

Seq2Seq in Action: Column Segmentation

Part 1

Bill Watson

S&P Global

January 23, 2020

Types of Recurrent Models

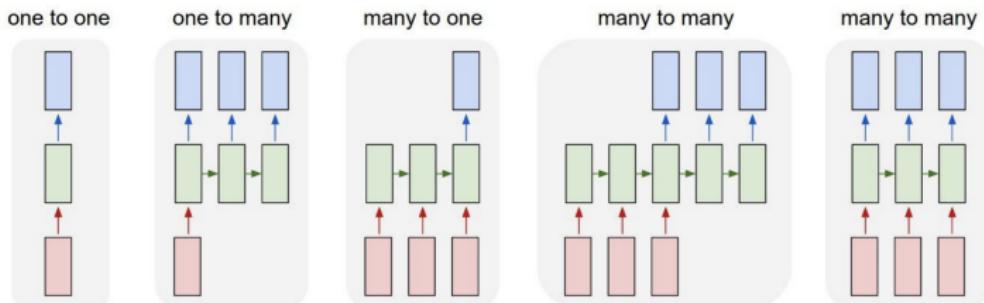


Figure: Source: Fei-Fei Li & Justin Johnson & Serena Yeung, Standford CS231, Recurrent Neural Networks

Type	Input T_x	Output T_y	Example
One-to-One	$T_x = 1$	$T_y = 1$	Vanilla NN
One-to-Many	$T_x = 1$	$T_y > 1$	Image Captioning
Many-to-One	$T_x > 1$	$T_y = 1$	Sentiment Classification
Many-to-Many	$T_x \neq T_y$	$T_x \neq T_y$	Machine Translation
Many-to-Many	$T_x = T_y$	$T_x = T_y$	Name Entity Recognition

What are Sequence to Sequence Models?

- ▶ Generally used to convert one set of tokens into another
 - ▶ Many - to - Many RNN
- ▶ Bleak view: map a sequence of indexes to another independent set of indexes



Figure: The pile gets soaked with data and starts to get mushy over time, so it's technically recurrent.

Use Case: Extracting Tables from Images

Primer: What are we trying to accomplish

- ▶ Tabular data is locked in PDFs
 - ▶ Either as a typeset document
 - ▶ Or as a scanned image
- ▶ Sometimes copy & paste won't suffice (or impossible with images)
- ▶ Let's take a quick look at the full pipeline to see where Seq2Seq models can fit in

What is our data?

Schedule 2 1 of 41				
CITY OF COFFEYVILLE, KANSAS GENERAL FUND Schedule of Receipts and Expenditures - Actual and Budget For the Year Ended December 31, 2017 (With Comparative Actual Amounts for the Prior Year Ended December 31, 2016)				
Current Year				
Year	Actual	Budget	Variance Over (Under)	Total
Receipts				
Taxes and Shared Receipts In Addition to Property Tax	\$ 1,779,884.36	\$ 1,873,543.39	\$ 2,205,362.00	\$ (255,814.63)
Delinquent Tax	357,335.16	85,439.02	265,436.00	23,436.02
Motor Vehicle Tax	188,143.40	200,000.00	170,000.00	(11,856.60)
Recreational Vehicle Tax	1,275.20	-	1,521.00	255.98
(A) 2016 Motor Vehicle Tax Refund Received	20.25	1,244.44	861.00	(275.58)
Commercial Vehicle Tax	95.35	51.83	117.00	21.62
Commercial Vehicle Tax Refund Received	94.95	94.95	117.00	22.05
Special Assessment	607.49	-	1,796.32	(1,188.83)
Debt Service	33,217.47	31,620.34	40,000.00	43,836.34
Interest on Debt	6,045.47	5,649.00	6,420.00	5,964.57
State Tax	3,137,239.37	4,850,214.72	5,145,316.00	(285,301.38)
Local Alcohol Liquor Tax	18,091.23	28,978.71	16,327.00	32,531.71
Special Highway Tax	203,010.00	208,967.36	206,000.00	(205,044.00)
Highway User Tax Highway User Tax Links	78,750.00	78,750.00	78,750.00	0.00
Highway County Aid	65,094.05	65,975.73	65,830.00	145.73
Landfill Tax	194,125.00	194,125.00	194,125.00	0.00
Fines, Forfeitures and Penalties	163,440.84	207,414.76	308,445.00	(181,206.26)
Charitable Contribution	1,000.00	2,000.00	2,000.00	(1,000.00)
Use of Money and Property	50,461.28	23,205.49	13,000.00	52,025.49
Interest Income	67,425.00	38,490.00	67,000.00	(30,907.00)
Sale of Equipment and Scrap	3,259.49	3,294.94	5,000.00	(1,705.00)
Other Income	-	-	-	-
Donations	-	-	-	-
Administrative Expense	44,132.89	49,100.82	49,400.00	(40,262.13)
Insurance Premiums	8,239.70	36,750.00	36,750.00	0.00
Miscellaneous	2,230.79	2,400.00	6,000.00	(3,470.20)
Operating Transfers From:				
City of Coffeyville	2,327,005.74	2,485,862.76	2,265,000.00	(178,868.00)
Water and Sewer Utility Fund	459,324.64	777,654.10	723,775.00	53,375.10
Community Development Fund	1,626,632.00	-	-	-
Total Receipts	\$ 13,853,737.06	\$ 15,249,775.00	\$ 15,259,394.00	\$ 1,202,005.96
Expenditures				
General Government	836,060.11	782,088.99	987,381.00	\$ (299,299.01)
Personal Services	208,200.00	240,000.00	215,777.00	51,777.00
Contractual Services	19,780.03	12,130.62	14,240.00	(2,399.38)
Commodities	1,842.98	1,803.93	18,600.00	(16,817.08)
Capital Outlay	-	-	-	-

- 23 -

5. CAPITAL LEASE OBLIGATIONS

The City has entered into a capital lease agreement in order to finance the purchase of a Fire Truck. Payments for the year ended December 31, 2017 were \$77,193.00. The remaining amount at approximately 5.02%. Final maturity for the lease is in 2028. Future minimum lease payments are as follows:

Year Ended December 31		Total
2018	\$ 155,458.46	
2019	155,458.46	
2020	155,458.46	
2021	155,458.46	
2022	155,458.46	
2023	155,458.46	
2024	155,458.46	
2025	155,458.46	
2026	155,458.46	
2027	155,458.46	
2028	155,458.46	
Less Impaired Interest	1,283,162.91	
Net Present Value of Minimum Lease Payments	(1,125,804.45)	
Less Current Maturing Obligation	111,238.87	
Long-Term Capital Lease Obligation	\$ 390,063.62	

The City has entered into a capital lease agreement in order to finance the purchase of a Fire Truck. Payments for the year ended December 31, 2016 were \$30,033.18. The remaining amount at approximately 3.27%. Final maturity for the lease is in 2027. Future minimum lease payments are as follows:

Year Ended December 31		Total
2018	\$ 60,094.36	
2019	60,094.36	
2020	60,094.36	
2021	60,094.36	
2022	60,094.36	
2023	60,094.36	
2024	60,094.36	
2025	60,094.36	
2026	60,094.36	
2027	60,094.36	
Less Impaired Interest	600,342.51	
Net Present Value of Minimum Lease Payments	(539,250.15)	
Less Current Maturing Obligation	50,663.60	
Long-Term Capital Lease Obligation	\$ 143,858.93	
Less Current Maturing Obligation	\$ 143,858.93	

6. OPERATING LEASES

As of December 31, 2017 the City has entered into an operating lease for office equipment. Rent expense for the year ended December 31, 2017 was \$10,369.36. Under the current lease agreements, the future minimum lease payments are as follows:

Period	Total
2018	\$ 10,369.36
2019	3,182.31
2020	1,141,270.00
2021	1,141,270.00
2022	1,141,270.00
2023	1,141,270.00
2024	1,141,270.00
2025	1,141,270.00
2026	1,141,270.00
2027	1,141,270.00
Less Current Maturing Obligation	1,141,270.00
Future Minimum Lease Payments	\$ 10,369.36
Less Current Maturing Obligation	\$ 10,369.36
Future Minimum Lease Payments	\$ 0.00

- 16 -

Schedule 1

CITY OF COFFEYVILLE, KANSAS Statement of Expenditures and Revenues and Budget For the Year Ended December 31, 2017	
Total Expenditures	\$ 1,202,005.96

Period	Total
2018	\$ 10,369.36
2019	3,182.31
2020	1,141,270.00
2021	1,141,270.00
2022	1,141,270.00
2023	1,141,270.00
2024	1,141,270.00
2025	1,141,270.00
2026	1,141,270.00
2027	1,141,270.00
Less Current Maturing Obligation	1,141,270.00
Future Minimum Lease Payments	\$ 10,369.36
Less Current Maturing Obligation	\$ 10,369.36
Future Minimum Lease Payments	\$ 0.00



Current Pipeline for Table Extraction

- ▶ Convert to an Image
- ▶ Binary Page Classification: Is there a table on this page?
- ▶ Table Segmentation: U-Net Model
- ▶ Optical Character Recognition via Tesseract
- ▶ **Column Segmentation**
- ▶ Table Alignment
- ▶ Dump to CSV

Visual: Table Segmentation

Schedule 2 1 of 41					
CITY OF COFFEYVILLE, KANSAS GENERAL FUND Schedule of Receipts, Disbursements, Actual and Budget Regulatory Basis For the Year Ended December 31, 2017 (With Comparative Actual Amounts for the Prior Year Ended December 31, 2016)					
Description	Current Year			Previous Year	
	Year Actual	Year Actual	Budgeted	Over (Under)	Over (Under)
Taxes and Shared Receipts					
Ad Valorem Property Tax	\$ 1,770,894.38	\$ 1,875,245.38	\$ 2,025,360.00	\$ (250,864.62)	
Delinquent Tax	157,235.16	45,459.02	55,000.00	(7,527.91)	
Motor Vehicle Tax	1,261.00	175,000.00	200,000.00	(23,739.00)	
Business Personal Tax	1,375.00	1,296.00	1,821.00	(525.00)	
1/2/20 M Vehicle Tax	827.30	566.64	841.00	(173.66)	
Use Tax	92.00	1,000.00	1,000.00	(908.00)	
Commercial Vehicle Tax	9,660.34	9,861.23	8,127.00	1,732.33	
Special Assessments	4,625.00	4,625.00	5,125.00	(500.00)	
Special Assessments Special Assessments	43,573.47	35,635.34	30,000.00	43,633.34	
Fire Protection Tax	400,470.00	400,470.00	450,000.00	(49,530.00)	
Sales Tax	5,117,229.57	4,825,219.72	5,145,516.00	(323,296.42)	
Federal Income	4,645.00	4,645.00	4,645.00		
State Income	18,551.55	18,879.71	16,521.71	3,358.04	
Employment Lagoon Tax	10,000.00	10,000.00	10,000.00		
Special Highway Tax	209,220.61	208,026.38	206,330.00	(205,494.41)	
Emergency Service Units	76,700.00	76,700.00	76,700.00		
Highway County Aid	45,478.00	45,975.73	40,820.00	5,140.73	
Leisure Tax	10,125.00	10,125.00	10,000.00	925.00	
Fines, Penalties and Recoveries	181,400.84	201,419.73	208,441.00	(10,031.25)	
Disbursements	40,612.00	40,612.00	40,612.00		
Use of Money and Property	40,412.28	28,000.00	32,000.00	(2,888.72)	
Interest Income	40,412.28	28,000.00	32,000.00	(2,888.72)	
Rents	87,435.00	86,000.00	87,400.00	(1,300.00)	
Sale of Equipment and Scrap	3,238.40	2,296.94	0,000.00	(1,730.46)	
Other Disbursements	1,180.00	-	1,180.00		
Disbursements	46,155.88	46,155.88	46,155.88		
Reimbursement Expense	1,000.00	-	1,000.00		
Insurance Premiums	38,750.00	-	38,750.00		
Reserves	2,036.78	4,526.00	6,000.00	(1,473.78)	
Operating Transfers In:					
General Fund	2,036.78	4,526.00	6,000.00	(1,473.78)	
Waste and Sewer Utility Fund	658,334.64	717,056.13	732,775.00	(55,379.35)	
Community Development Fund	1,403.83	-	1,403.83		
Special Assessment Fund	17,000.00	-	17,000.00		
Disbursements	808,686.11	762,088.99	987,262.00	(200,293.81)	
General Fund	808,686.11	762,088.99	987,262.00	(200,293.81)	
Emergency Service	100,000.00	100,000.00	100,000.00		
Commodities	10,780.03	12,280.62	14,245.00	(2,155.38)	
Capital Outlay	1,403.83	18,032.93	18,450.00	(16,617.08)	

Schedule 2 1 of 41					
5. CAPITAL LEASES OBLIGATIONS					
The City has entered into a capital lease agreement in order to finance the purchase of a Fire Truck. Payment of principal and interest of \$77,739.18 per year, beginning January 1, 2018, at approximately 5.00%. Final maturity for the lease is in 2036. Future minimum lease payments are as follows:					
Total Stated Obligation: \$1,551,458.46					
Year Ended December 31:					
2018	\$ 77,739.18				
2019	75,518.46				
2020	73,307.73				
2021	71,097.00				
2022	68,886.28				
2023	66,675.55				
2024	64,464.82				
2025	62,254.09				
2026	60,043.36				
2027	57,832.63				
2028	55,621.90				
2029	53,411.17				
2030	51,199.44				
2031	48,987.71				
2032	46,776.98				
2033	44,566.25				
2034	42,355.52				
2035	40,144.79				
2036	37,934.06				
Less Interest Accrued:					
2018 \$1,551,458.46					
2019 \$115,062.00					
2020 \$115,062.00					
2021 \$115,062.00					
2022 \$115,062.00					
2023 \$115,062.00					
2024 \$115,062.00					
2025 \$115,062.00					
2026 \$115,062.00					
2027 \$115,062.00					
2028 \$115,062.00					
2029 \$115,062.00					
2030 \$115,062.00					
2031 \$115,062.00					
2032 \$115,062.00					
2033 \$115,062.00					
2034 \$115,062.00					
2035 \$115,062.00					
2036 \$115,062.00					
Less Present Value of Minimum Lease Payments:					
\$115,062.00					
Less Current Mortisities:					
\$115,062.00					
Less Future Capital Lease Obligations:					
\$115,062.00					

