A Cross-Domain Transferable Neural Coherence Model

Peng Xu[1], Hamidreza Saghir[1], Jin Sung Kang[1], Teng Long[1], Avishek Joey Bose*[,1,2], Yanshuai Cao[1], Jackie Chi Kit Cheung[1,2,3]

[1]Borealis AI, [2]McGill University, [3]Canada CIFAR Chair, Mila

Motivations

- Coherence is a discourse property that is concerned with the logical and semantic organization of a passage, such that the overall meaning of the passage is expressed fluidly and clearly.
- It is an important quality measure for text generated by humans or machines
- Modelling coherence can benefit many applications, including summarization, question answering, essay scoring and text generation.

Key Take-aways

- Much of global coherence can be decomposed into a series of local decisions.
- Simpler local model enables better cross-domain generalization.
- A new dataset with increasingly harder cross-domain evaluation protocols.

Discriminative vs. Generative

Corpus as $\mathcal{C} = \{d_k\}_{k=1}^{N}$, with N documents; each document d_k with sentences $\{s_i\}$.

- **Discriminative models** try to assign a higher coherence score to d_i than a randomly permuted version d_i^- of the same document. Let S_i denote the set of all possible permutations of d_i. The learning minimizes L with respect to θ:

$$\min_{\theta} \sum_{d_i \in \mathcal{C}} \sum_{d_i^- \in S_i} L(d_i, d_i^-; \theta). \tag{1}$$

Disadvantages of discriminative models:

1. Prune to overfitting to dataset and domain, especially for big neural nets.
2. $n!$ possible sentence orderings for a document with n sentences, thus the sampled negative instances can only cover a tiny proportion of this space.

- **Generative models** are based on the idea that in a coherent document, subsequent sentences should be predictable given their preceding sentences, and vice versa. Trained to maximize log-likelihood:

$$\max_{\theta} \sum_{d \in \mathcal{C}} \sum_{s \in d} \log p(s|c_s; \theta), \tag{2}$$

where c_s is the context of the sentence s.

Hidden assumptions behind this maximum likelihood approach:

1. Conditional log likelihood is a good proxy for coherence.
2. The long-range dependencies can be well captured by the generative model.

Potential problems for generative models:

1. A coherent sentence does not need to have high conditional log likelihood. Fluency, grammaticality, and word frequency in s all affect log likelihood.
2. Learning long-range dependencies in neural nets still an active research area.

*Work done while the author was an intern at Borealis AI.

Local Discriminative Model

Our operating assumption is that the global coherence of a document can be well approximated by the average of coherence scores between consecutive pairs of sentences, in light of theories like Centering Theory [1] and Rhetorical Structure Theory [2].

- **Training objectives:** Formally, our discriminative model $f_\theta(.,.)$ takes a sentence pair and returns a score. Then our training objective is:

$$\mathcal{L}(\theta) = \sum_{d \in \mathcal{C}} \sum_{s_i \in d} \mathop{\mathbb{E}}_{p(s'|s_i)} \left[L(f_\theta(s_i, s_{i+1}), f_\theta(s_i, s')) \right] \tag{3}$$

where $\mathbb{E}_{p(s'|s_i)}$ denotes expectation with respect to negative sampling distribution p which could be conditioned on s_i; and $L(.,.)$ is a loss function that takes two scores.

- **Negative samples:** $(n-1)*(n-2)/2$ negative pairs instead of $n!$. The quadratic number of negatives provides a rich enough learning signal, while not too prohibitively large to be effectively covered by a sampling procedure.
- **Pre-trained generative model as the sentence encoder:** Since generative models can often be turned into sentence encoder, generative coherence model can be leveraged by our model to benefit from the advantages of both generative and discriminative training.

Datasets and Protocols

- Closed domain: the Wall Street Journal (WSJ) portion of Penn Treebank.
- Open domain: a new dataset based on Wikipedia and design three cross-domain evaluation protocols with increasing levels of difficulty.

