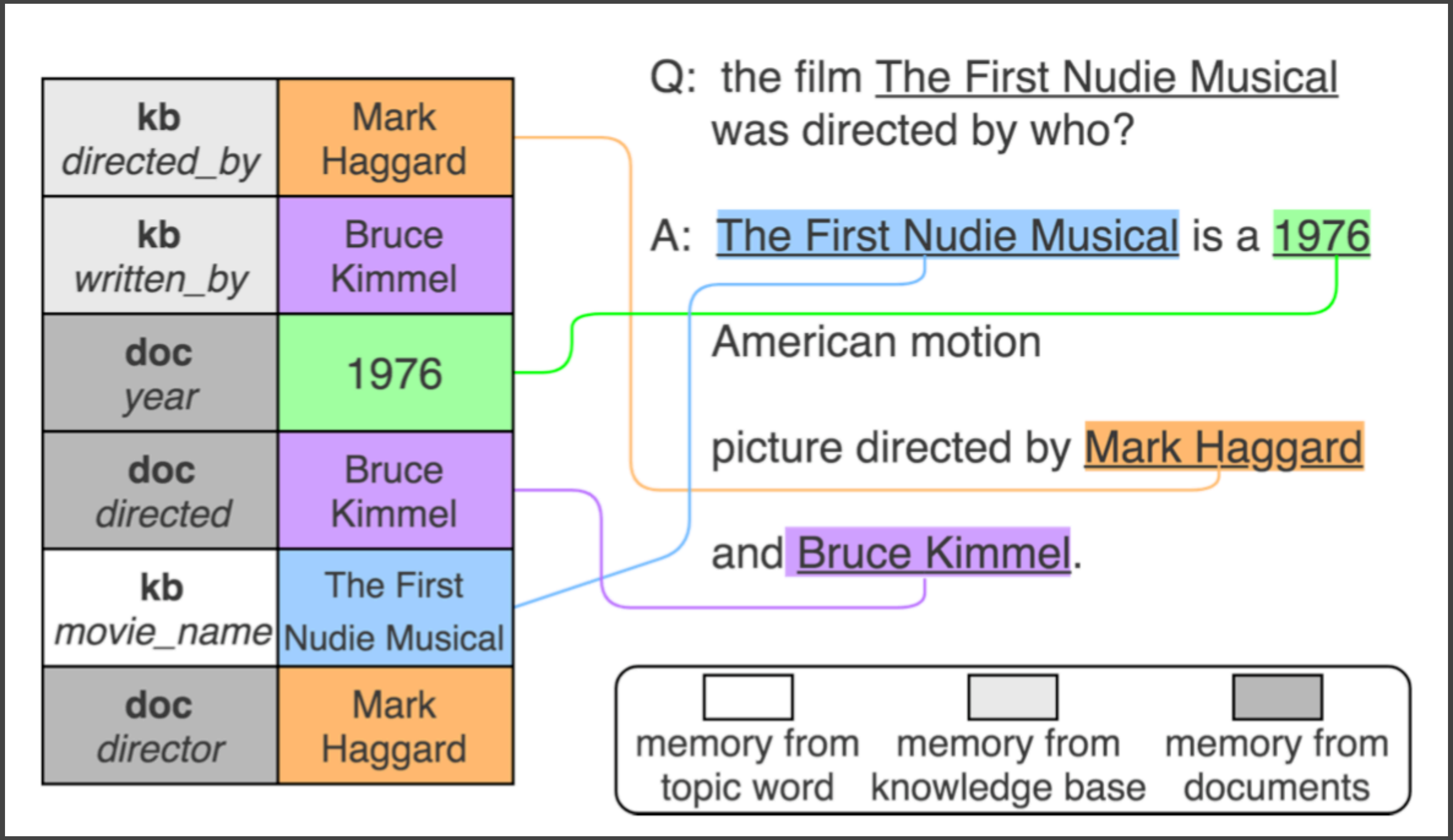
MOTIVATION

- Sentence generation and metrics:
  - Close ended: machine translation& summarization
    - Well-defined, meaningful metrics
  - Open ended: structural data to text; chit-chat; visual story telling
    - No perfect metrics, multi-dimensional evaluation
    - Require world knowledge

- World knowledge: structural data, different sources, heterogeneous natural
  - Fully structural: knowledge graph
  - Semi structural: web table; OpenIE
  - Unstructured: need to organize, data cleaning