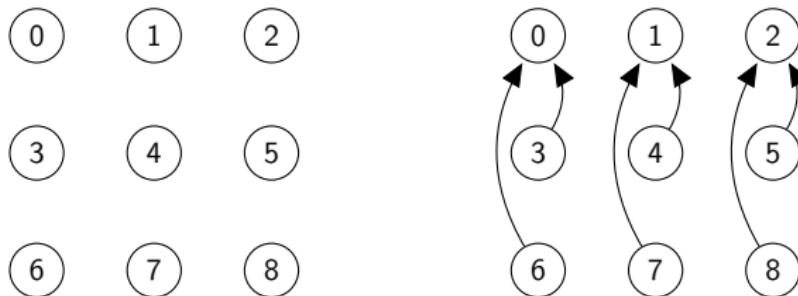
Schedule 2 1 of 41					
6. OPERATING LEASES					
As of December 31, 2017 the City has entered into an operating lease for office equipment. Rent expense for the year ended December 31, 2017, was \$4,216.68. Under the current lease agreement, the future minimum lease payments are as follows:					
Total \$4,216.68					
Year Ended December 31:					
2018	\$ 4,216.68				
2019	3,165.51				
2020	2,114.34				
2021	1,063.17				
2022	516.00				
2023	258.00				
2024	129.00				
2025	64.50				
2026	32.25				
2027	16.13				
2028	8.06				
2029	4.03				
2030	2.01				
2031	1.01				
2032	0.50				
2033	0.25				
2034	0.13				
2035	0.06				
2036	0.03				
Less Future Capital Lease Obligations:					
\$4,216.68					
Less Future Capital Lease Payments:					
\$4,216.68					

Visual: Tesseract OCR Output

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
5	1	1	1	1	1	1092	100	62	27	95	For
5	1	1	1	1	2	1165	100	55	26	95	the
5	1	1	1	1	3	1239	100	82	25	96	Year
5	1	1	1	1	4	1335	100	114	24	96	Ended
5	1	1	1	1	5	1463	100	82	24	93	June
5	1	1	1	1	7	1630	100	88	22	36	ZU18
5	1	1	1	2	1	1295	146	219	43	96	(Continued)
5	1	1	1	3	1	460	263	128	32	96	NOTE
5	1	1	1	3	2	605	263	16	31	90	1
5	1	1	1	3	3	637	280	13	5	87	-
5	1	1	1	3	4	665	260	245	34	96	SUMMARY
5	1	1	1	3	5	925	260	61	32	95	OF
5	1	1	1	3	6	1002	257	303	34	96	SIGNIFICANT
5	1	1	1	3	7	1319	254	318	34	95	ACCOUNTING
5	1	1	1	3	8	1651	252	220	34	95	POLICIES
5	1	1	1	3	9	1886	251	232	42	96	(Continued)
5	1	1	1	4	1	537	372	39	31	93	H.
5	1	1	1	4	2	612	369	206	34	96	Inventories
5	1	1	1	4	3	833	368	64	33	96	and
5	1	1	1	4	4	911	367	141	43	96	Prepaid
5	1	1	1	4	5	1066	368	98	31	96	Items
5	1	1	1	5	1	611	477	207	34	96	Inventories
5	1	1	1	5	2	831	487	54	22	96	are
5	1	1	1	5	3	896	476	123	32	96	valued
5	1	1	1	5	4	1031	480	32	27	96	at
5	1	1	1	5	5	1073	475	55	32	96	the
5	1	1	1	5	6	1141	474	101	33	96	lower
5	1	1	1	5	7	1253	473	41	33	96	of
5	1	1	1	5	8	1301	477	73	28	93	cost
5	1	1	1	5	9	1385	471	147	42	90	(first-in,
5	1	1	1	5	10	1545	470	163	41	96	first-out)

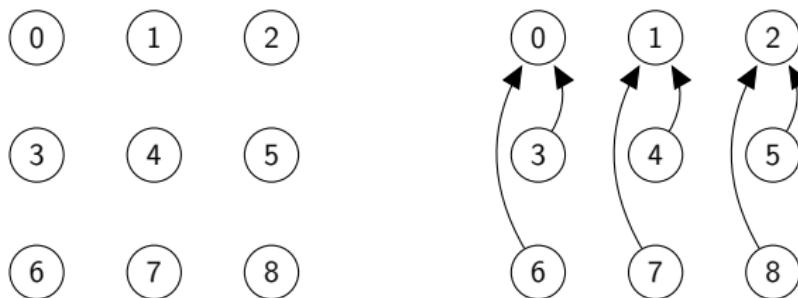
Column Segmentation

- ▶ Naive approach is to merge words close together
 - ▶ But spacing is **variable** in tables
 - ▶ Text justification (left, center, right)
- ▶ **Union-Find Disjoint Set (UFDS) Algorithm** to group columns by **Intersection Over Union (IOU)**



Column Segmentation

- ▶ Naive approach is to merge words close together
 - ▶ But spacing is **variable** in tables
 - ▶ Text justification (left, center, right)
- ▶ **Union-Find Disjoint Set (UFDS) Algorithm** to group columns by **Intersection Over Union (IOU)**



- ▶ **Idea:** Use a Seq2Seq to learn correct segmentation

Visual: Alignment

-- -- -- #SumYear1-(SOSCS-#BS)					
	Prior				Variance -
	Year				Over
	Actual		Actual	Budget	(Under)
Receipts					
Taxes and Shared Receipts					
Ad Valorem Property Tax	\$1,779,884.26		\$1,872,543.39	\$8	2,225,362 \$ -352,818.61
Delinquent Tax	107,335.16		83,439.02	60,000	23,439.02
Motor Vehicle Tax	185,104.64		198,703.03	206,631	-7,927.97
Recreational Vehicle Tax	1,372.3		1,786.98	1,521	265.98
16/20 M Vehicle Tax	827.2		566.64	841	-274.36
Vehicle Rental Excise Tax	95.1		51.82	177	-125.18
Commercial Vehicle Tax	9,660.34		9,861.33	8,137	1,724.33
Watercraft Tax	657.49		532.7	517	15.7
Special Assessments	63,572.47		73,635.34	30,000	43,635.34
Franchise Tax	604,481.17		688,423.06	476,972	211,491.06
Sales Tax	5,137,239.57		4,852,214.72	5,145,516	-293,301.28
Federal Grants	8,650.01		2,705.56	-	2,705.56
Local Alcohol Liquor Tax	18,091.55		28,878.71	16,327	12,561.71
Special Highway Tax	259,225.61		256,024.36	256,330	-305.64
Highway Connecting Links	76,750.82		76,645.19	76,700	-54.81
Highway County Aid	45,694.05		45,975.73	40,830	5,145.73
Licenses and Permits	94,108.5		92,606	-	92,606
Fines, Forfeitures and Penalties	183,460.84		207,414.75	308,445	-101,030.25
Charges for Services	99,811.84		90,317.95	96,596	-6,278.05
Use of Money and Property					
Interest Income	20,461.28		21,263.49	12,000	9,263.49
Rents	67,423		36,490.5	67,400	-30,907.5
Sale of Equipment and Scrap	2,228.4		3,294.94	5,000	-1,705.06
Other Receipts					
Donations	-		1,180	-	1,180
Reimbursed Expense	46,151.89		40,804.82	-	40,804.82
Insurance Proceeds	-		38,750	-	38,750
Miscellaneous	8,239.7		4,526.25	6,000	-1,473.75
Operating Transfers from:					
Electric Utility Fund	2,327,045.74		2,408,082.7	2,583,695	-175,612.3
Water and Sewer Utility Fund	694,334.64		777,054.1	723,775	53,279.1
Community Development Fund	218,655.33		-	-	-
Total Receipts	12,060,562.9		11,913,775.08	\$8	12,348,772 \$5 -434,996.92
Expenditures			TT		
General Government					
Personal Services	838,080.11		782,088.99	\$	987,382 \$8 -205,293.01
Contractual Services	259,202.84		269,145.94	313,797	-44,651.06
Commodities	10,788.03		12,130.62	14,340	-2,169.58
Capital Outlay	1,442.98		1,832.92	18,450	-16,617.08

Problem Statement: Column Segmentation

Year Ended December 31	Totals	Year Ended December 31	Totals
2018	\$ 155,458.46	2018	\$155,458.46
2019	155,458.46	2019	155,458.46
2020	155,458.46	2020	155,458.46
2021	155,458.46	2021	155,458.46
2022	155,458.46	2022	155,458.46
2023-2026	<u>505,870.61</u>	2023-2026	<u>505,870.61</u>
	1,283,162.91		1,283,162.91
Less imputed interest	(174,862.64)		
Net Present Value of Minimum Lease Payments	1,108,300.27	Less imputed interest	-174,862.64
Less: Current Maturities	<u>(118,236.87)</u>	Net Present Value of Minimum Lease Payments	
Long-Term Capital Lease Obligations	<u>\$ 990,063.40</u>	Less: Current Maturities	-118,236.87
		Long-Term Capital Lease Obligations	\$990,063.40

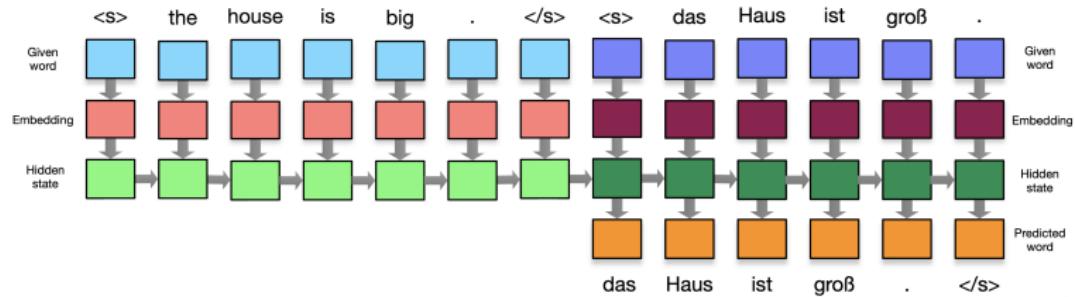
- Given a sequence of text, can we predict where to insert columns?

Revisiting Column Segmentation as a Seq2Seq Problem

- ▶ Translation Task
- ▶ Given a sequence of tokens
- ▶ Translate into the same set of tokens, but with a special **(SEG)** token

Less imputed interest (174,862.64)
↓
Less imputed interest **(SEG)** (174,862.64)

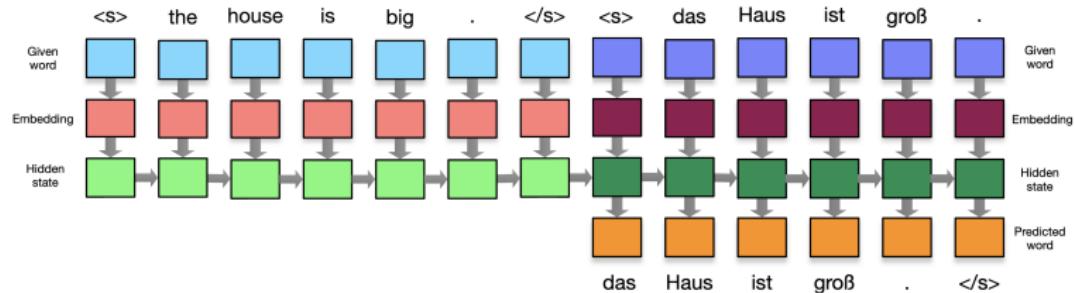
Seq2Seq Architecture



- ▶ We can use a **Encoder-Decoder** style model!

