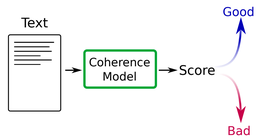
Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks

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

♠Nanyang Technological University ♣Salesforce Research Asia *Equal Contribution

Coherence Analysis

- Coherence models - distinguish a coherent text from incoherent ones
- Proposed with the aim of applying them in text generation and ranking - *e.g.*, machine translation, essay scoring, dialog systems, summarization, ...
- But most work on coherence modeling **ignores** downstream applications



Motivation

- Common evaluation: synthetic discrimination tasks
- Unclear if existing models are capable of modeling coherence of generated text from downstream applications

Goal: Assess coherence models on standard discrimination tasks and compare against performance on downstream use-cases.

Standard Training and Synthetic Tasks

Training:

- WSJ news dataset
- Standard pairwise training setup: coherent vs. incoherent document ranking
- 20 random permutations (incoherent documents) for training and testing

Global Discrimination Tasks

1. **Standard**: Incoherent document is a random permutation of sentences
2. **Inverse**: Incoherent document has a fully reversed sentence order

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

Model	Standard	Inverse
EGRID	81.60	75.78
NEURALEGRID	84.36	83.94
LEXNEUEGRID	88.51	88.13
TRANSMODEL	91.77	**99.62**
UNIFIEDMODEL	**93.19**	96.78

Table 1. Results: Accuracies of the coherence models in the **Global Discrimination** task.

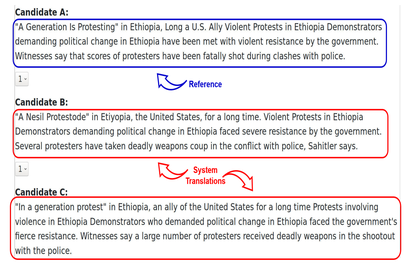
Downstream Tasks

Machine Translation Coherence

1. Reference translations more coherent than system translations at document level. **Test if models score reference higher**
2. User study: get pairwise coherence ranks of translations from different systems. **Check model agreement with humans**

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

Table 2. Machine Translation Results



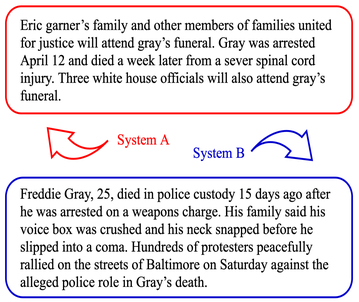
⇒ Models like TRANSMODEL and UNIFIEDMODEL with strong global discrimination performance do not perform well

Summarization

1. **Abstractive:** User study for coherence ranking of summaries from abstractive summarization systems. **Check model agreements with humans**
2. **Extractive:** Human coherence ratings for extractive summaries from Document Understanding Conference. **Check model agreements with human ranks converted from ratings**

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

Table 3. Text Summarization Results



⇒ Similar pattern; models with high performance in synthetic tasks perform poorly

Task-specific Training

- Investigate whether a change in the training setup might help models learn more useful task-specific features

Machine Translation Setup

- Train coherence models with reference text as positive and system translations as negative documents using WMT data
- Report accuracy and agreement on the same test data as previous experiment

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 improves from 43.36% to 77.35%, high agreement of 0.83

Conclusions

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, failing 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

Conclusions:

- **Standard training paradigm may need reform** to build more generalizable models
- **Standard evaluation needs reform** to be more indicative of real-world performance

Resources

Please scan the QR code for more information and resources:

SCAN ME