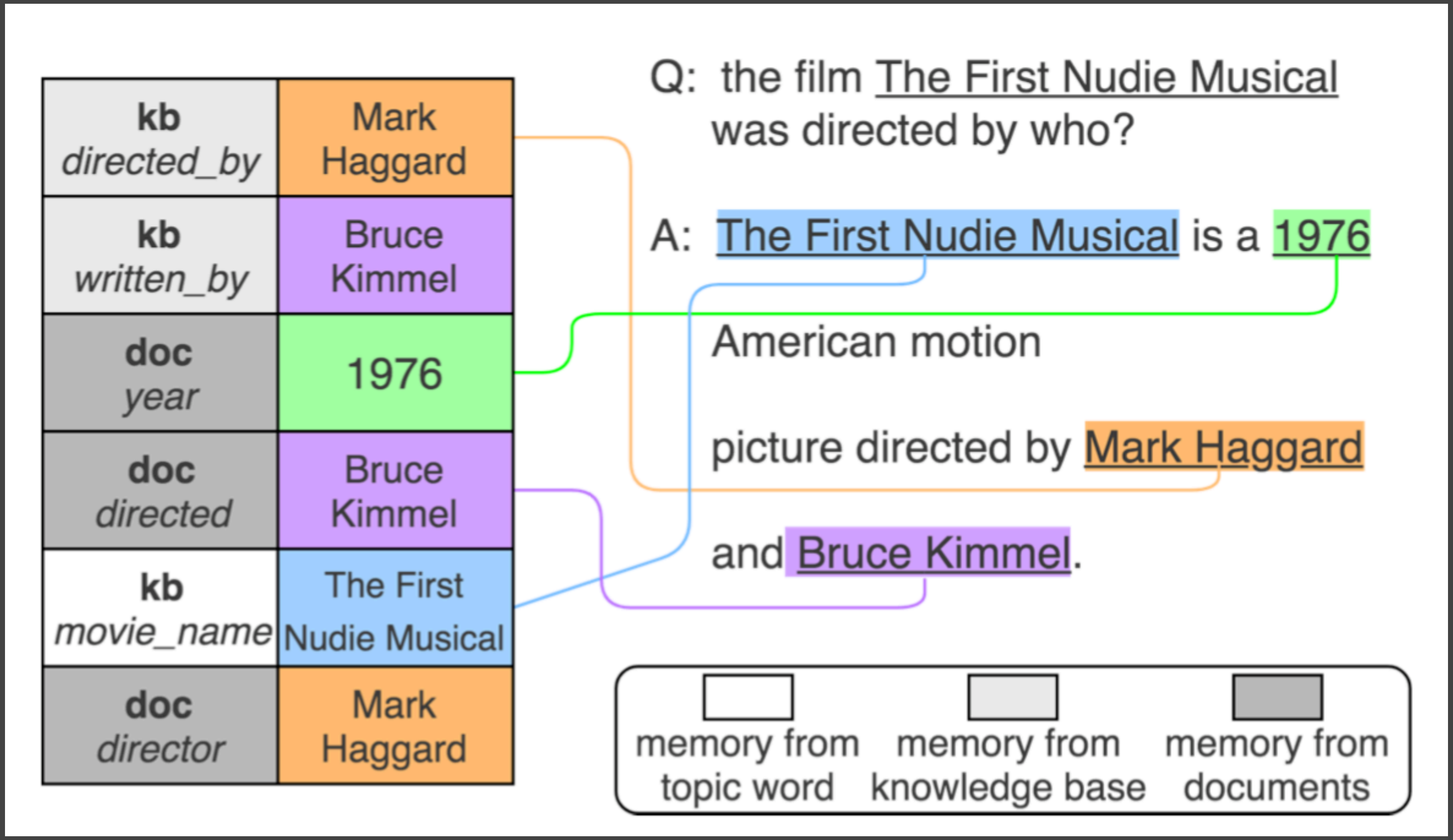
- Generation application:
  - Dialog response — any simplification?
  - First simplification: answer sentence generation from a table — this talk
  - Further simplification: table to text — this talk

THE ANSWER SENTENCE GENERATION TASK



- Given a key-value table, a question
- Find the answer
- Compose it into a sentence
- Similar to single-round dialog