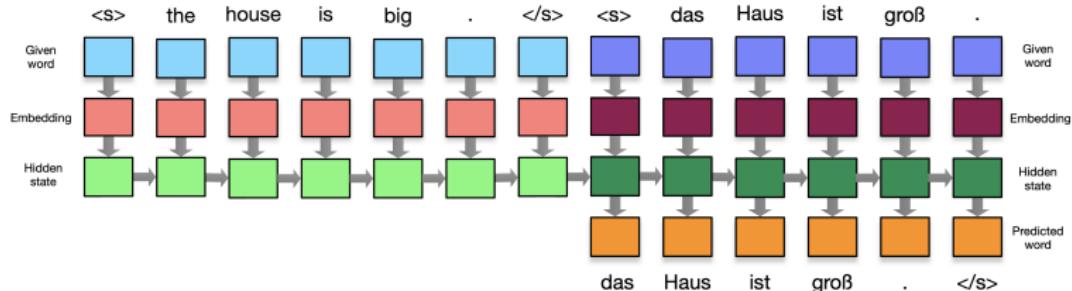
Encoder-Decoder Architecture

Overview: Encoders and Decoders



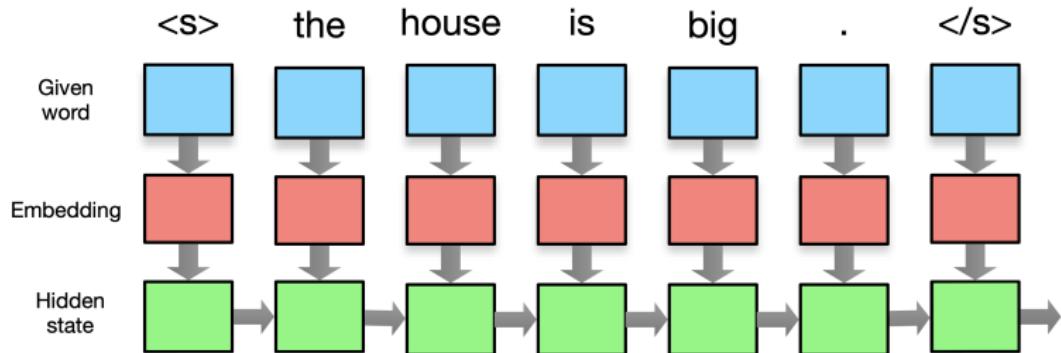
- ▶ We have 2 sub-networks: an **Encoder** and a **Decoder**
- ▶ Encoders
 - ▶ Give the source sentence meaning
- ▶ Decoders
 - ▶ Emit a variable-length sequence
- ▶ We will discuss how to connect the two for joint training

Overview: Encoders and Decoders



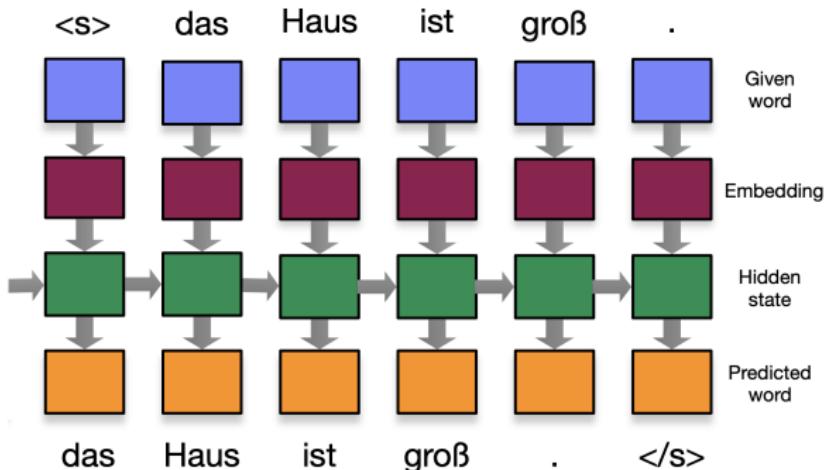
- ▶ Encapsulation allows for flexible design choices
- ▶ **Embeddings**
 - ▶ Pre-trained
 - ▶ DIY
- ▶ **Recurrent Layer**
 - ▶ Type
 - ▶ Depth
 - ▶ Directionality

What makes an Encoder?



- ▶ Recall: **Encoders** give the source sentence meaning
- ▶ Effectively a language model, without a layer to predict the next word
- ▶ Idea is to pass on the hidden state, and possibly use the encodings directly

What makes a Decoder?



- ▶ Recall: **Decoders** provide a new sequence conditioned on the Encoder's hidden state
- ▶ Starts with the Encoder's hidden state, and predicts one token at a time
- ▶ Re-feed the predicted token back into the decoder

Word Embeddings: Practical Considerations

Column Segmentation Issue: Vocabulary Size

- ▶ Is the text actually important?
- ▶ Over 10K **unique tokens**, mostly numbers

Column Segmentation Issue: Vocabulary Size

- ▶ Is the text actually important?
- ▶ Over **10K unique tokens**, mostly numbers
- ▶ **Solutions:**
 - ▶ Token → Part-of-Speech
 - ▶ Numeric Value → **(NUM)**
- ▶ Reduces vocabulary size to **21 tokens**
- ▶ But lets take a look at alternative solutions to managing vocab size

Recap: Word Embeddings

$$\begin{array}{l} \text{the} \rightarrow \begin{bmatrix} 1.23 & -0.58 & 0.22 & -0.80 & 0.61 \end{bmatrix} \\ \text{cat} \rightarrow \begin{bmatrix} -1.10 & 1.23 & 0.17 & -0.21 & 1.43 \end{bmatrix} \\ \text{is} \rightarrow \begin{bmatrix} -0.26 & 0.70 & 0.27 & 0.59 & -1.04 \end{bmatrix} \\ \text{black} \rightarrow \begin{bmatrix} -1.13 & -0.81 & -0.53 & -0.59 & 0.26 \end{bmatrix} \end{array}$$

- ▶ A trick to map tokens to vector representations

Recap: Pretrained Word Embeddings

- ▶ Co-occurrence Matrix
- ▶ Pointwise Mutual Information
- ▶ SVD Co-occurrence
- ▶ Ngram
- ▶ CBOW, Skip Gram
- ▶ GloVe
- ▶ ELMo
- ▶ BERT

DIY Embeddings

- ▶ We can always train our own!

Special Tokens for Sequence Modeling

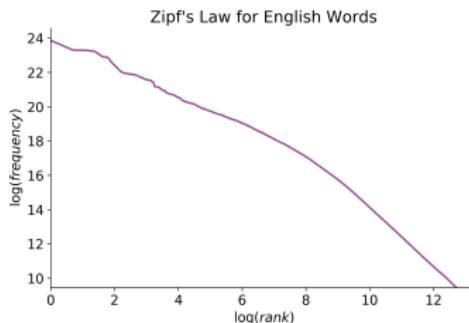
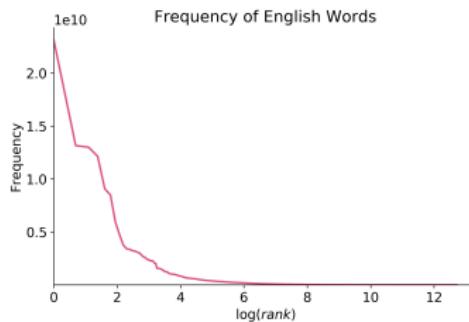
Special Tokens for Sequence Modeling

- ▶ $\langle \text{PAD} \rangle$ → Padding / Masking
- ▶ $\langle \text{UNK} \rangle$ → Unknown words
- ▶ $\langle \text{SOS} \rangle$ → Start of Sequence
- ▶ $\langle \text{EOS} \rangle$ → End of Sequence

```
<SOS> Data Science is the <UNK> ! <EOS>
<SOS> I <UNK> for S&P . <EOS> <PAD>
<SOS> Hello World ! <EOS> <PAD> <PAD> <PAD>
```

Managing the Vocab Size

- ▶ Languages are unevenly distributed
- ▶ Many rare words, names → inflates the size of the vocabulary
- ▶ **Problem:**
 - ▶ Large embedding matrices for source, target language
 - ▶ Large output layers for prediction and softmax



Morphology, Compounding, and Transliteration

- ▶ Morphological Analysis

tweet, tweets, tweeted, tweeting, retweet, ...

- ▶ Compound Splitting

- ▶ `homework` → `home·work`
- ▶ `website` → `web·site`

- ▶ Names, Places, Proper Nouns

- ▶ `Hoboken, Baltimore, Obama, Michelle`
- ▶ Can do Transliteration

Handling Numbers

- ▶ Do we really need to encode every number? **NO!**

I pay **950.00** in May **2007** > I pay **2007** in May **950.00**

Handling Numbers

- ▶ Do we really need to encode every number? **NO!**

I pay **950.00** in May **2007** > I pay **2007** in May **950.00**

- ▶ **Solution 1:** Replace with a **<NUM>** token, but

I pay **<NUM>** in May **<NUM>** = I pay **<NUM>** in May **<NUM>**

Handling Numbers

- ▶ Do we really need to encode every number? **NO!**

I pay 950.00 in May 2007 > I pay 2007 in May 950.00

- ▶ **Solution 1:** Replace with a $\langle \text{NUM} \rangle$ token, but

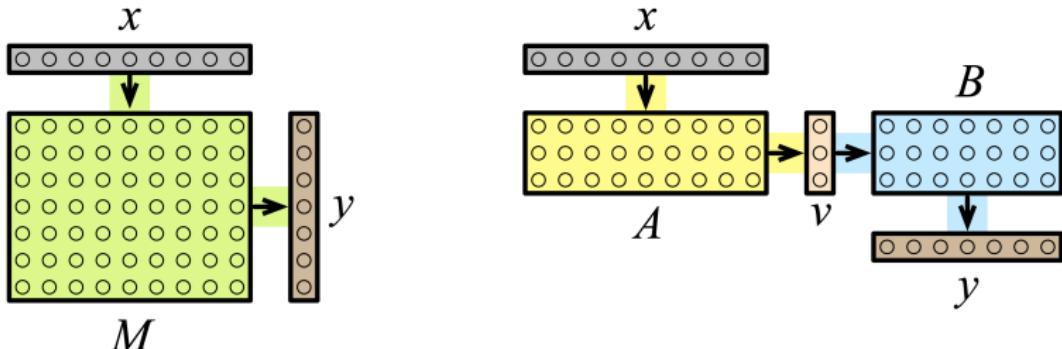
I pay $\langle \text{NUM} \rangle$ in May $\langle \text{NUM} \rangle$ = I pay $\langle \text{NUM} \rangle$ in May $\langle \text{NUM} \rangle$

- ▶ **Solution 2:** Replace each digit with a unique symbol, e.g. 5

I pay 555.55 in May 5555 > I pay 5555 in May 555.55

- ▶ This reduces the need for embeddings, when we can simply do transliteration

Factored Decomposition



- ▶ **Problem:** Large input and output vectors
 - ▶ $|x| = 20,000, |y| = 50,000 \rightarrow |M| = 1,000,000,000$
- ▶ **Solution:** Use a bottleneck with smaller matrices A, B
 - ▶ $|v| = 100 \rightarrow |A| = 2,000,000, |B| = 5,000,000$
 - ▶ Total Parameters: 7,000,000

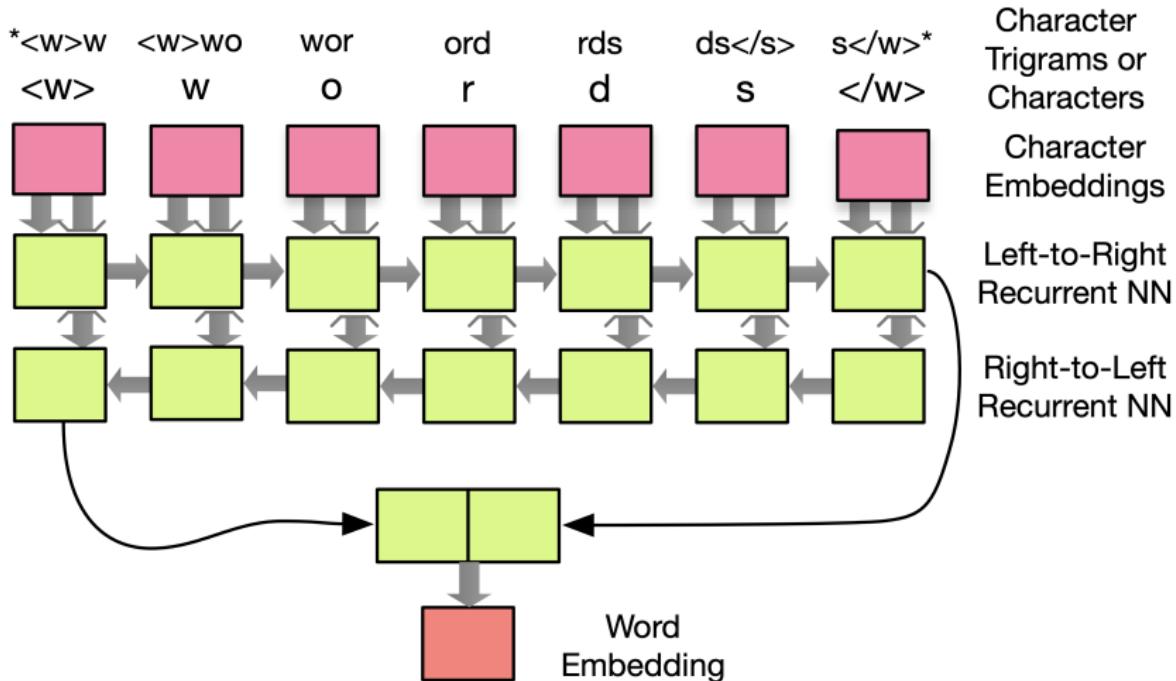
Character-Based Models

- ▶ Instead use embeddings for character string
`b e a u t i f u l`
- ▶ Idea is to induce embeddings for unseen morphological variants:
`beautiful`
- ▶ Tokens are single characters, symbols, whitespace

