Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks

Tasnim Mohiuddin*, Prathyusha Jwalapuram*, Xiang Lin* and Shafiq Joty*

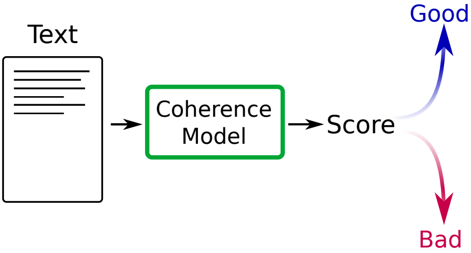
School of Computer Science and Engineering
Nanyang Technological University

Outline

1. Introduction
   - Motivation
2. Coherence Models
3. Experiments
   - Synthetic Tasks
   - Coherence Evaluation Tasks
   - Task-specific Training for Dialog
4. Task-specific Training for MT
5. Conclusions

Coherence Models

- Build models that can distinguish a coherent text from incoherent ones
- Key problem in discourse analysis



Motivation

- Coherence models have been proposed with the aim of applying them in text generation and ranking - *e.g.*,
  - Summarization
  - Machine translation
  - Essay scoring
  - Dialog systems
  - ....
- But most work on coherence modeling ignores downstream applications

Motivation

- Coherence models are commonly evaluated on **synthetic** discrimination tasks
  - Testing on *readability assessment* or *essay scoring* is far less common

Motivation

- Coherence models are commonly evaluated on **synthetic** discrimination tasks
  - Testing on *readability assessment* or *essay scoring* is far less common
- There have been claims of
  - human parity in MT [Hassan et al., 2018]
  - fluency in summarization [Celikyilmaz et al., 2018]
  - fluency in context-consistent response generation [Zhang et al., 2020]

⇒ Coherence modeling of machine generated text is now more crucial than ever [Läubli et al., 2018]

Motivation

- But unclear if existing coherence models are capable of this task
    - Their performance on **downstream applications** is rarely studied

Motivation

- But unclear if existing coherence models are capable of this task
  - Their performance on **downstream applications** is rarely studied
- Our goal is to bridge this gap
  ⇒ Assess coherence models on standard discrimination tasks and compare against performance on downstream use-cases

Coherence Models

- Both traditional and neural models included:

EGrid Entity-grid model [Barzilay and Lapata, 2005]

NeuralEGrid Neural entity-grid model [Nguyen and Joty, 2017]

LexNeuEGrid Neural entity-grid with lexicalised entity transitions [Mohiuddin et al., 2018]

TransModel A neural local coherence model that considers only adjoining sentences [Xu et al., 2019]

UnifiedModel Unified model that captures syntax, discourse relations, entity attention and global topic structures [Moon et al., 2019]

Training

	Sections	# Doc.	# Pairs
Train	00-13	1,378	26,422
Test	14-24	1,053	20,411

Table: Statistics of the WSJ news dataset used for the **Global discrimination** task.

- Standard pairwise training setup: coherent vs. incoherent document ranking
- Incoherent document is a random permutation of the sentences in the coherent document (from the WSJ corpus)
- 20 random permutations for training and testing, with 10% held out for development

Synthetic Tasks

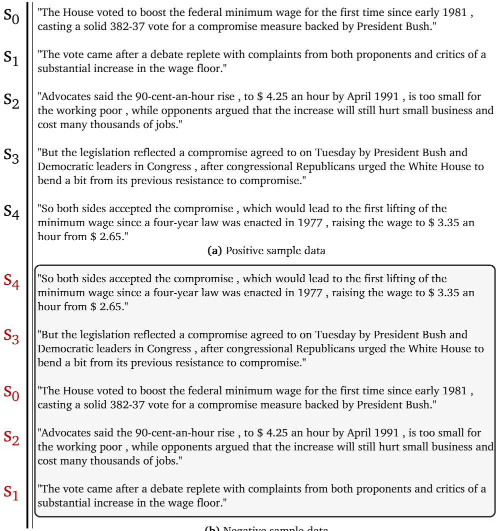
Global Discrimination Tasks

Two synthetic tasks that test coherence at a global level

Standard: sentences randomly permuted to create incoherent document

Inverse: sentence order is reversed to create incoherent document

Global Discrimination Task



S_0 "The House voted to boost the federal minimum wage for the first time since early 1981 , casting a solid 382-37 vote for a compromise measure backed by President Bush."

