Discourse Coherence in the Wild: A Dataset, Evaluation and Methods Supplementary Material

Alice Lai
University of Illinois at Urbana-Champaign*
aylai2@illinois.edu

Joel Tetreault
Grammarly
joel.tetreault@grammarly.com

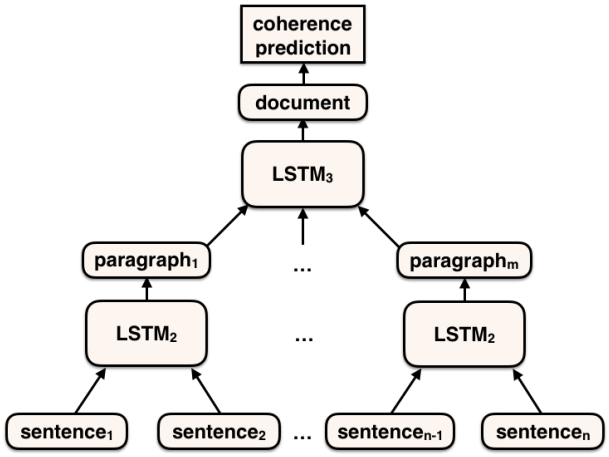

Figure 1: Structure of PARSEQ model. The sentence vectors are the output from the first LSTM (not pictured), which takes GloVe word embeddings as input.

A Supplementary Material

A.1 Corpus Examples

Table 1 contains additional examples of texts from our corpus, specifically from the Yahoo Answers domain, with their coherence labels.

A.2 Annotator Instructions

The annotation instructions in Section 2.4 are the simplified instructions that we provided to Mechanical Turk workers. The expert annotators received a longer version of those instructions, which are available in Table 2.

A.3 Model Details

Figure 1 shows the structure of PARSEQ. The sentence vectors pictured are the output at the final timestep from the first LSTM (not pictured), which takes GloVe word embeddings as input. A second LSTM takes these sentence vectors as input and produces paragraph vectors, and a third LSTM takes a sequence of paragraph vectors and produces a single document vector.

A.4 Additional Results

Table 3 contains the classification test results of all systems when the consensus labels come from the Mechanical Turk judgments rather than the expert judgments.

Table 4 contains the precision and recall results for the minority class classification test. For neural models, we report precision and recall for one run on test (F0.5 scores in Section 4.4 were averaged over 10 runs).

To compare all models on an established dataset, we report results on the sentence ordering task using the Wall Street Journal (WSJ) portion of the Penn Treebank. Following previous work, we use 20 random permutations of each article and the train/test split defined by Tien Nguyen and Joty (2017) (train = Section 00-13, test = 14-24). Table 5 contains the results of all models on WSJ. These results verify our re-implementation of the EGRID model, as well as establishing the reasonable performance of our neural sequence model on news text.

A.5 Model Parameters

We specify the parameters for all models and experiments in Tables 6 and 7. Additionally, for the combined training data experiment (Table 10 in the paper), we train parseq with LSTM dimensionality = 100, hidden layer = 200, dropout = 0.5.

EGRID *Sequence length* is the length of the transition sequences used to compute the feature vector from the entity grid. For salience, we follow Barzilay and Lapata (2008) and split entities into two salience classes (doubling the number

*Research performed while at Grammarly.

Domain	Score	Text
Yahoo	Low	I see it, but then again almost every war entered by the U.S. is connected to gaining something. The U.S. is just using politically correct was of taking over a country without anybody noticing it. They enter a war and some how we come out better than the country we went in to help. We say we are helping but if the country has nothing for us then we don't bother with it. For example: Korea stated and I quote "we have nuclear weapons and we plan to use them" so how come we are in Iraq who have no weapons? Well maybe the U.S. sees no threat but then again somebody did sneak into the country and take over planes. Also not to long ago it was common for somebody to hijack a plane. Well that is all I have to say on the matter.
Yahoo	High	Don't be intimidated by Impressionism. It is simply a style worked in loose strokes. The idea is to give an "impression" of the subject. Choose a simple subject, like a still life or bowl of fruit. Then layout your palette using the colors you see (make sure to look for subtle colors only an artist might see...such as the "blue" in an apple), and with a larger than usual brush, stroke the basic shapes in a medium value, then add shadows, then a highlight layer. That should do for a class project in Impressionism. The danger would come from over-working the painting. You don't want fine strokes or details, remember just the "impression" of your subject. The whole idea is to stay loose and free. A lot of people struggle with it. The trick is to just paint without worrying too much. Good luck.