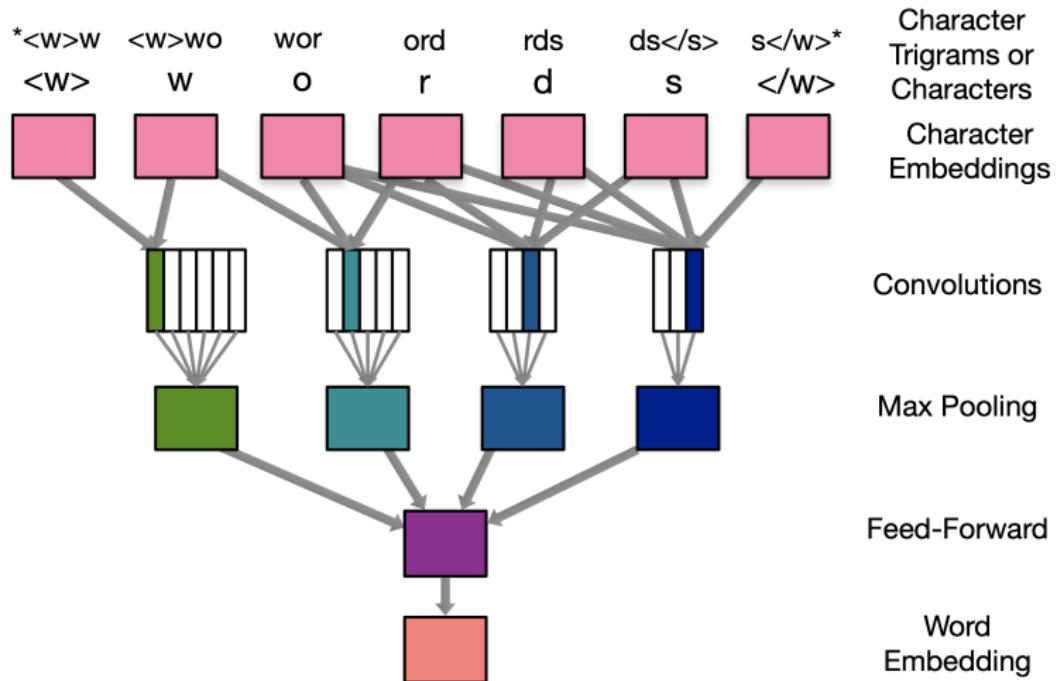
Character-Based Models

- ▶ Instead use embeddings for character string
`b e a u t i f u l`
- ▶ Idea is to induce embeddings for unseen morphological variants:
`beautiful`
- ▶ Tokens are single characters, symbols, whitespace
- ▶ Generally poor performance

Character-Based Models



Character-Based Models



BPE Subwords

```
the · f a t · c a t · i s · i n · t h e · t h i n · b a g  
t h → th     the · f a t · c a t · i s · i n · t h e · t h i n · b a g  
a t → at     the · f a t · c a t · i s · i n · t h e · t h i n · b a g  
i n → in     the · f a t · c a t · i s · i n · t h e · t h i n · b a g  
t h e → the   the · f a t · c a t · i s · i n · t h e · t h i n · b a g
```

- ▶ Breaks words into subwords
 - ▶ Starts with the character set
 - ▶ Merges the most frequent pairs, one per iteration
- ▶ Unsupervised (accidental) morphology (frequency suffixes)

BPE Tokenization, The Great Gatsby

- ▶ Unique Characters: 39
- ▶ Unique BPE Tokens (1K iterations): 801
- ▶ Unique BPE Tokens (5K iterations): 3,371
- ▶ Unique Tokens: 5,856

100 s@@ o w@@ e be@@ at on , b@@ o@@ at@@ s a@@ g@@ a@@ in@@ s@@ t
the c@@ ur@@ r@@ en@@ t@@ ,
b@@ or@@ n@@ e b@@ ac@@ k c@@ e@@ as@@ el@@ es@@ s@@ ly
in@@ to the p@@ a@@ st@@ .

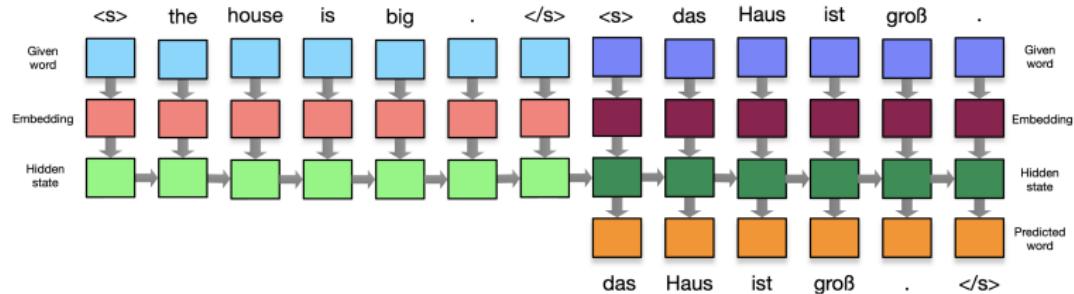
1K so we be@@ at on, bo@@ at@@ s against the c@@ ur@@ r@@ ent@@ ,
bor@@ ne@@ e back c@@ e@@ as@@ el@@ es@@ s@@ ly into the pa@@ st@@ .

10K so we beat on , bo@@ ats against the curr@@ ent ,
bor@@ ne back ceaselessly into the past .

50K so we beat on , boats against the current ,
borne back ceaselessly into the past .

Column Segmentation Architecture

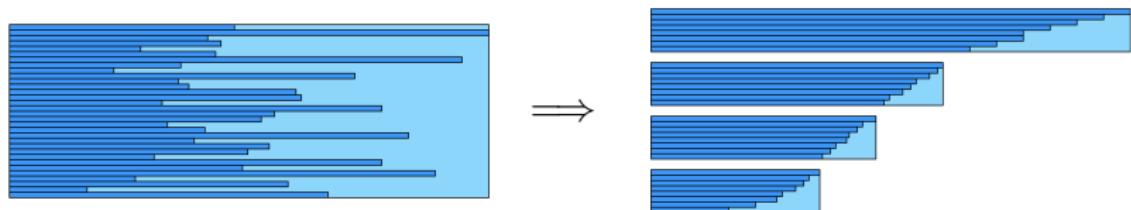
Seq2Seq Architecture



```
Seq2Seq(  
    encoder: Encoder(  
        embedding: Embedding(21, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
    )  
    decoder: Decoder(  
        embedding: Embedding(22, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
        (fc): Linear(in_features=128, out_features=22, bias=True)  
    )  
)  
The model has 508,438 trainable parameters
```

Training Considerations

Increasing Throughput through Batching



- ▶ We can pad sentences of different lengths to increase batch size
- ▶ While also minimizing the use of padding
- ▶ Matrix Operations are faster
- ▶ **Warning!** Make sure padding is not influencing the hidden state

Teacher Forcing

- ▶ Instead of refeeding the predicted token, replace it with the true token randomly
- ▶ This is only done during training, not decoding

$$y_{i+1} = \begin{cases} \operatorname{argmax}_j \theta_i & \mathcal{U}(0, 1) > \text{TF} \\ t_{i+1} & \text{else} \end{cases}$$

- ▶ t_{i+1} is the true token
- ▶ TF is the teacher forcing ratio

Cross Entropy and Label Smoothing

$$\begin{aligned}\ell(\mathbf{x}, y_i) &= -\log \left(\frac{\exp x_{y_i}}{\sum_j \exp x_j} \right) \\ &= -\underbrace{x_{y_i}}_{\max} + \log \underbrace{\sum_j \exp x_j}_{\min}\end{aligned}$$

- ▶ Softmax and Cross-Entropy loss assign all the probability mass to a single word
 - ▶ LogSumExp is minimized on confident predictions
- ▶ Solution: **smooth** the distribution

Quick Maths: LogSumExp Bounds

$$\begin{aligned}\max \{x_1, \dots, x_n\} &= \log (\exp(\max x_i)) \\ &\leq \log (\exp(x_1) + \dots + \exp(x_n)) \\ &\leq \log (n \cdot \exp(\max x_i)) \\ &= \max \{x_1, \dots, x_n\} + \log(n).\end{aligned}$$

- ▶ The lower bound is *approached* when **all but one** of the arguments approach $-\infty$
- ▶ The upper bound is met when **all** the arguments are equal (uniform)
- ▶ Hence LogSumExp is minimized on a confident prediction

Cross Entropy and Label Smoothing

- ▶ Softmax

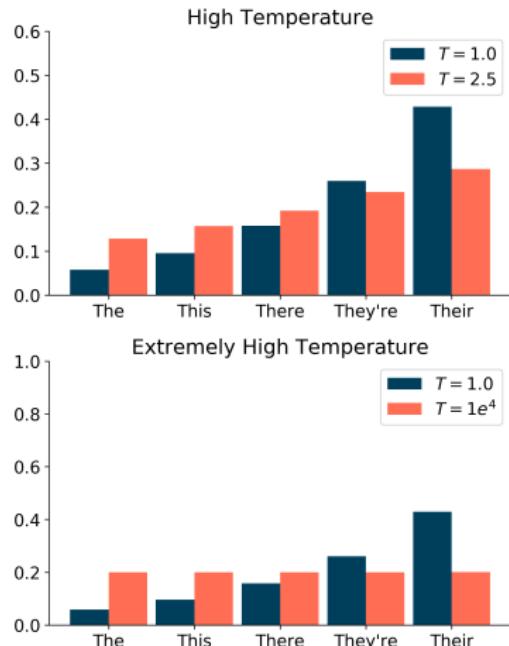
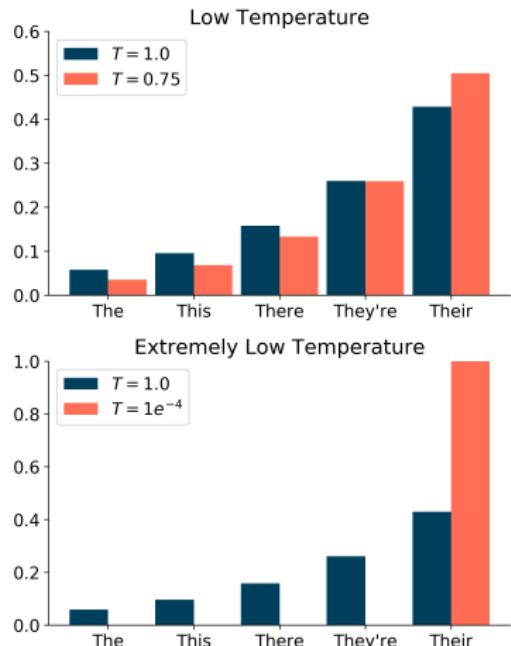
$$p(y_i) = \frac{\exp x_{y_i}}{\sum_j \exp x_j}$$

- ▶ Smoothed Softmax with Temperature T

$$p(y_i) = \frac{\exp (x_{y_i}/T)}{\sum_j \exp (x_j/T)}$$

- ▶ As $T \rightarrow \infty$, the distribution is smoother, uniform
- ▶ As $T \rightarrow 0$, the distribution approaches a kronecker delta centered on the class with the most mass
- ▶ **Question:** Why do we divide instead of adding/subtracting?

Visualizing Temperature



Monte Carlo Decoding

- ▶ Recall how we select the next token:
 - ▶ **Greedy:** Top token weight
 - ▶ **Teacher Forcing:** Randomly select the true token
- ▶ Note that the outputs are a distribution over the target vocabulary
- ▶ **Use these weights in a multinomial to randomly select a continuation**

$$y_{i+1} \sim \text{Multinomial}(\theta_i)$$

Different Token Decoding Schemes

- ▶ Greedy:

$$y_{i+1} = \operatorname{argmax}_j \theta_i$$

- ▶ Teacher Forcing:

$$y_{i+1} = \begin{cases} \operatorname{argmax}_j \theta_i & \mathcal{U}(0, 1) > \text{TF} \\ t_{i+1} & \text{else} \end{cases}$$

- ▶ Monte Carlo:

$$y_{i+1} \sim \text{Multinomial}(\theta_i)$$

- ▶ θ_i are the output weights from the Decoder
- ▶ t_{i+1} is the true token at position $i + 1$
- ▶ TF is the teacher forcing ratio

Masked Loss

- ▶ Remember, we don't care what gets predicted after seeing a $\langle \text{EOS} \rangle$
- ▶ Hence, we need to mask out the loss for predicted tokens associated with $\langle \text{PAD} \rangle$

pred	\rightarrow	<i>le</i>	<i>le</i>	<i>chat</i>	<i>chat</i>	<i>chat</i>	<i>chat</i>
		\downarrow	\downarrow	\downarrow	\downarrow	\downarrow	\downarrow
		ℓ	ℓ	ℓ	ℓ	ℓ	0
		\uparrow	\uparrow	\uparrow	\uparrow	\uparrow	\uparrow
true	\rightarrow	<i>le</i>	<i>chat</i>	<i>est</i>	<i>noir</i>	$\langle \text{EOS} \rangle$	$\langle \text{PAD} \rangle$

- ▶ **Solution:** Zero out elements by either:
 - ▶ Multiply pad outputs by 0
 - ▶ Specify the label to ignore in Cross Entropy call

$$\ell(\mathbf{x}, y_i) = \mathbb{1}_{\{y_i \neq \langle \text{PAD} \rangle\}} \cdot \left(-x_{y_i} + \log \sum_j \exp x_j \right)$$

Automatic Evaluation: BLEU

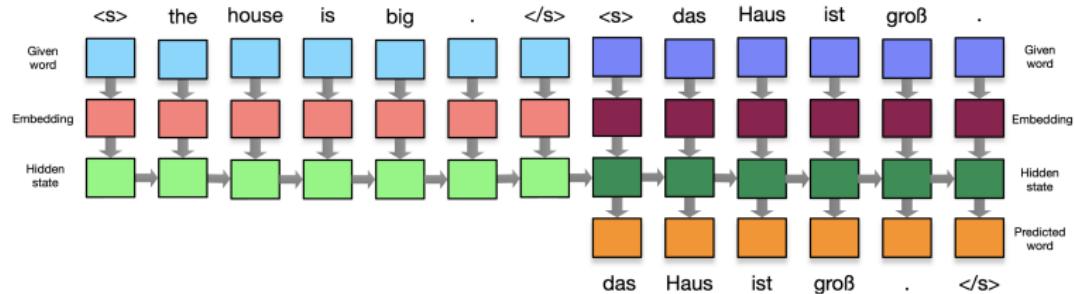
- ▶ Bilingual Evaluation Understudy
- ▶ For a candidate and reference translation:
 - ▶ Compare n -gram precision of sizes 1 to 4
 - ▶ Brevity penalty for short translations

Source	У меня есть навыки управления временем моркови	
Reference	I have the time management skills of a carrot	1.00
A	You got the scheduling skills of a banana	0.28
B	I have the time to plant lettuce	0.39
C	I have time management skills for a carrot	0.35

- ▶ Key: 4-gram, Trigram, Bigram, Unigram, No Match

Initial Training Results

Recall: Seq2Seq Architecture



```
Seq2Seq(  
    encoder: Encoder(  
        embedding: Embedding(21, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
    )  
    decoder: Decoder(  
        embedding: Embedding(22, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
        (fc): Linear(in_features=128, out_features=22, bias=True)  
    )  
)  
The model has 508,438 trainable parameters
```

Seq2Seq: Sample Results

```
> num num num punct
= num <SEG> num <SEG> num <SEG> punct
< num <SEG> num <SEG> num <SEG> punct
```

Seq2Seq: Sample Results

```
> num num num punct
= num <SEG> num <SEG> num <SEG> punct
< num <SEG> num <SEG> num <SEG> punct

> adj num num num
= adj <SEG> num <SEG> num <SEG> num
< adj <SEG> num <SEG> num <SEG> num num num num num
```

Seq2Seq: Sample Results

> num num num punct

= num <SEG> num <SEG> num <SEG> punct

< num <SEG> num <SEG> num <SEG> punct

> adj num num num

= adj <SEG> num <SEG> num <SEG> num

< adj <SEG> num <SEG> num <SEG> num num num num num

> verb noun noun noun noun noun verb

= verb <SEG> noun noun <SEG> noun noun noun <SEG> verb

< verb noun <SEG> <SEG> noun noun noun noun <SEG> noun ...