THE ANSWER SENTENCE GENERATION TASK



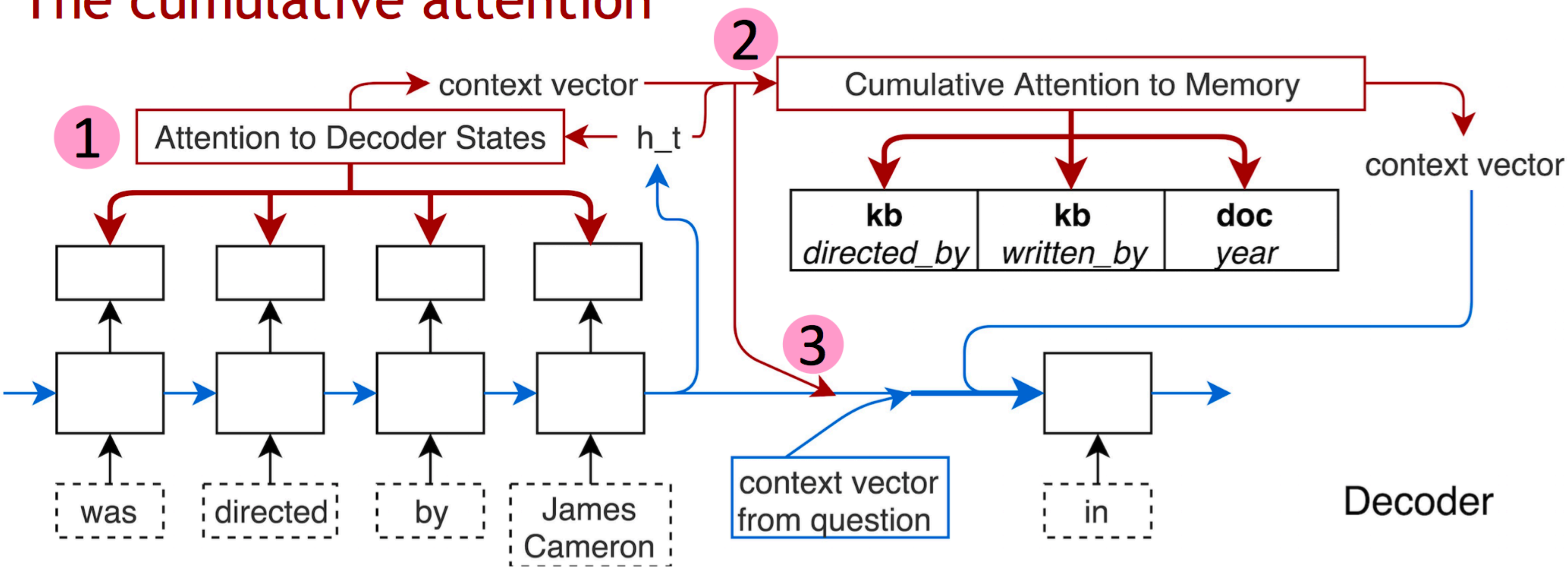
- The Goal
  - Primary: answer coverage, sentence quality
  - Additionally: incorporate background information - allows the conversation to continue, more human-like
- First intuition: key-value memory + seq2seq-Attn + pointer
- Problem: redundancy v.s. informativeness

THE REDUNDANCY - INFORMATIVENESS TRADEOFF

- Redundancy: the generated sentence will repeat certain words
- Source of redundancy:
  - From the data - different sources
  - From the decoder
- Informativeness: the sentence should discuss the subject
  - Background information
- The tradeoff:
  - Think about: when generating an additional word, either a new word, or an existing word
  - The longer the sentence is, the more information it can provide, the more redundancy there might be.
  - Intrinsic of human language: more informativeness, more redundancy
  - Will discuss more about the multi-aspects nature of human language
- The Goal
  - Primary: Answer coverage, Sentence quality
  - Secondary: less redundancy, more information — on the decoder
- What kind of structural inductive bias we want to inject into this decoder?