Table 1: Examples of texts with coherence scores.

You will be given a short text (100-300 words) to read. We will specify which one of several domains the text comes from, and in some domains we will provide additional context for the text.

Your task is to rate the coherence of the text from 1 to 3 (1 means low coherence, 3 means high coherence).

Coherence in writing refers to how well ideas flow from one sentence to the next, and from one paragraph to the next. A text that is highly coherent is easy to understand and easy to read. This usually means the text is well-organized, logically structured, and presents only information that supports the main idea. On the other hand, a text with low coherence is difficult to understand. This may be because the text is not well organized, contains unrelated information that distracts from the main idea, or lacks transitions to connect the ideas in the text.

Try to ignore the effects of grammar or spelling errors when assigning a coherence rating, as long as the errors do not significantly interfere with your ability to read and understand the text. In the email data, assume that jargon and acronyms are used correctly, and do your best to judge coherence despite that.

You should assign a coherence rating to the text based on whether it is a coherent example of text *in that domain*. A reader has different expectations about how a business email should be written compared to a post on an online forum, and the coherence rating should reflect this difference. A business email with a score of 1 is not necessarily incoherent in the same way that a very incoherent Yahoo Answers post is, but it is not very coherent *for a business email*.

Table 2: The annotation instructions we provided to expert annotators.

	Accuracy			
System	Yahoo	Clinton	Enron	Yelp
Majority class	39.5	40.5	44.0	40.5
Baseline	35.0	43.5	45.0	41.5
EGRID	43.0	41.0	45.5	43.0
EGRAPH	39.5	41.5	44.5	40.5
EGRIDCONV	41.0	43.5	44.5	54.0
LEXGRAPH	38.0	36.0	48.0	45.5
CLIQUE	48.0	45.0	52.5	51.0
SENTAVG	**52.0**	48.5	55.5	49.0
PARSEQ	47.5	**51.0**	**56.5**	**57.5**

Table 3: Three-way classification results on test data. Untrained rater judgments.

of features) based on whether their frequency is greater than the *salience threshold*. (Salience = off means that there is only one salience class containing all entities.) *Syntax* indicates whether we consider grammatical roles (subject, object, other) in building the entity grid.

EGRAPH The *graph type* specifies whether we use an unweighted graph (u), a graph weighted by the number of entities shared between sentences (w), or a graph weighted by syntactic role information (syn). *Distance* indicates whether edge weights are decreased according to the distance between sentences.

EGRIDCONV We specify dropout rate, batch size, and entity role embedding size. For the convolution layer, we specify filter number, window size, and pooling length.

LEXGRAPH We define the similarity *threshold* used to filter out edge weights between sentences, and k as the size of the subgraphs we consider when extracting features from the document graph.

CLIQUE We define the dropout rate, the LSTM dimensionality, and the hidden layer dimensionality. *Window size* is the number of sentences in a

System	Yahoo		Clinton		Enron		Yelp	
	p	r	p	r	p	r	p	r
Baseline	0.25	0.61	0.26	0.24	0.33	0.38	0.17	0.42
EGRID	0.31	0.16	0.36	0.12	0.57	0.10	0.33	0.05
EGRAPH	0.26	0.94	0.35	0.58	0.25	0.45	0.10	0.68
EGRIDCONV	0.31	0.41	0.16	0.24	0.22	0.40	0.50	0.05
LEXGRAPH	0.26	0.29	0.20	0.03	0.55	0.15	0.00	0.00
CLIQUE	0.07	0.03	0.00	0.00	0.17	0.03	1.00	0.05
SENTAVG	0.38	0.73	0.39	0.36	0.42	0.33	0.36	0.21
PARSEQ	0.43	0.51	0.21	0.39	0.57	0.20	0.13	0.11

Table 4: Minority class predictions, precision/recall results on test data.

System	Accuracy
Random baseline	50.0
EGRID	83.0
EGRAPH	65.7
EGRIDCONV	82.2
LEXGRAPH	72.7
CLIQUE	60.9
SENTSEQ	74.1

Table 5: Sentence ordering results on WSJ test data.

clique.