Translation Results: TL;DR

- ▶ Reasonably for small sequences
- ▶ Struggles to terminate properly
 - ▶ Maybe more training?
- ▶ Long Sequences are a mess

Translation Results: TL;DR

- ▶ Reasonably for small sequences
- ▶ Struggles to terminate properly
 - ▶ Maybe more training?
- ▶ Long Sequences are a mess
- ▶ **However** this indicates that enough signal can be transferred via the hidden state

Review

- ▶ We have covered:
 - ▶ What are Seq2Seq Models?
 - ▶ How can I reduce my vocabulary size?
 - ▶ How do I train a Seq2Seq model?
- ▶ Part 2 will explore more advanced concepts such as:
 - ▶ Decoding with Beam Search
 - ▶ Modeling Recurrent Relations
 - ▶ Other applications of Sequence Modeling

Thank You!

Seq2Seq in Action: Column Segmentation

Part 2

Bill Watson

S&P Global

January 30, 2020

Quick Review: Table Extraction

Schedule 1

CITY OF COFFEYPVILLE, KANSAS
 Summary of Revenues, Expenditures, and Budgets
 (Indirect Funds Only)
 Regulatory Basis

For the Year Ended December 31, 2017

Funds	Total	Adjustments for Qualifying Budgetary Activities		Budget for Current Period	Expenditures Charged to Current Year	Variances - Over Budget
		Certified	Budgeted			
General Fund	\$ 13,032,133.00	\$ 13,032,133.00	\$ 13,032,133.00	\$ 13,032,133.00	\$ 13,032,133.00	\$ 11,150,000.00*
Special Purpose Funds:						
Library	390,713.00	-	1,352,713.00	390,713.00	365,950,113	(24,760,87)
Local Activities	130,278.00	-	80,278.00	130,278.00	98,445,265	(66,166,97)
Police Department	51,274.00	-	22,274.00	51,274.00	38,445,101	(15,170,93)
Airport Building	82,274.00	-	82,274.00	82,274.00	81,817,930	(45,263,11)
Golf Course	333,313.00	-	333,313.00	333,313.00	299,260,211	(34,349,79)
Youth Activities Center	34,158.00	-	4,158.00	34,158.00	27,668,514	(6,531,46)
Assaults On Other Operations	128,961.00	-	128,961.00	128,961.00	101,216,477	(27,754,53)
Business Development Training Center	135,919.00	-	135,919.00	135,919.00	135,919.00	(5,318)
Veterans Memorial Stadium	25,897.00	9,335.72	35,232.72	25,897.00	20,003,712	(5,397,004)
Other Special Funds	1,387,366.00	-	1,387,366.00	1,387,366.00	1,042,291,341	(340,055,621)
City of Coffeyville Tax	518,265.00	-	518,265.00	518,265.00	518,265.00	-
1/2 Cent Sales Tax BBR Data Service	277,195.00	-	277,195.00	277,195.00	277,195.00	(17,465,00)
Community Improvement District Sales Tax						
Gas and Oil Production						
Gas and Oil Tax						
Business Parks						
Electric Utility	60,092,640.00	-	60,092,640.00	56,101,000,617	11,991,547,231	
Water and Sewer Utility	5,146,333.00	-	5,146,333.00	5,146,333.00	5,146,333.00	(694,94)
Water and Sewer Data Service	1,161,729.00	-	1,161,729.00	1,161,729.00	1,161,729.00	(231,899,48)
Stormwater Utility	481,991.00	-	481,991.00	481,991.00	463,849,313	(18,141,67)
Refuse/Trash Utility	606,026.00	-	606,026.00	606,026.00	538,268,1	(69,467,79)
Internet Utility	279,466.00	144.60	279,466.00	279,466.00	271,734,463	(7,734,53)

- 22 -

Quick Review: Table Extraction

Schedule 1

CITY OF COFFEYPVILLE, KANSAS
 Summary of Expenditures, Capital and Budget
 (Planned Funds Only) and Budget
 Regulatory Basis

For the Year Ended December 31, 2017

Funds	Total			Adjustments for Qualifying Budgeted Contributions			Budgeted Disbursements Charged to Current Year			Variances - Over Budget		
	Total	Certified	Budgeted	Contributions	Budgeted	Contributed	Budgeted	Contributed	Budgeted	Contributed	Budgeted	Contributed
General Fund	\$ 12,952,133.00	\$ 12,952,133.00	\$ 12,952,133.00	\$ 40,500.00	\$ 29,713.00	\$ 29,713.00	\$ 365,950.13	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93
Special Purpose Funds:												
Library	310,278.00	-	-	1,352,716.00	800,000.00	800,000.00	384,445.26	665,126.72	665,126.72	665,126.72	665,126.72	665,126.72
Local Alcohol License	1,140,000.00	-	-	22,000.00	22,000.00	22,000.00	1,000.00	1,000.00	1,000.00	1,000.00	1,000.00	1,000.00
Airport Building	82,274.00	-	-	80,274.00	80,274.00	80,274.00	81,817.90	81,817.90	81,817.90	81,817.90	81,817.90	81,817.90
Golf Course	333,713.00	-	-	333,713.00	-	-	299,260.21	27,249.79	27,249.79	27,249.79	27,249.79	27,249.79
Youth Activities Center	34,158.00	-	-	4,158.00	4,158.00	4,158.00	27,000.00	7,000.00	7,000.00	7,000.00	7,000.00	7,000.00
Assets in Other Operations	128,961.00	-	-	128,961.00	-	-	101,608.40	16,531.46	16,531.46	16,531.46	16,531.46	16,531.46
Business Development Training Center	135,919.00	-	-	135,919.00	-	-	135,919.00	135,919.00	135,919.00	135,919.00	135,919.00	135,919.00
Veterans Memorial Stadium	25,897.00	-	-	25,897.00	-	-	20,000.00	5,897.00	5,897.00	5,897.00	5,897.00	5,897.00
Other	1,827,466.00	-	-	1,827,466.00	-	-	850,000.00	850,000.00	850,000.00	850,000.00	850,000.00	850,000.00
City of Coffeyville Tax ICBR Debt Service	518,265.00	-	-	518,265.00	-	-	518,265.00	-	-	-	-	-
Community Improvement District Sales Tax	1,232,000.00	-	-	1,232,000.00	-	-	-	-	-	-	-	-
Gas & Electric Utility	1,140,000.00	-	-	1,140,000.00	-	-	-	-	-	-	-	-
Business Parks	60,092,640.00	-	-	60,092,640.00	-	-	56,101,000.00	6,091,640.00	6,091,640.00	6,091,640.00	6,091,640.00	6,091,640.00
Electric Utility	5,146,333.00	-	-	5,146,333.00	-	-	5,146,333.00	5,146,333.00	5,146,333.00	5,146,333.00	5,146,333.00	5,146,333.00
Gasoline Fuel Service	1,161,720.00	-	-	1,161,720.00	-	-	1,161,720.00	1,161,720.00	1,161,720.00	1,161,720.00	1,161,720.00	1,161,720.00
Water and Sewer Data Service	481,991.00	-	-	481,991.00	-	-	465,849.33	16,141.67	16,141.67	16,141.67	16,141.67	16,141.67
Stormwater Utility	606,020.00	-	-	606,020.00	-	-	558,206.81	99,467.19	99,467.19	99,467.19	99,467.19	99,467.19
Refuse/Tire Utility	279,466.00	-	-	279,466.00	-	-	271,374.45	8,091.55	8,091.55	8,091.55	8,091.55	8,091.55
Incentive Colony	144,60	-	-	144,60	-	-	-	-	-	-	-	-
1/2 Curb Sale Tax ICBR Debt Service	1,232,000.00	-	-	1,232,000.00	-	-	-	-	-	-	-	-
Community Improvement District Sales Tax	277,193.00	-	-	277,193.00	-	-	-	-	-	-	-	-
Gas & Electric Utility	60,092,640.00	-	-	60,092,640.00	-	-	-	-	-	-	-	-
Business Parks	5,146,333.00	-	-	5,146,333.00	-	-	-	-	-	-	-	-
Electric Utility	1,161,720.00	-	-	1,161,720.00	-	-	-	-	-	-	-	-
Gasoline Fuel Service	481,991.00	-	-	481,991.00	-	-	-	-	-	-	-	-
Water and Sewer Data Service	606,020.00	-	-	606,020.00	-	-	-	-	-	-	-	-
Stormwater Utility	279,466.00	-	-	279,466.00	-	-	-	-	-	-	-	-
Refuse/Tire Utility	8,091.55	-	-	8,091.55	-	-	-	-	-	-	-	-
Incentive Colony	144,60	-	-	144,60	-	-	-	-	-	-	-	-

- 22 -



Schedule 1

CITY OF COFFEYPVILLE, KANSAS
 Summary of Expenditures, Capital and Budget
 (Planned Funds Only) and Budget
 Regulatory Basis

For the Year Ended December 31, 2017

Funds	Total			Adjustments for Qualifying Budgeted Contributions			Budgeted Disbursements Charged to Current Year			Variances - Over Budget		
	Total	Certified	Budgeted	Contributions	Budgeted	Contributed	Budgeted	Contributed	Budgeted	Contributed	Budgeted	Contributed
General Fund	\$ 12,952,133.00	\$ 12,952,133.00	\$ 12,952,133.00	\$ 40,500.00	\$ 29,713.00	\$ 29,713.00	\$ 365,950.13	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93	\$ 11,150,837.93
Special Purpose Funds:												
Library	310,278.00	-	-	1,352,716.00	800,000.00	800,000.00	384,445.26	665,126.72	665,126.72	665,126.72	665,126.72	665,126.72
Local Alcohol License	1,140,000.00	-	-	22,000.00	22,000.00	22,000.00	1,000.00	1,000.00	1,000.00	1,000.00	1,000.00	1,000.00
Airport Building	82,274.00	-	-	80,274.00	80,274.00	80,274.00	81,817.90	81,817.90	81,817.90	81,817.90	81,817.90	81,817.90
Golf Course	333,713.00	-	-	333,713.00	-	-	299,260.21	27,249.79	27,249.79	27,249.79	27,249.79	27,249.79
Youth Activities Center	34,158.00	-	-	34,158.00	-	-	30,351.30	3,196.50	3,196.50	3,196.50	3,196.50	3,196.50
Assets in Other Operations	128,961.00	-	-	128,961.00	-	-	101,608.40	16,531.46	16,531.46	16,531.46	16,531.46	16,531.46
Business Development Training Center	135,919.00	-	-	135,919.00	-	-	135,919.00	135,919.00	135,919.00	135,919.00	135,919.00	135,919.00
Veterans Memorial Stadium	25,897.00	-	-	25,897.00	-	-	20,000.00	5,897.00	5,897.00	5,897.00	5,897.00	5,897.00
Other	1,827,466.00	-	-	1,827,466.00	-	-	850,000.00	850,000.00	850,000.00	850,000.00	850,000.00	850,000.00
City of Coffeyville Tax ICBR Debt Service	518,265.00	-	-	518,265.00	-	-	518,265.00	-	-	-	-	-
Community Improvement District Sales Tax	1,232,000.00	-	-	1,232,000.00	-	-	-	-	-	-	-	-
Gas & Electric Utility	1,140,000.00	-	-	1,140,000.00	-	-	-	-	-	-	-	-
Business Parks	60,092,640.00	-	-	60,092,640.00	-	-	-	-	-	-	-	-
Electric Utility	5,146,333.00	-	-	5,146,333.00	-	-	-	-	-	-	-	-
Gasoline Fuel Service	1,161,720.00	-	-	1,161,720.00	-	-	-	-	-	-	-	-
Water and Sewer Data Service	481,991.00	-	-	481,991.00	-	-	-	-	-	-	-	-
Stormwater Utility	606,020.00	-	-	606,020.00	-	-	-	-	-	-	-	-
Refuse/Tire Utility	279,466.00	-	-	279,466.00	-	-	-	-	-	-	-	-
Incentive Colony	8,091.55	-	-	8,091.55	-	-	-	-	-	-	-	-
1/2 Curb Sale Tax ICBR Debt Service	1,232,000.00	-	-	1,232,000.00	-	-	-	-	-	-	-	-
Community Improvement District Sales Tax	277,193.00	-	-	277,193.00	-	-	-	-	-	-	-	-
Gas & Electric Utility	60,092,640.00	-	-	60,092,640.00	-	-	-	-	-	-	-	-
Business Parks	5,146,333.00	-	-	5,146,333.00	-	-	-	-	-	-	-	-
Electric Utility	1,161,720.00	-	-	1,161,720.00	-	-	-	-	-	-	-	-
Gasoline Fuel Service	481,991.00	-	-	481,991.00	-	-	-	-	-	-	-	-
Water and Sewer Data Service	606,020.00	-	-	606,020.00	-	-	-	-	-	-	-	-
Stormwater Utility	279,466.00	-	-	279,466.00	-	-	-	-	-	-	-	-
Refuse/Tire Utility	8,091.55	-	-	8,091.55	-	-	-	-	-	-	-	-
Incentive Colony	144,60	-	-	144,60	-	-	-	-	-	-	-	-