S_1 "The vote came after a debate replete with complaints from both proponents and critics of a substantial increase in the wage floor."

S_2 "Advocates said the 90-cent-an-hour rise , to $ 4.25 an hour by April 1991 , is too small for the working poor , while opponents argued that the increase will still hurt small business and cost many thousands of jobs."

S_3 "But the legislation reflected a compromise agreed to on Tuesday by President Bush and Democratic leaders in Congress , after congressional Republicans urged the White House to bend a bit from its previous resistance to compromise."

S_4 "So both sides accepted the compromise , which would lead to the first lifting of the minimum wage since a four-year law was enacted in 1977 , raising the wage to $ 3.35 an hour from $ 2.65."

(a) Positive sample data

S_4 "So both sides accepted the compromise , which would lead to the first lifting of the minimum wage since a four-year law was enacted in 1977 , raising the wage to $ 3.35 an hour from $ 2.65."

S_3 "But the legislation reflected a compromise agreed to on Tuesday by President Bush and Democratic leaders in Congress , after congressional Republicans urged the White House to bend a bit from its previous resistance to compromise."

S_0 "The House voted to boost the federal minimum wage for the first time since early 1981 , casting a solid 382-37 vote for a compromise measure backed by President Bush."

S_2 "Advocates said the 90-cent-an-hour rise , to $ 4.25 an hour by April 1991 , is too small for the working poor , while opponents argued that the increase will still hurt small business and cost many thousands of jobs."

S_1 "The vote came after a debate replete with complaints from both proponents and critics of a substantial increase in the wage floor."

(b) Negative sample data

Global Discrimination Task Results

Model	**Emb.**	**Standard**	**Inverse**
EGrid	–	81.60	75.78
NeuralEGrid	–	84.36	83.94
LexNeuEGrid	word2vec	88.51	88.13
TransModel	Avg. Glove	91.77	**99.62**
UnifiedModel	ELMo	**93.19**	96.78

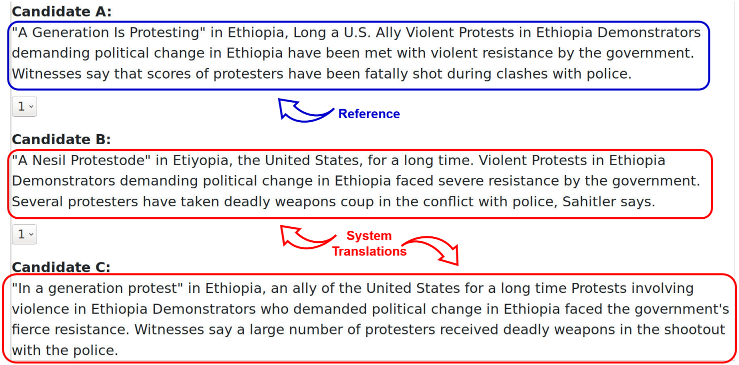
- UnifiedModel best at the standard task
- TransModel best at the inverse task

Downstream Tasks

Machine Translation Coherence

- At a document level, reference translations shown to be more coherent than system translations [Smith et al., 2016]
  ⇒ Test if the models score the reference higher
- Conduct a user study to get pairwise coherence ranks of translations from different systems
  ⇒ Check agreement with humans with respect to the ranks

Machine Translation Coherence



Candidate A:

"A Generation Is Protesting" in Ethiopia, Long a U.S. Ally Violent Protests in Ethiopia Demonstrators demanding political change in Ethiopia have been met with violent resistance by the government. Witnesses say that scores of protesters have been fatally shot during clashes with police.

1

Reference

Candidate B:

"A Nesil Protestode" in Etiyopia, the United States, for a long time. Violent Protests in Ethiopia Demonstrators demanding political change in Ethiopia faced severe resistance by the government. Several protesters have taken deadly weapons coup in the conflict with police, Sahitler says.

1

System Translations

Candidate C:

"In a generation protest" in Ethiopia, an ally of the United States for a long time Protests involving violence in Ethiopia Demonstrators who demanded political change in Ethiopia faced the government's fierce resistance. Witnesses say a large number of protesters received deadly weapons in the shootout with the police.