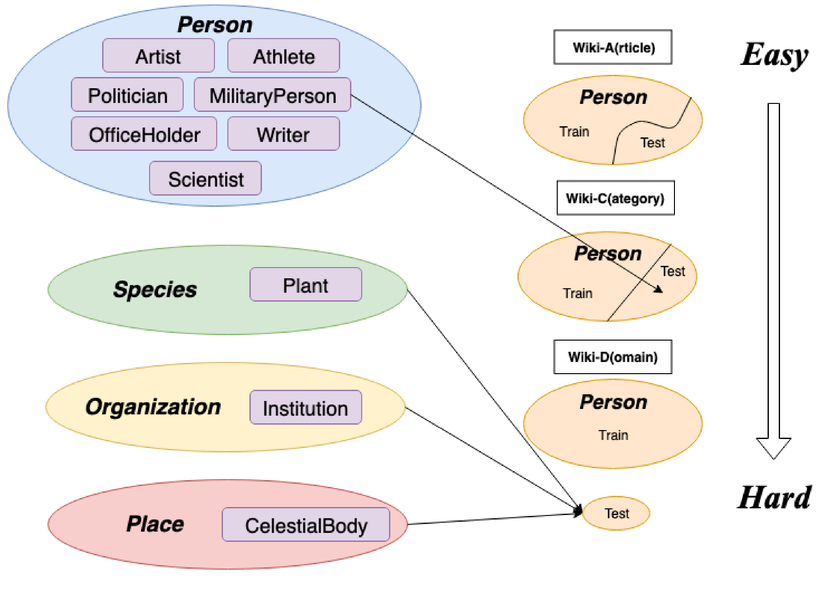

Fig. 1: Overview of the new dataset and three evaluation protocols based on it.

Results

1. **LCD-G**: use averaged GloVe vectors as the sentence representation;
2. **LCD-I**: use pre-trained InferSent as the sentence encoder;
3. **LCD-L**: apply max-pooling on the hidden state of the language model to get the sentence representation.

	Discr.	Ins.
Clique-Discr. (3)	70.91	11.53
Clique-Discr. (7)	70.30	5.01
Grid-CNN	85.57 (85.13)	23.12
Extended Grid-CNN	88.69 (87.51)	25.95
Seq2Seq	86.95	27.28
Vae-Seq2Seq	87.01	26.73
LM	86.50	26.33
LCD-G	92.51	30.30
LCD-I	94.54	32.34
LCD-L	**95.49**	**33.79**

Tab. 1: Accuracy of **Discr**imination and **Ins**ertion tasks evaluated on WSJ.

Clique-Discr. (3)	76.17
Clique-Discr. (7)	73.86
Seq2Seq	86.63
Vae-Seq2Seq	82.40
LM	93.83
LCD-G	91.32
LCD-I	94.01
LCD-L	**96.01**

Tab. 2: Accuracy of discrimination task under Wiki-A

Model	Artist	Athlete	Politician	Writer	MilitaryPerson	OfficeHolder	Scientist	Average
Clique-Discr. (3)	73.01	68.90	73.82	73.28	72.86	73.74	74.56	72.88
Clique-Discr. (7)	71.26	66.56	73.72	72.01	72.67	72.62	71.86	71.53
Seq2Seq	82.72	73.45	84.88	85.99	81.40	83.25	85.27	82.42
Vae-Seq2Seq	82.58	74.14	84.70	84.94	81.07	82.66	85.09	82.17
LM	88.18	78.79	88.95	90.68	87.02	87.35	91.92	87.56
LCD-G	89.66	86.06	90.98	90.26	89.23	89.86	90.64	89.53
LCD-I	92.14	89.03	93.23	92.07	91.63	92.39	93.03	91.93
LCD-L	**93.54**	**90.13**	**94.04**	**93.68**	**93.20**	**93.01**	**94.81**	**93.20**

Tab. 3: Accuracy of discrimination task under Wiki-C setting.

Model	Plant	Institution	CelestialBody	WSJ	Average
Clique-Discr. (3)	66.14	66.51	60.38	64.71	64.44
Clique-Discr. (7)	65.47	69.14	61.44	66.66	65.68
Seq2Seq	82.58	80.86	69.44	74.62	76.88
Vae-Seq2Seq	81.90	78.00	69.10	73.27	75.57
LM	81.88	83.82	74.78	79.78	80.07
LCD-G	86.57	86.10	79.16	82.51	83.59
LCD-I	**89.07**	88.58	80.41	**83.27**	85.33
LCD-L	88.83	**89.46**	**81.31**	82.23	**85.48**

Tab. 4: Accuracy of discrimination task under Wiki-D setting.

- Diminishing return with greater coverage of samples beyond certain points.
- Similar comparisons for sentence re-ordering tasks.
- Correlated with Wikipedia's "rewrite" flags as proxy of human judgement.

Misc

[1] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*.

[2] Sandra A Thompson and William C Mann. 1987. Rhetorical structure theory. *IPRA Papers in Pragmatics*.

arXiv