- 22 -



Quick Review: Table Extraction

Schedule 1

CITY OF COFFEETOWN, KANSAS
Summary of Revenues and Expenses
(Produced Pursuant Only)
Regulatory Basis

For the Year Ended December 31, 2017

Funds	Total	Adjustments Are Qualifying		Budget for Current Year	Requisitions Charged to Current Year	Variance - Over Budget (Under)
		Overfunded	Underfunded			
General Fund	\$ 12,622,13,00			29,671,13,00	26,932,13,13	(\$ 1,639,97,87)
Special Purpose Funds				1,38,278,00	44,46,12,00	(\$ 94,21,90)
Literacy	299,713,00	-		8,02,714,00	8,1,81,90	(\$ 1,19,79,70)
Local Auto League	140,278,00	-		33,513,00	27,42,60	(\$ 6,09,40)
Airport Building	82,274,00	-		29,26,21,00	13,4,29,70	(\$ 16,97,20)
Golf Course	203,513,00	-		4,12,00	4,6,53,40	(\$ 5,41,40)
Youth Activities Center	34,158,00	-		12,0,00	10,7,24,00	(\$ 1,74,80)
Health	128,061,00	-		13,0,00	10,21,47	(\$ 2,74,53)
Business Development Training Center	135,919,00	-		15,9,19,00	13,9,19,00	(\$ 1,10,10)
Veterans Memorial Stadium	25,897,00	9,235,72		32,32,72	9,03,27,72	(\$ 26,19,20)
Other	1,272,00	1,272,00		1,272,00	1,26,87,72	(\$ 1,10,23,28)
CIP Fund	51,8,265,00	-		86,8,265,00	51,8,265,00	(\$ 35,00,00)
1/2 Cent Sales Tax 2018 Debt Service	51,8,265,00	-		51,8,265,00	51,8,265,00	-
Community Improvement District & Other Tax	-	-		-	-	-
Gas and Liquid Public	277,145,00	-		277,145,00	277,145,00	(\$ 17,665,00)
Zoning Board of Adjustment	60,079,240,00	-		56,131,029,67	56,131,029,67	(\$ 1,991,347,33)
Electric Utility	5,136,333,00	-		5,136,333,00	5,136,333,00	(\$ 54,54)
Facility Debt Service	1,16,279,00	3,272,42		1,16,279,00	1,16,279,00	(\$ 21,11)
Water and Sewer Debt Service	481,991,00	-		481,991,00	481,991,00	(\$ 99,648)
Stormwater Utility	606,026,00	-		606,026,00	558,266,81	(\$ 47,765,19)
Benton/Twent Utility	297,046,00	14,436,00		297,046,00	271,36,43	(\$ 46,679,54)
Balancem CCR	-	-		-	-	-

- 22 -

Quick Review: Table Extraction

CITY OF COFFEYVILLE, KANSAS
 Summary of Financials
 (Unaudited - Partial Only) and Budget
 Recreational Basis

For the Year Ended December 31, 2017

Funds	Total	Adjustments for Qualifying Budget Changes		Budget for Current Year	Reappropriations Changed to Current Year	Variance - Over (Under)
		Overfunded	Underfunded			
General Fund	\$ 12,582,113.00			\$ 11,593,817.00	\$ 11,593,817.00	\$ (1,532,300.00) (99)
Special Purpose Funds	290,713.00			290,713.00	290,713.00	
Library	50,078.00			50,078.00	50,078.00	
Local Auto League	50,078.00			50,078.00	50,078.00	
Airport Building	82,274.00			82,274.00	82,274.00	
Golf Course	203,313.00			203,313.00	203,313.00	
Youth Activities Center	34,158.00			34,158.00	34,158.00	
Health	40,000.00			40,000.00	40,000.00	
Business Development Training Center	128,001.00			128,001.00	128,001.00	
Veterans Memorial Stadium	155,919.00			155,919.00	155,919.00	
Other	25,897.00			25,897.00	25,897.00	
CIP	1,272,266.00			1,272,266.00	1,272,266.00	
CIP - Capital Outlay Tax	518,265.00			518,265.00	518,265.00	
1/2 Cent Sales Tax SBH Debt Service						
Community Improvement District & River Tax						
Gas and Liquid Public						
Zoo/Brewery/Park Assessment						
Electric Utility						
Facilities Dept Services						
Police Dept Services						
Fire and EMS Dept Service						
Water and Sewer Dept Service						
Stormwater Utility						
Bullock/Tread Utility						
Bullock/City						
	297,145.00			297,145.00	297,145.00	
	60,079,240.00			60,079,240.00	60,079,240.00	
	5,196,333.00			5,196,333.00	5,196,333.00	
	1,162,798.00			1,162,798.00	1,162,798.00	
	481,991.00			481,991.00	481,991.00	
	606,026.00			606,026.00	606,026.00	
	297,046.00			297,046.00	297,046.00	
	14,450.00			14,450.00	14,450.00	

- 22 - →

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text	
5	1	1	1	1	1	1092	100	62	27	95	For	
5	1	1	1	1	1	21165	100	55	26	95	the	
5	1	1	1	1	3	1239	100	82	25	96	Year	
5	1	1	1	1	4	1335	100	114	24	96	Ended	
5	1	1	1	1	5	1463	100	82	24	93	June	
5	1	1	1	1	7	1630	100	88	22	36	ZU18	
5	1	1	1	1	2	11295	146	219	43	95	(Continued)	
5	1	1	1	1	3	1	460	263	128	32	96	NOTE
5	1	1	1	1	2	605	263	16	31	90		
5	1	1	1	3	3	637	280	13	5	87		
5	1	1	1	3	4	665	260	245	34	96	SUMMARY	
5	1	1	1	3	5	925	260	61	32	95	OF	
5	1	1	1	3	6	1002	257	303	34	96	SIGNIFICANT	
5	1	1	1	3	7	1319	254	318	34	95	ACCOUNTING	
5	1	1	1	3	8	1651	252	220	34	95	POLICIES	
5	1	1	1	3	9	1886	251	232	42	96	(Continued)	
5	1	1	1	4	1	537	372	39	31	93	H.	
5	1	1	1	4	2	612	369	206	34	95	Inventories	
5	1	1	1	4	3	833	368	64	33	96	and	
5	1	1	1	4	4	911	367	141	43	96	Prepaid	
5	1	1	1	4	5	1096	368	98	31	95	Items	
5	1	1	1	5	1	611	477	207	34	96	Inventories	
5	1	1	1	5	2	831	487	54	22	96	are	
5	1	1	1	5	3	896	476	123	32	95	valued	
5	1	1	1	5	4	1031	480	32	27	96	at	
5	1	1	1	5	5	1073	475	55	32	96	the	
5	1	1	1	5	6	1141	474	101	33	95	lower	
5	1	1	1	5	7	1253	473	41	33	96	of	
5	1	1	1	5	8	1301	477	73	28	93	cost	
5	1	1	1	5	9	1385	471	147	42	90	(first-in,	
5	1	1	1	5	10	1545	470	163	41	96	(first-out)	

Quick Review: Table Extraction

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
5	1	1	1	1	1	1092	100	62	27	95	For
5	1	1	1	1	2	1165	100	55	26	95	the
5	1	1	1	1	3	1239	100	82	25	95	Year
5	1	1	1	1	4	1335	100	114	24	96	Ended
5	1	1	1	1	5	1463	100	82	24	93	June
5	1	1	1	1	7	1630	100	88	22	36	ZU18
5	1	1	1	2	1	1295	146	219	43	96	(Continued)
5	1	1	1	3	1	460	263	128	32	96	NOTE
5	1	1	1	3	2	605	263	16	31	90	
5	1	1	1	3	3	637	280	13	5	87	-
5	1	1	1	3	4	665	260	245	34	96	SUMMARY
5	1	1	1	3	5	925	260	61	33	95	OF
5	1	1	1	3	6	1002	257	303	34	96	SIGNIFICANT
5	1	1	1	3	7	1319	254	318	34	95	ACCOUNTING
5	1	1	1	3	8	1651	255	220	34	95	POLICIES
5	1	1	1	3	9	1886	251	232	42	96	(Continued)
5	1	1	1	4	1	537	372	39	31	93	H.
5	1	1	1	4	2	612	369	206	34	96	Inventories
5	1	1	1	4	3	833	368	64	33	96	and
5	1	1	1	4	4	911	367	141	43	96	Prepaid
5	1	1	1	4	5	1056	368	98	31	96	Items
5	1	1	1	5	1	611	477	207	34	96	Inventories
5	1	1	1	5	2	831	487	54	22	96	are
5	1	1	1	5	3	896	476	123	32	96	valued
5	1	1	1	5	4	1031	480	32	27	95	at
5	1	1	1	5	5	1073	475	55	32	96	the
5	1	1	1	5	6	1141	474	101	33	96	lower
5	1	1	1	5	7	1253	473	41	33	96	of
5	1	1	1	5	8	1301	477	73	28	93	cost
5	1	1	1	5	9	1385	471	147	42	90	(first-in,
5	1	1	1	5	10	1545	470	163	41	96	first-out)

Quick Review: Table Extraction

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
5	1	1	1	1	1	1092	100	62	27	95	For
5	1	1	1	1	2	1165	100	55	26	95	the
5	1	1	1	1	3	1239	100	62	25	95	Year
5	1	1	1	1	4	1335	100	114	24	95	Ended
5	1	1	1	1	5	1463	100	82	24	93	June
5	1	1	1	1	7	1630	100	88	22	36	ZU18
5	1	1	1	2	1	1295	146	219	43	95	(Continued)
5	1	1	1	3	1	460	263	128	32	95	NOTE
5	1	1	1	3	2	605	266	16	31	90	
5	1	1	1	3	3	637	280	13	5	87	-
5	1	1	1	3	4	665	260	245	34	95	SUMMARY
5	1	1	1	3	5	925	260	61	36	95	OF
5	1	1	1	3	6	1002	257	303	34	95	SIGNIFICANT
5	1	1	1	3	7	1319	254	318	34	95	ACCOUNTING
5	1	1	1	3	8	1651	255	220	34	95	POLICIES
5	1	1	1	3	9	1886	251	232	42	95	(Continued)
5	1	1	1	4	1	537	372	39	31	93	H.
5	1	1	1	4	2	612	369	206	34	95	Inventories
5	1	1	1	4	3	833	368	64	33	95	and
5	1	1	1	4	4	911	367	141	43	95	Prepaid
5	1	1	1	4	5	1056	368	98	31	95	Items
5	1	1	1	5	1	611	477	207	34	95	Inventories
5	1	1	1	5	2	831	487	54	22	95	are
5	1	1	1	5	3	896	476	123	32	95	valued
5	1	1	1	5	4	1031	480	32	27	95	at
5	1	1	1	5	5	1073	475	55	32	95	the
5	1	1	1	5	6	1141	474	101	33	95	lower
5	1	1	1	5	7	1253	473	41	35	95	of
5	1	1	1	5	8	1301	477	73	28	93	cost
5	1	1	1	5	9	1385	471	147	42	90	(first-in,
5	1	1	1	5	10	1545	470	163	41	95	first-out)



Receipts	2018 - 2019 General Fund - (BOGCS - w/o)			Variance - Over (Under)
	Prior Year	Actual	Budget	
	Actual			
Taxes and Shared Receipts				
Ad Valorem Property Tax	\$1,779,884.26	\$1,672,543.39	\$1,225,362	\$ -552,818.61
Delinquent Tax	107,336.00	83,439.02	65,000	25,439.02
Motor Vehicle Tax	185,104.64	198,703.03	206,851	-7,107.97
Recreational Vehicle Tax	1,372.3	1,796.98	1,521	255.68
1/2/20 M Vehicle Tax	872.7	566.64	841	-124.36
Vehicle Rental Excise Tax	98.1	51.82	177	-125.18
Commercial Vehicle Tax	9,690.34	9,861.33	8,107	1,754.00
Watercraft Tax	657.49	520.7	517	15.7
Special Assessments	8,032.47	7,934.54	30,000	-22,067.54
Permit Tax	654,481.17	686,420.05	475,932	\$11,481.00
Entire Tax	6,121,208.67	4,610,514.73	4,462,516	\$95,001.98
Federal Grants	8,600.01			
Local Alcohol Liquor Tax	10,091.99	28,879.71	15,327	12,551.71
Special Highway Tax	250,225.61	256,024.36	256,330	-205.64
Highway Connecting Links	76,750.80	76,446.19	74,700	-54.61
Highway County Aid	40,894.05	40,875.73	40,830	514.73
Licenses and Permits	94,105.8	92,656	92,656	
Flows, Forfeitures and Penalties	182,460.84	207,414.75	308,445	-101,030.25
Charges for Services	95,811.84	90,217.85	86,559	-5,279.05
Use of Money and Property				
Interest Income	20,461.28	21,263.49	12,000	9,263.49
Rents	67,423	34,480.5	67,400	-56,907.5
Sale of Equipment and Scrap	2,229.4	3,294.94	5,000	-1,705.06
Other Receipts				
Donations	-	1,180	-	1,180
Reimbursed Expenses	46,191.89	40,804.82	-	40,804.82
Insurance Proceeds	-	38,750	-	38,750
Interest	8,239.7	4,526.25	6,300	-1,473.75
Operating Transfers from:				
Electric Utility Fund	2,327,245.74	2,406,982.7	2,965,899	-179,912.3
Water and Sewer Utility Fund	694,534.64	777,054.1	723,775	53,278.1
Community Development Fund	216,605.33	-	-	-
Total Receipts	12,095,362.9	11,813,775.08	\$12,348,772	\$ -434,006.02
Expenditures			TT	
General Government				
Personal Services	836,080.11	782,088.89	\$ 987,380	\$ -209,280.01
Contractual Services	259,252.84	269,145.84	313,797	-44,651.00
Commodities	10,788.03	12,130.62	14,240	-2,058.38
Capital Outlay	1,442.98	1,832.82	16,450	-16,617.08
	40			

