



Problem statement

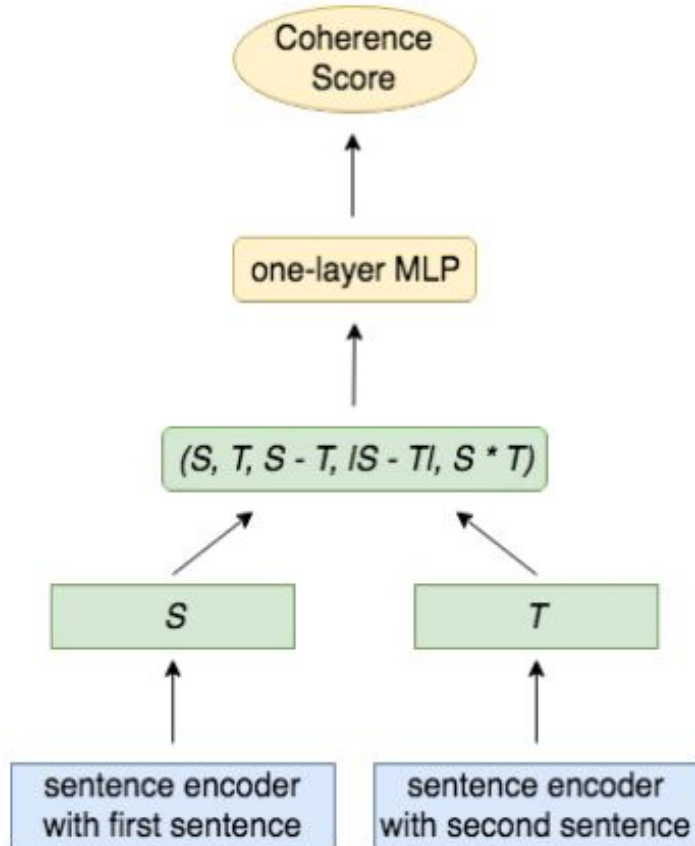
- Coherence is a discourse property used to refer to sense relations between single units (sentences or propositions) of a text. Due to these relations, the text appears to be logically and semantically consistent for the reader. Text analysis focusing on coherence is primarily concerned with the construction and configuration of sense in the text i.e. how its single constituents are connected so that the text becomes meaningful for the addressee rather than being a random sequence of unrelated sentences and clauses.
- Coherence is different from cohesion
- It is an important quality measure for text generated by humans or machines, and modelling coherence can benefit many applications, including summarization, question answering, essay scoring etc.
- In our project, we experiment with neural models in measuring textual coherence.



Our implementation

- We refer to the paper: [A Cross-Domain Transferable Neural Coherence Model](#) (Xu et Al, 2019)
- This paper comes up with a novel local coherence discriminator model (LCD), which performs better than previous global and generative neural coherence models in most cases.
- Extremely simple idea, easy to implement and get intuition
- LCD is based on the assumption that the global coherence of a document can be well approximated by the average of coherence scores between consecutive pairs of sentences. This assumption allows us to cast the learning problem as discriminating consecutive sentence pairs s_i, s_{i+1} in the training documents.

Model Architecture





Dataset

- Benchmark dataset
- Wall Street Journal (WSJ) portion of Penn Treebank



Evaluation methods

- We measure performance for the tasks of **discrimination** and **insertion**.
- In the discrimination task, a document is compared to a random permutation of its sentences, and the model is considered correct if it scores the original document higher than the permuted one.
- In the insertion task, we evaluate models based on their ability to find the correct position of a sentence that has been removed from a document



Experiments

- Different **encoders** including Averaged GloVe and SBERT
- Alter **scoring** by appending functions like sigmoid and tanh
- **Bidirectional** and unidirectional models

Comparison of baseline results

- Results with the recommended hyperparameters

Output function	Encoder	Discrimination	Insertion
None	Avg_Glove	0.92537	0.29847
Sigmoid	Avg_Glove	0.80393	0.21469
TanH	Avg_Glove	0.10488	0.72060
None	SBERT	0.93851	0.33034

Analysis of Hyperparameter tuning

Hyperparameters tuned:

- input_dropout: [0.5, 0.6, 0.7]
- hidden_layers: [1, 2]
- hidden_dropout: [0.2, 0.3, 0.4]
- margin: [4.0, 5.0, 6.0]
- weight_decay: [0.0, 0.1]
- dpout_model: [0.0, 0.05, 0.1]