Machine Translation Coherence Results

Model	Acc. (%)	AC1 Agr.
EGrid	51.75	**0.80**
NeuralEGrid	**54.75**	0.77
LexNeuEGrid	49.34	0.76
TransModel	48.67	0.77
UnifiedModel	43.36	0.78

- Overall, EGrid model performing better than others
- Models like TransModel and UnifiedModel with strong global discrimination performance do not perform well

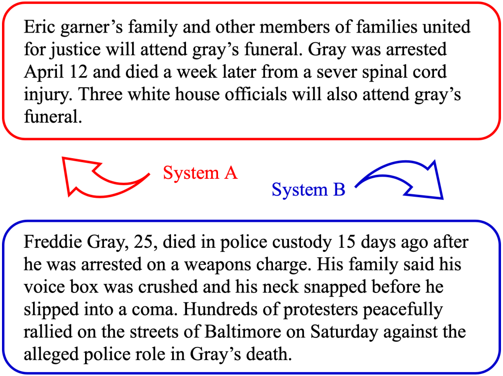
Downstream Tasks

Summarization

Abstractive: Conduct user study to rank the summaries from different abstractive summarization systems in terms of coherence
$\Rightarrow$ Check agreements between coherence model ranks and human ranks

Extractive: Use human coherence ratings for extractive summaries from Document Understanding Conference (2003)
$\Rightarrow$ Check agreements between coherence model ranks and human ranks converted from ratings

Summary Coherence



Eric garner's family and other members of families united for justice will attend gray's funeral. Gray was arrested April 12 and died a week later from a sever spinal cord injury. Three white house officials will also attend gray's funeral.

System A

System B

Freddie Gray, 25, died in police custody 15 days ago after he was arrested on a weapons charge. His family said his voice box was crushed and his neck snapped before he slipped into a coma. Hundreds of protesters peacefully rallied on the streets of Baltimore on Saturday against the alleged police role in Gray's death.

Figure: Summaries from two abstractive summarization systems

Summarization Results

Models	**Abs. Agr.**	**Ext. Agr.**
EGrid	**0.71**	0.52
NeuralEGrid	0.68	**0.70**
LexNeuEGrid	**0.71**	0.57
TransModel	0.55	0.38
UnifiedModel	0.68	0.35

- Similar pattern emerges; models which had high performance in synthetic tasks perform poorly

Task-specific Training for Dialog

- Previous task setup has a training and testing mismatch
  ⇒ Re-train and test on task-specific setup for next utterance ranking for dialog

Task-specific Training for Dialog

- Previous task setup has a training and testing mismatch
  ⇒ Re-train and test on task-specific setup for next utterance ranking for dialog
- Why next utterance ranking?
  - Non-synthetic task with task-specific training data
  - Similar to the synthetic task of insertion [Elsner and Charniak, 2011a]

Task-specific Training for Dialog

Next Utterance Ranking

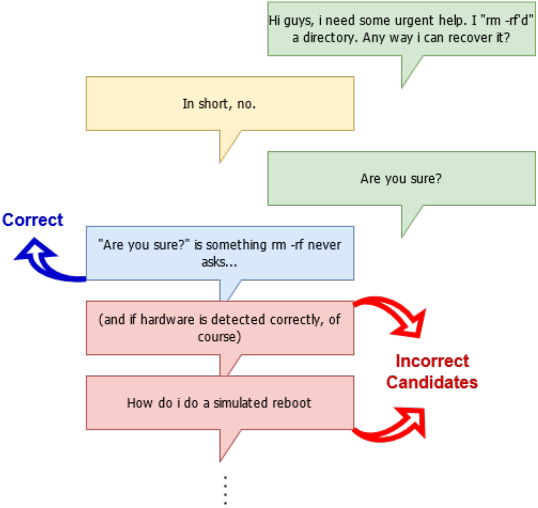
- Use the data from Noetic End-to-End Response Selection Challenge II from DSTC 8
- Each sample has conversational context $U = (u_1, \ldots, u_{|U|})$ and potential candidate utterances $C = \{c_1, \ldots, c_{|C|}\}$
    - Model needs to select correct next utterance $r \in C$

Task-specific Training for Dialog

Next Utterance Ranking

- A good coherence model should rank coherent dialog $P = (u_1, \ldots, u_{|U|}, r)$ higher than incoherent one $N = (u_1, \ldots, u_{|U|}, c_j)$
- Train model to score coherent (P) higher than incoherent samples (N)
- Candidate pool has 100 utterances
  $\Rightarrow$ Report accuracy and DSTC8 metrics of Recall@k and Mean Reciprocal Rank

Next Utterance Ranking



Hi guys, i need some urgent help. I "rm -rf'd" a directory. Any way i can recover it?