Outline Review

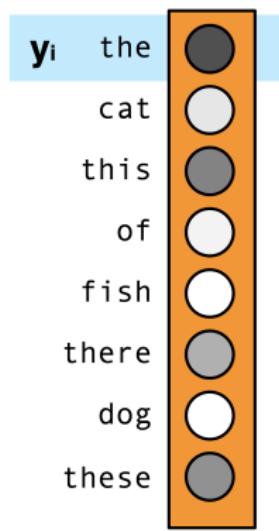
- ▶ We have covered:
 - ▶ What are Seq2Seq Models?
 - ▶ How can I reduce my vocabulary size?
 - ▶ How do I train a Seq2Seq model?
- ▶ Part 2:
 - ▶ Decoding with Beam Search
 - ▶ Modeling Recurrent Relations (RNNs)
 - ▶ Updating the Column Segmentation Model
 - ▶ Other applications of Sequence Modeling

Decoding: Making better Translations

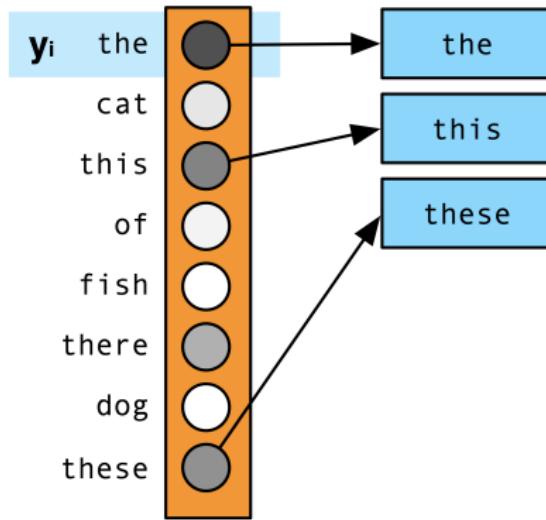
Beam Search

- ▶ Beam Search is a technique to explore different translation paths
- ▶ Based on the idea that you **shouldn't take the greedy path**
- ▶ And that the best path may be sub-optimal during early iterations

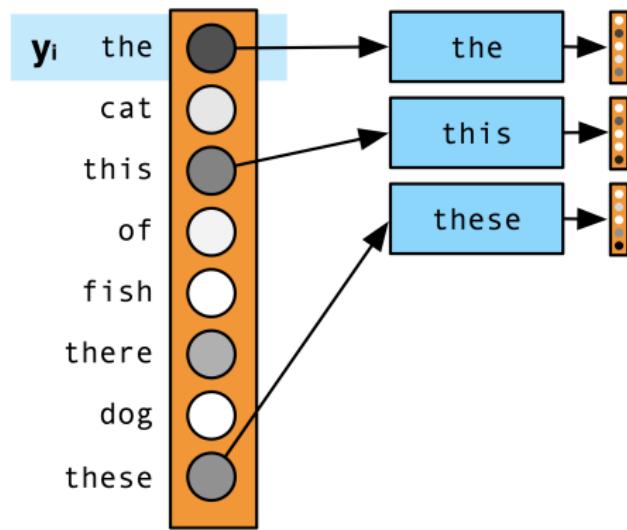
Beam Search



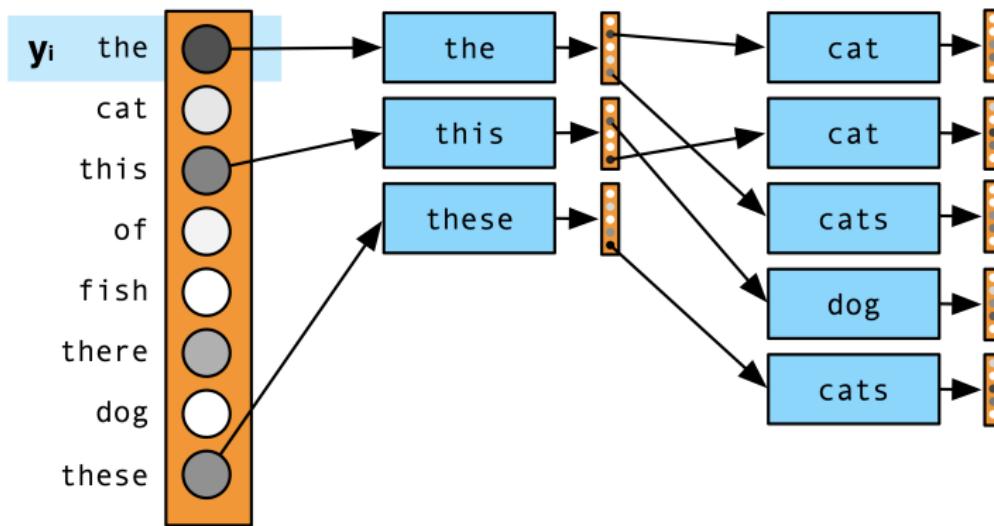
Beam Search



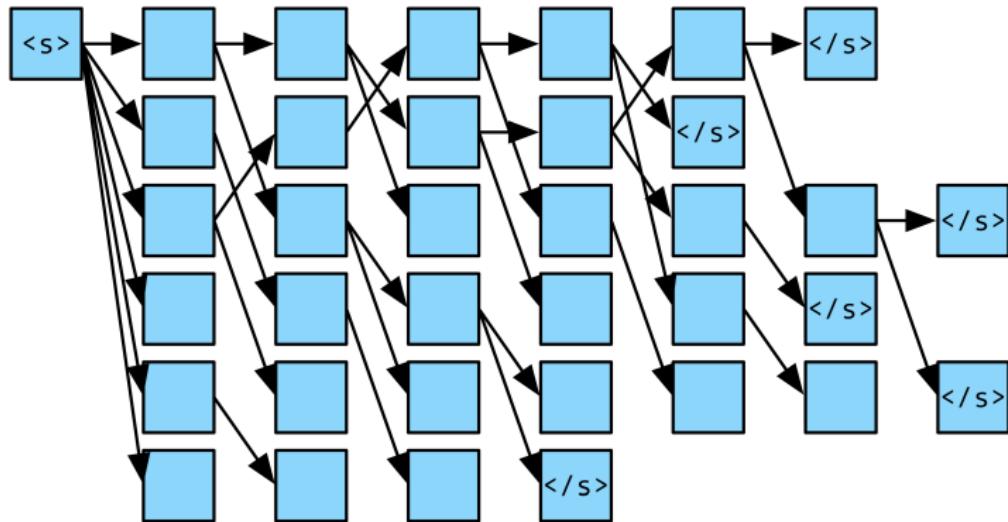
Beam Search



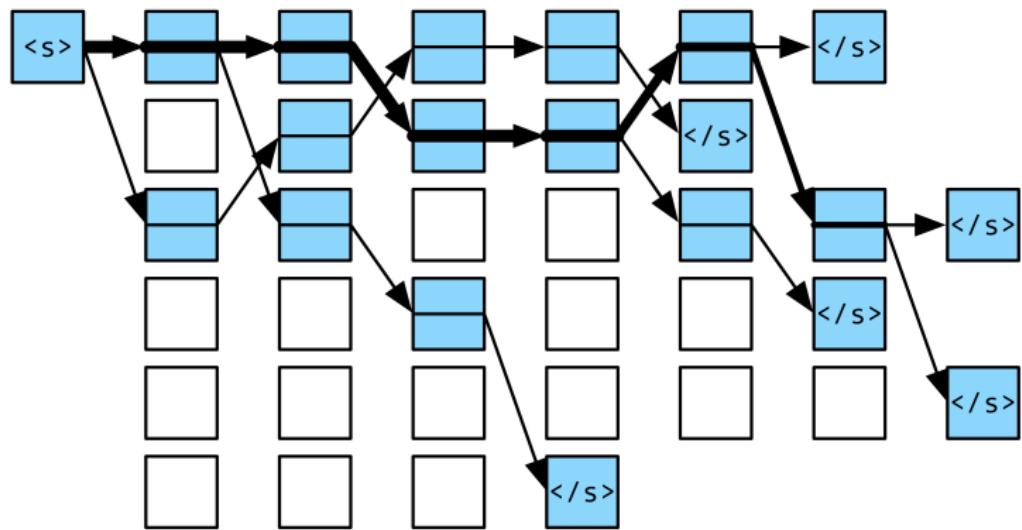
Beam Search



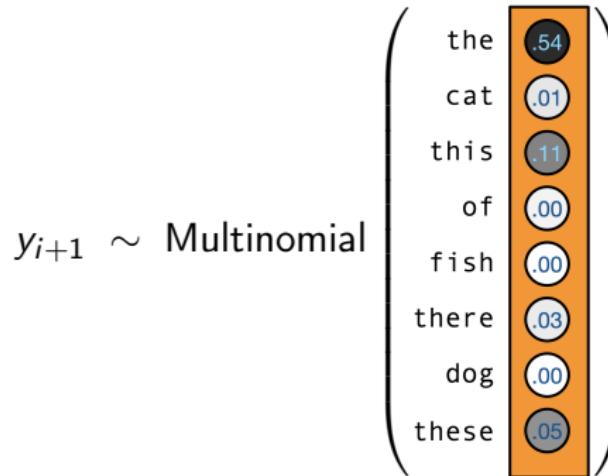
Beam Search



Beam Search

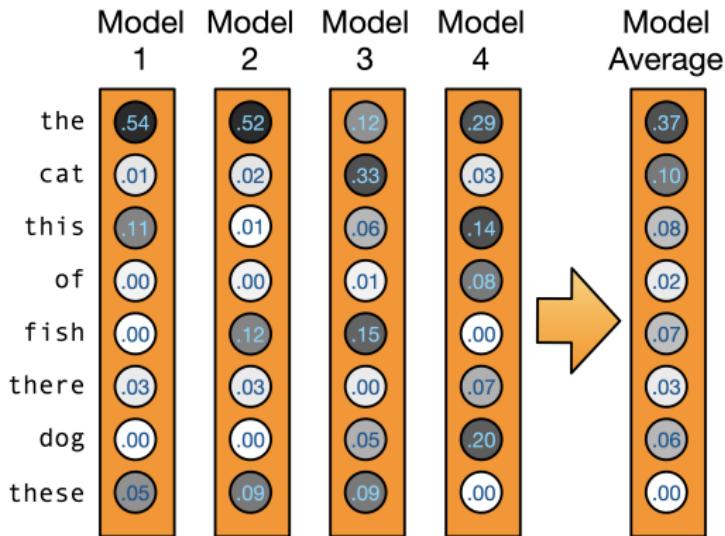


Monte Carlo Beam Search



- ▶ Why not sample n words based on their probabilities?
- ▶ Adds more diversity to beam search results

Ensembling



- ▶ Why not average different models?
- ▶ Random initialization leads to different local solutions
- ▶ Could also use model dumps from different iterations

Modeling Recurrent Relations

Vanilla RNNs

$$h_t = \tanh \left(\underbrace{W_{ih}x_t + b_{ih}}_{\text{input}} + \underbrace{W_{hh}h_{t-1} + b_{hh}}_{\text{hidden}} \right)$$

- ▶ h_t is the hidden state at time t
- ▶ x_t is the input at time t
- ▶ h_{t-1} is the previous hidden state
- ▶ h_0 is initialized to 0

Long Short Term Memory (LSTM)

$$\begin{aligned} \text{Gates} \rightarrow & \left\{ \begin{array}{l} i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \end{array} \right. \\ \text{Outputs} \rightarrow & \left\{ \begin{array}{l} c_t = f_t \odot c_{t-1} + i_t \odot g_t \\ h_t = o_t \odot \tanh(c_t) \end{array} \right. \end{aligned}$$

- ▶ h_t is the hidden state at time t
- ▶ c_t is the cell state at time t
- ▶ x_t is the input at time t

Gated Recurrent Units (GRU)

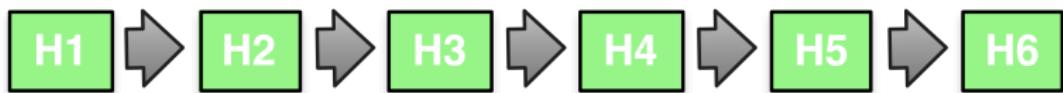
$$\text{Gates} \rightarrow \begin{cases} r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \\ z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \\ n_t = \tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{(t-1)} + b_{hn})) \end{cases}$$

$$\text{Outputs} \rightarrow \begin{cases} h_t = (1 - z_t) \odot n_t + z_t \odot h_{(t-1)} \end{cases}$$

- ▶ h_t is the hidden state at time t
- ▶ x_t is the input at time t

Aside: Different Perspectives on Deep Recurrent Models

- ▶ So far we've only seen Left to Right Sequencing



Aside: Different Perspectives on Deep Recurrent Models

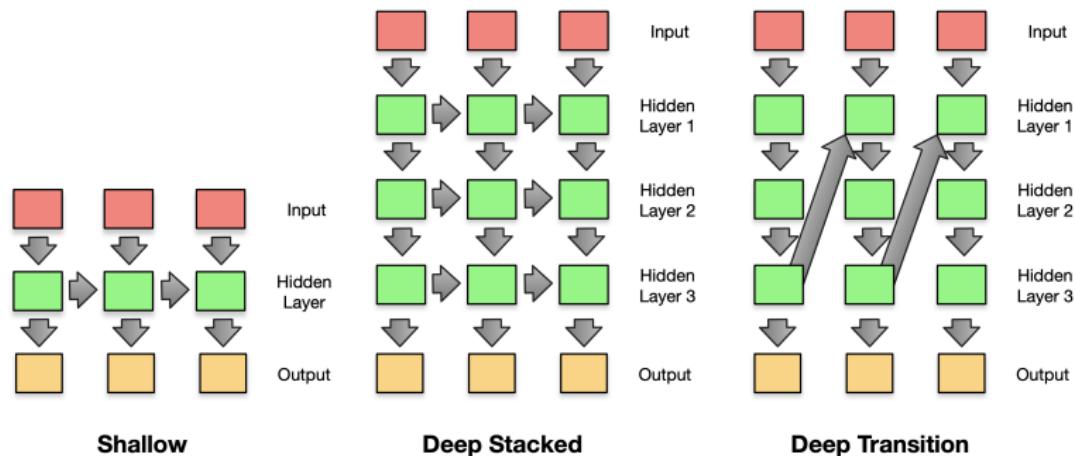
- ▶ So far we've only seen Left to Right Sequencing



- ▶ Why not Right to Left?