Discrimination validation (Best 20 results)

Scores

- Discrimination

Description	Value
mean	0.92649
std	0.00107
min	0.925
max	0.9291
25%	0.92585
50%	0.92635
75%	0.9269

- Insertion

Description	Value
mean	0.30562
std	0.00156
min	0.3036
max	0.3091
25%	0.304275
50%	0.3054
75%	0.306825

Trends

- input_dropout: 0.5 > 0.6
- hidden_layers: 2 > 1
- margin: 6.0 > 4.0 > 5.0
- weight_decay: 0.0
- dpout_model: 0.0 > 0.1 > 0.05
- hidden_dropout
 - Discrimination: 0.3 > 0.2 > 0.4
 - Insertion: 0.3 > 0.4 > 0.2

Best model

- Same model performs best in both tasks

```
{
  "hparams": {
    "loss": "margin",
    "input_dropout": 0.5,
    "hidden_state": 500,
    "hidden_layers": 2,
    "hidden_dropout": 0.3,
    "num_epochs": 50,
    "margin": 6.0,
    "lr": 0.001,
    "weight_decay": 0.0,
    "use_bn": False,
    "task": "discrimination",
    "bidirectional": False,
    "dpout_model": 0.0
  },
}
```

Insertion validation (Best 20 results)

Scores

- Discrimination

Description	Value
mean	0.926410
std	0.001753
min	0.923900
max	0.930300
25%	0.925100
50%	0.925900
75%	0.927425

- Insertion

Description	Value
mean	0.3041
std	0.0029
min	0.301
max	0.3124
25%	0.3024
50%	0.3033
75%	0.305

Trends


- input_dropout: 0.5
- hidden_layers: 1 > 2
- hidden_dropout: 0.4 > 0.2 > 0.3
- margin: 6.0 > 5.0 > 4.0
- weight_decay: 0.0
- dpout_model: 0.05 > 0.1 > 0.0

Best model

- Same model performs best in both tasks

```
{
  "hparams": {
    "loss": "margin",
    "input_dropout": 0.5,
    "hidden_state": 500,
    "hidden_layers": 1,
    "hidden_dropout": 0.4,
    "num_epochs": 50,
    "margin": 6.0,
    "lr": 0.001,
    "l2_reg_lambda": 0.0,
    "use_bn": False,
    "task": "insertion",
    "bidirectional": False,
    "dpout_model": 0.05
  }
}
```

Results after hyperparameter tuning



Validation Task	Bidirectional	Output function	Encoder	Discr	Ins	Avg
discrimination	False	None	average_glove	0.9204	0.3028	0.6116
discrimination	False	None	sbert	0.9232	0.3139	0.61855
discrimination	False	tanh	average_glove	0.0952	0.8153	0.45525
discrimination	False	tanh	sbert	0.2989	0.2798	0.28935
discrimination	False	sigmoid	average_glove	0.6148	0.4915	0.55315
discrimination	False	sigmoid	sbert	0.2711	0.3196	0.29535
discrimination	True	None	average_glove	0.9259	0.3091	0.61749
discrimination	True	None	sbert	0.9268	0.313	0.6199
discrimination	True	tanh	average_glove	0.0001	1.0	0.50005
discrimination	True	tanh	sbert	0.807	0.3134	0.5602
discrimination	True	sigmoid	average_glove	0.713	0.3809	0.54695
discrimination	True	sigmoid	sbert	0.8113	0.2484	0.52985

Validation Task	Bidirectional	Output function	Encoder	Discr	Ins	Avg
insertion	False	None	average_glove	0.9185	0.2993	0.6089
insertion	False	None	sbert	0.9226	0.3164	0.61949
insertion	False	tanh	average_glove	0.0	1.0	0.5
insertion	False	tanh	sbert	0.5872	0.4079	0.49755
insertion	False	sigmoid	average_glove	0.0581	0.9511	0.50459
insertion	False	sigmoid	sbert	0.6704	0.1794	0.4249
insertion	True	None	average_glove	0.9191	0.2929	0.606
insertion	True	None	sbert	0.9245	0.3239	0.6242
insertion	True	tanh	average_glove	0.7525	0.3163	0.5344
insertion	True	tanh	sbert	0.7852	0.3211	0.55315
insertion	True	sigmoid	average_glove	0.6717	0.3519	0.5118
insertion	True	sigmoid	sbert	0.8125	0.3122	0.56235