In short, no.

Are you sure?

"Are you sure?" is something rm -rf never asks...

(and if hardware is detected correctly, of course)

How do i do a simulated reboot

⋮

Next Utterance Ranking Results

	R@1	R@5	R@10	MRR	Acc.
Advising dataset					
Official Evaluation					
Best	0.564	0.81	0.88	0.68	X
Median	0.14	0.37	0.51	0.26	X
Worst	0.01	0.05	0.09	0.05	X
Coherence Model					
EGrid	0.004	0.03	0.07	0.04	47.16
NeuralEGrid	0.057	0.17	0.23	0.13	56.15
LexNeuEGrid	0.046	0.17	0.26	0.13	57.66
TransModel	0.067	0.20	0.30	0.14	**66.62**
UnifiedModel	0.022	0.06	0.19	0.11	54.33

	R@1	R@5	R@10	MRR	Acc.
Ubuntu dataset					
Official Evaluation					
Best	0.761	0.96	0.98	0.85	X
Median	0.55	0.86	0.93	0.68	X
Worst	0.24	0.38	0.46	0.32	X
Coherence Model					
EGrid	0.007	0.05	0.09	0.05	47.48
NeuralEGrid	0.18	0.39	0.49	0.29	73.18
LexNeuEGrid	0.15	0.31	0.39	0.24	74.39
TransModel	0.045	0.14	0.26	0.12	70.94
UnifiedModel	0.035	0.17	0.33	0.13	**74.49**

- Pairwise accuracies are higher than random baseline, but skewed due to large number of negative candidates
- Official evaluation metrics of Recall@k and Mean Reciprocal Rank show very poor performance despite task-specific training

Task-specific Training for MT

Machine Translation Coherence

- Investigate whether a change in the training setup might help models learn more useful task-specific features
- Train coherence models with reference text as positive and system translations as negative documents using WMT data $\Rightarrow$ Report accuracy and agreement on the same test data as previous experiment

Task-specific Results for MT

Model	Acc. (%)	AC1 Agr.
EGRID	48.74	0.797
NEURALEGRID	52.58	0.76
LEXNEUEGRID	56.84	0.795
TRANSMODEL	57.65	0.751
UNIFIEDMODEL	**77.35**	**0.828**

- Performance improves across models
    - UNIFIEDMODEL jumps up from 43.36% to 77.35%, and high agreement of 0.83

Discussion

- Models trained on permuted sentences **may not be learning features useful for downstream applications**
  - Models based on synthetic tasks may be overfitting on the these tasks

⇒ May fail to find coherence issues more subtle than permuted text

Discussion

- Models trained on permuted sentences **may not be learning features useful for downstream applications**
    - Models based on synthetic tasks may be overfitting on the these tasks

  ⇒ May fail to find coherence issues more subtle than permuted text

- Only considering incoherence from permuted documents may be a **poor approximation of real-world coherence** problems
    - *e.g.*, MT output often produced sentence by sentence - unlikely to be out of order

Discussion

- Models fail on next-utterance ranking despite task-specific re-training
  - Best performance on similar synthetic task of insertion also barely reaches 26% [Elsner and Charniak, 2011a, Nguyen and Joty, 2017]
- **Training procedures may not be helpful for learning generic features** that apply in a harder setup
  ⇒ A change of training setup may be needed for actual usage

Conclusions

- Similar results from Elsner and Charniak [2011b] showed **lack of generalizability** of coherence models to the task of chat disentanglement
- **Standard training paradigm may need reform** to build more generalizable models
- **Standard evaluation needs reform** to be more indicative of real-world performance

Thank you!

SCAN ME