MODELING

The cumulative attention



- The inductive bias for the decoder: how do I know I already said this?
- Generation mode: decoder self attention - help avoid repetition
- Copy mode: cumulative attention - help content selection, i.e. select new words instead of mentioned
- Comparison with the transformer model, the shared inductive bias
  - Decoder self attention
  - Cumulative Attention

AUTOMATIC EVALUATION

Model	Redundancy	C_{single}	C_{part}	$C_{perfect}$	Enrich
GenQA	0.1109	91.25%	69.19%	38.92%	0.1535
HS-GenQA	0.1218	94.10%	76.47%	50.10%	0.1951
GenQA-AttnHist	0.1280	95.99%	73.44%	44.94%	0.1903
CheckList	0.1176	93.80%	76.32%	50.04%	0.1963
HS-AttnHist	0.1295	97.17%	**77.90%**	**51.55%**	**0.1996**
HS-CumuAttn	**0.0983**	**98.15%**	**77.28%**	50.79%	0.1665

Table 3: Results on the `WikiMovies-Synthetic` dataset

Model	BLEU	Redundancy	C_{part}	$C_{perfect}$	Enrich
GenQA	42.50	0.2603	62.80%	18.24%	0.5903
CheckList	43.69	0.2744	63.42%	18.23%	0.6094
HS-CumuAttn	**44.97**	**0.2385**	**64.06%**	**19.09%**	**0.6218**

Table 4: Results on the `WikiMovies-Wikipedia` dataset

- Redundancy = % repeated words
- Informativeness = Enrichment = % related facts
- Redundancy - informativeness tradeoff in baseline models
- Performance gain from decoder's modeling power in all aspects

GENERATION SAMPLES

Question 1	who starred in Cemetery Man ?			
Memory	0 *ans_actor*	Rupert Everett	1 *ans_actor*	Anna Falchi
	2 *starred_actors*	Rupert Everett	3 *starred_actors*	Anna Falchi
	4 *movie*	Cemetery Man		
Answer	The film stars Rupert Everett$_0$, _UNK , and Anna Falchi$_1$.			
Question 2	who was Dying Breed written by ?			
Memory	0 *ans_release_year*	2008	1 *ans_writer*	Jody Dwyer
	2 *ans_actor*	Nathan Phillips	3 *ans_writer*	Leigh Whannell
	4 *written_by*	Jody Dwyer	5 *movie*	Dying Breed
Answer	Dying Breed$_5$ is a 2008$_0$ Australian horror film that was directed by Jody Dwyer$_1$ and stars Leigh Whannell$_3$ and Nathan Phillips$_2$.			
Question 3	who is the director that directed Livid ?			
Memory	0 *ans_director*	Julien Maury	1 *directed_by*	Alexandre Bustillo
	2 *ans_release_year*	2011	3 *ans_director*	Alexandre Bustillo
	4 *movie*	Livid	5 *directed_by*	Julien Maury
	6 *ans_language*	French		
Answer	Livid$_4$ ( ) is a 2011$_2$ French$_6$ supernatural horror film directed and written by Julien Maury$_0$ and Alexandre Bustillo$_3$.			
Question 4	Drag Me to Hell , when was it released?			
Memory	0 *ans_director*	Sam Raimi	1 *ans_wiki*	Scream
	2 *release_year*	2009	3 *ans_genre*	Horror
	4 *ans_release_year*	2009	5 *movie*	Drag Me to Hell
Answer	Scream$_1$ is a 2009$_4$ film			
Question 5	the movie Lights in the Dusk starred who ?			
Memory	0 *starred_actors*	Janne Hyytiäinen	1 *ans_language*	Finnish
	2 *starred_actors*	Maria Järvenhelmi	3 *ans_actor*	Janne Hyytiäinen
	4 *starred_actors*	Ilkka Koivula	5 *movie*	Lights in the Dusk
	6 *ans_actor*	Ilkka Koivula	7 *ans_release_year*	2006
	8 *ans_actor*	Maria Järvenhelmi		
Answer	Lights in the Dusk$_5$ (,) is a 2006$_7$ Finnish$_1$ drama film starring Janne Hyytiäinen$_3$, Ilkka Koivula$_6$ and Maria Järvenhelmi$_8$.			