SENTAVG, PARSEQ For both models, we specify the dropout rate, the LSTM dimensionality, and the hidden layer dimensionality. For PARSEQ, the LSTM dimensionality applies to all 3 LSTMs.

Model	Parameter	Classification				Score Prediction			
		Yahoo	Clinton	Enron	Yelp	Yahoo	Clinton	Enron	Yelp
Baseline	threshold1	6.5	6.5	6.0	2.5	–	–	–	–
	threshold2	7.0	7.0	6.5	3.0	–	–	–	–
EGRID	sequence length	4	3	4	2	2	2	4	3
	salience threshold	off	2	4	4	2	off	3	2
	syntax	on	off	on	on	off	off	on	on
EGRAPH	graph type	syn	syn	syn	syn	u	w	w	syn
	distance	no	no	no	no	yes	yes	yes	no
	threshold1	15.0	0.1	0.1	0.5	–	–	–	–
	threshold2	16.0	1.1	1.1	1.6	–	–	–	–
EGRIDCONV	dropout	0.2	0.2	0.5	0.2	0.2	0.5	0.5	0.2
	filter	100	100	100	200	200	200	200	100
	window	4	2	2	6	2	2	2	4
	pool	3	7	3	5	5	3	3	3
	batch	128	128	32	128	32	32	32	32
	embedding size	100	100	100	200	100	200	200	100
LEXGRAPH	threshold	0.7	0.5	0.7	0.9	0.5	0.3	0.7	0.9
	k	6	6	6	5	6	6	4	5
CLIQUE	dropout	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	LSTM dim	100	100	200	100	100	100	200	100
	hidden dim	200	200	200	200	200	200	200	100
	window size	3	3	3	7	7	7	7	4
SENTAVG	dropout	0.5	0.5	0.5	0.2	0.2	0.5	0.5	0.2
	LSTM dim	200	50	200	50	300	300	300	300
	hidden dim	200	50	100	300	100	100	50	50
PARSEQ	dropout	0.5	0.5	0.5	0.5	0.5	0.2	0.5	0.2
	LSTM dim	200	300	50	50	300	200	100	300
	hidden dim	100	100	100	200	100	50	100	100

Table 6: Best parameter values for classification and score prediction experiments.

		Sentence Ordering					Minority Class			
Model	Parameter	Yahoo	Clinton	Enron	Yelp	WSJ	Yahoo	Clinton	Enron	Yelp
Baseline	threshold1	–	–	–	–	–	8.0	6.5	6.0	5.0
	threshold2	–	–	–	–	–	–	–	–	–
EGRID	sequence length	4	4	4	4	3	2	2	2	3
	salience threshold	4	off	4	off	4	off	off	2	2
	syntax	on	on	off	on	on	off	off	on	off
EGRAPH	graph type	syn	w	w	w	w	u	w	w	w
	distance	yes	yes	yes	yes	yes	yes	yes	yes	no
	threshold1	–	–	–	–	–	1.2	0.5	0.9	2.2
	threshold2	–	–	–	–	–	–	–	–	–
EGRIDCONV	dropout	0.2	0.2	0.2	0.2	0.5	0.2	0.5	0.5	0.5
	filter	100	100	100	100	150	100	200	200	200
	window	6	6	4	6	6	2	4	6	6
	pool	7	7	7	7	6	3	3	5	7
	batch	32	32	32	128	128	128	32	32	32
	embedding size	200	100	200	100	100	100	200	100	200
LEXGRAPH	threshold	0.9	0.9	0.9	0.9	0.3	0.5	0.7	0.5	0.9
	k	4	3	4	4	4	4	3	3	3
CLIQUE	dropout	0.5	0.2	0.2	0.2	0.2	0.2	0.2	0.5	0.2
	LSTM dim	100	100	100	300	300	50	50	50	50
	hidden dim	100	100	50	50	50	50	300	50	200
	window size	7	5	5	5	7	5	7	5	7
SENTAVG	dropout	–	–	–	–	–	0.2	0.5	0.2	0.2
	LSTM dim	–	–	–	–	–	200	200	50	200
	hidden dim	–	–	–	–	–	300	200	50	50
PARSEQ	dropout	0.5	0.2	0.2	0.2	0.5	0.5	0.2	0.5	0.2
	LSTM dim	50	300	300	300	300	100	50	200	300
	hidden dim	200	300	200	100	200	50	300	50	100

Table 7: Best parameter values for sentence ordering and minority class classification experiments.