GENERATION SAMPLE PROBLEMS

- Not template based, but learns template
  - Very common in NLG
  - Partially because of MLE& greedy/ beam search sampling

- Lack fact check (Q2)

- Dependency agreement (Q2, Q3), note: this is short-term dependency

- Too dull (Q4)

- -> New evaluation metrics? Some works do

- Again, the central role of the decoder
  - The decoder language model
  - The decoder pointer model
  - Sampling strategy, more considerations on this

ON DECODING: SAMPLING STRATEGIES

- If we want
  - A. Prevent repetition: rejection sampling
    - The first approach to cut repetition
    - Often combined with other techniques
  - B. Most probable: greedy decoding, beam Search decoding
    - Good for close ended tasks: MT, Summarization
      - restricted search space
    - Not for open ended tasks:
      - e.g. simple, short — larger prob. — safe — less informative
  - C. More diversity: top-k/ top-p sampling
    - Better for open ended generation
    - Increase diversity, decrease certainty — another tradeoff pair
    - But, do you want the generation do random walk over large search space, or do you want it to walk within the restricted target space?
  - D. Must contain answer words: constraint decoding
    - Grid beam search
    - Metropolis-hastings sampling
- All depend on a more powerful decoder - pre-training nowadays

THE EFFECTIVENESS OF PRE-TRAINING

	BERT Init.	Random Init.
Full Set	37.72	37.83
1/3 Set	34.74	35.96
1/10 Set	28.54	31.73

Table 2: The performance of the models on different size of training data with different initialization.

Pre-training Method	BLEU (1/10 Set)	BLEU (1K)
Left-to-right LM	31.72	17.66
Masked LM	29.83	13.60
Self Pre-train	30.60	2.53
No Pre-train	27.77	2.42

Table 3: The performance of our model with different size of training data and pre-training methods.

- Table to text generation, pre-trained transformer decoder
- Pre-training on different domain: Random > BERT
  - The domain gap and the objective gap
- In domain pre-training: Left to right > Masked LM > no pre-training; the objective gap
- Effectiveness on few shot learning
- Side node - BERT generation: Gibbs sampling, non-autoregressive sampling

II. EVALUATION

START FROM BLEU

- Bask to the task:
  - Close ended, quality = exactly describe the subject as the references do
  - Open ended, quality = describe anything about the subject fluently, no exact restriction
- BLEU and other reference matching based evaluation:
  - Meaningful only when you have good reference
  - Close ended: restricted reference space
  - Open ended: exponential reference space
- Extend the reference space:
  - More references for test set: hand written, IR
  - Match any sentence from the training/test corpus
  - Quality aspect: fluency ↑ exact matching ↓
  - LM perplexity — match any, soft version; interpretability
  - More quality aspects

THE MULTI-DIMENSIONAL NATURE OF HUMAN LANGUAGE

- What do we want from a NLG system?
  - Overall quality
  - Naturalness/ fluency
  - Diversity/ mode coverage
  - Redundancy
  - Informativeness/ information coverage
  - Fact check
  - Dependency agreement
  - Word choice
    - Use/ not use certain word
    - Prevent offensive language
  - Tradeoffs

ON NLG METRICS

- Targets v.s. qualifiers
- Overall quality
  - BLEU: still good for tasks less open-ended
  - Hard to produce 3-gram and 4-gram matching, 2-gram most sensible
  - Not correlated with human preference: acceptable, not preferable
- Naturalness/ fluency:
  - Perplexity, corpus level, qualifier
- Diversity/ mode coverage:
  - reverse perplexity, corpus level qualifier
- Redundancy:
  - Repeated word count, qualifier
- Informativeness/ information coverage:
  - Precision and recall, target measure
- Fact check:
  - How to do this?? target measure
- Dependency agreement:
  - Dependency parsing? target measure
- Word choice: use/ not use certain word:
  - Matching
  - But more importantly, how to rectify?

ON NLG METRICS, TRADEOFFS

- Informativeness v.s. redundancy
- Informativeness v.s. dependency
- Sentence length v.s. naturalness
- Sentence length v.s. dependency
- Easy to converge to short, safe sentences
- Challenge: longer sentences, complex dependency (either short term or long term), external knowledge (common sense)

ON NLG METRICS, THE ACCEPTANCE REGION

- What should we do with all these aspects? Acceptance region
- Goal-oriented:
  - Pick the target metrics, set up your lowest acceptance bar
  - Determine the qualifiers, set up your lowest acceptance bar
  - Accept all models satisfying the lower bound
  - Tune the target metrics tradeoffs
  - Leveraging more data/ better inductive bias that simultaneously increase all tradeoff factors

III. CONCLUSION & FINAL REMARKS

CONCLUSION & FINAL REMARKS

- The multi-dimensional nature of human language, no single perfect metrics
- Find the primary goal, set the target metrics, accept all within the lower bound
- Strike a balance between the tradeoff targets
- Better inductive bias on the model to simultaneously increase all targets