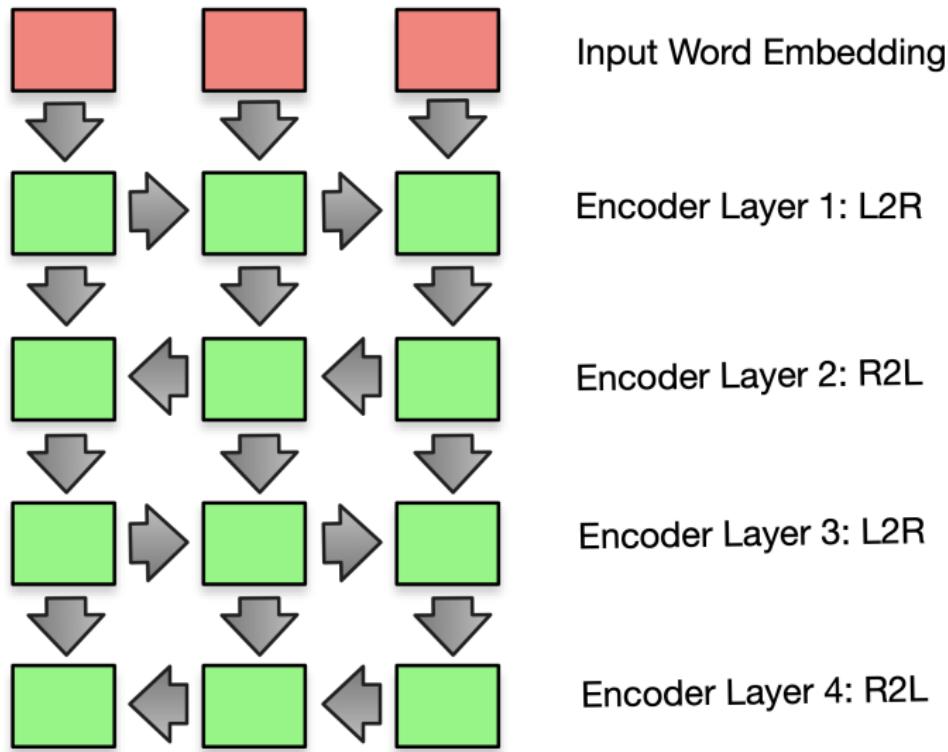


Aside: Different Perspectives on Deep Recurrent Models

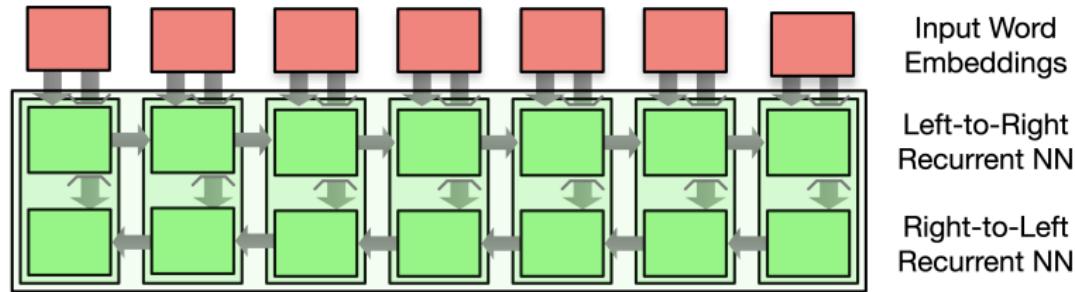


- ▶ Experiment with different stacking techniques

Alternating Recurrent Directions



Bidirectional Sequence Modeling



- ▶ Can capture both left and right context
- ▶ Implementation usually concatenates RNN states

Aside: Dimensionality of Inputs and Outputs (Batch First)

Type	RNN	LSTM	GRU
In	B, L, H_{in}	B, L, H_{in}	B, L, H_{in}
h_{t-1}	$N_L \cdot N_D, B, H_{out}$	$N_L \cdot N_D, B, H_{out}$	$N_L \cdot N_D, B, H_{out}$
c_{t-1}	-	$N_L \cdot N_D, B, H_{out}$	-
h_t	$N_L \cdot N_D, B, H_{out}$	$N_L \cdot N_D, B, H_{out}$	$N_L \cdot N_D, B, H_{out}$
c_t	-	$N_L \cdot N_D, B, H_{out}$	-
Out	$B, L, N_D \cdot H_{out}$	$B, L, N_D \cdot H_{out}$	$B, L, N_D \cdot H_{out}$

- ▶ B is the batch size
- ▶ L is the sequence length
- ▶ N_D is the number of directions
- ▶ N_L is the number of layers
- ▶ H_{in}, H_{out} are the input and hidden size

Aside: The Influence of Padding in RNNs

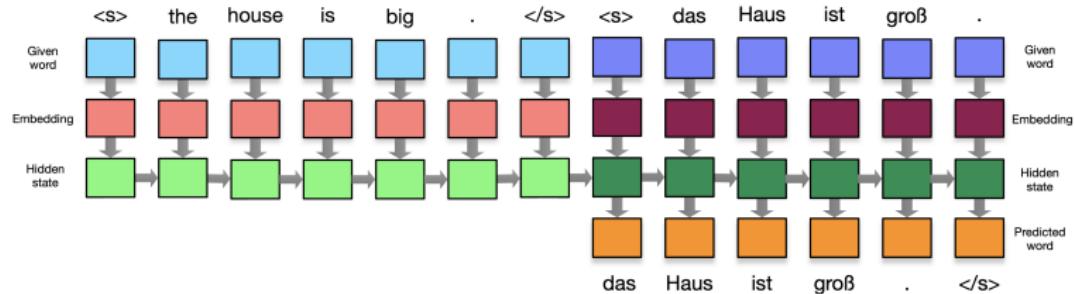
- ▶ Assume the embedding $E[\langle \text{PAD} \rangle] = \mathbf{0}$
- ▶ Are we safe?

Aside: The Influence of Padding in RNNs

- ▶ Assume the embedding $E[\langle \text{PAD} \rangle] = \mathbf{0}$
- ▶ Are we safe? NO!
- ▶ Because of the bias term, zero input does not result in a zero output
 - ▶ This alters the hidden state being passed onto the next iteration
- ▶ Don't even think about the mess bidirectionals become...
- ▶ **Question:** Does this mean we learn the amount of padding for a given sequence?

Applications to Column Segmentation

Applications to Column Segmentation



```
Seq2Seq(  
    (encoder): Encoder(  
        (embedding): Embedding(21, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
    )  
    (decoder): Decoder(  
        (embedding): Embedding(22, 256)  
        (gru): GRU(256, 128, num_layers=2, batch_first=True)  
        (fc): Linear(in_features=128, out_features=22, bias=True)  
    )  
)  
The model has 508,438 trainable parameters
```

Column Segmentation: Throw away the decoder

- ▶ Just use an encoder, with an output layer
- ▶ For every token, predict if we should insert a `\langle SEG \rangle`
- ▶ No longer have to learn the actual sequence

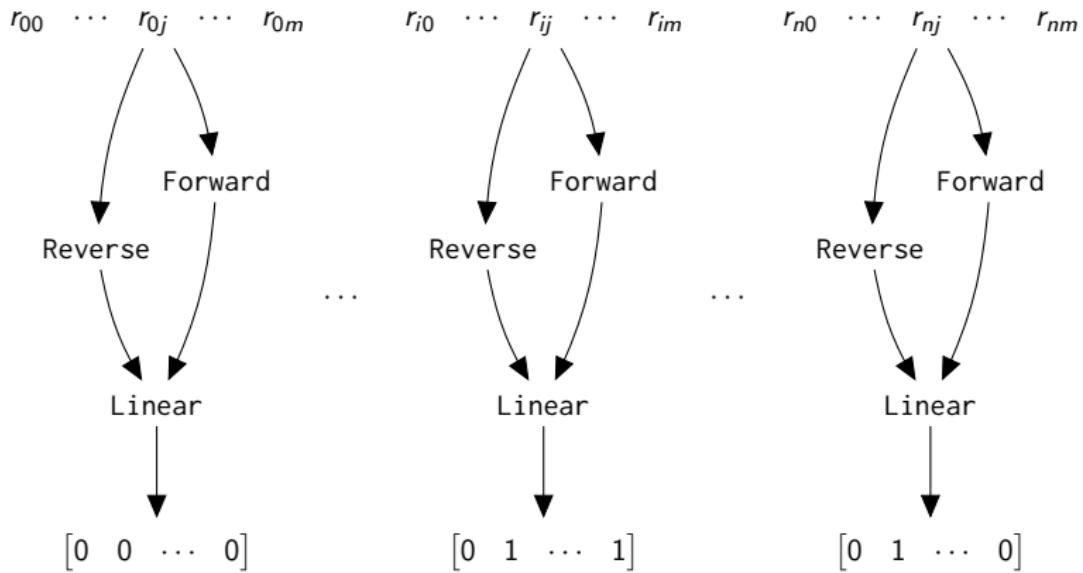
```
Classifier(  
    (encoder): Encoder(  
        (embedding): Embedding(21, 128)  
        (gru): GRU(128, 128, batch_first=True, bidirectional=True)  
    )  
    (linear): Linear(in_features=256, out_features=1, bias=True)  
)  
The model has 201,089 trainable parameters
```

Independent Row: Training Labels

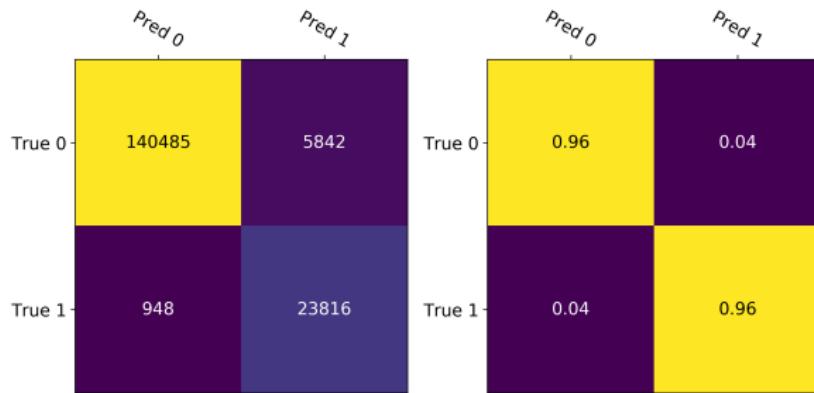
- ▶ Labels are now binary vectors, where $x_i = 1$ implies that there is a **(SEG)** between x_i and x_{i+1}

Less imputed interest	(174,862.64)		
0	0	1	0	0	0
Less imputed interest	(SEG)	(174,862.64)	

Independent Row Training Visual

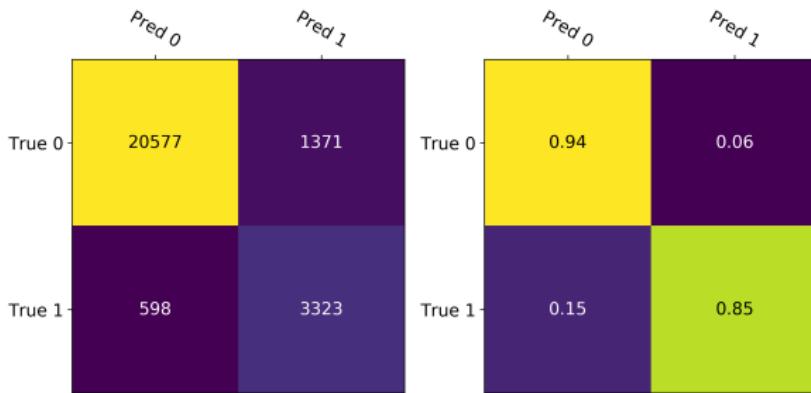


Independent Row: Sample Results (Training)



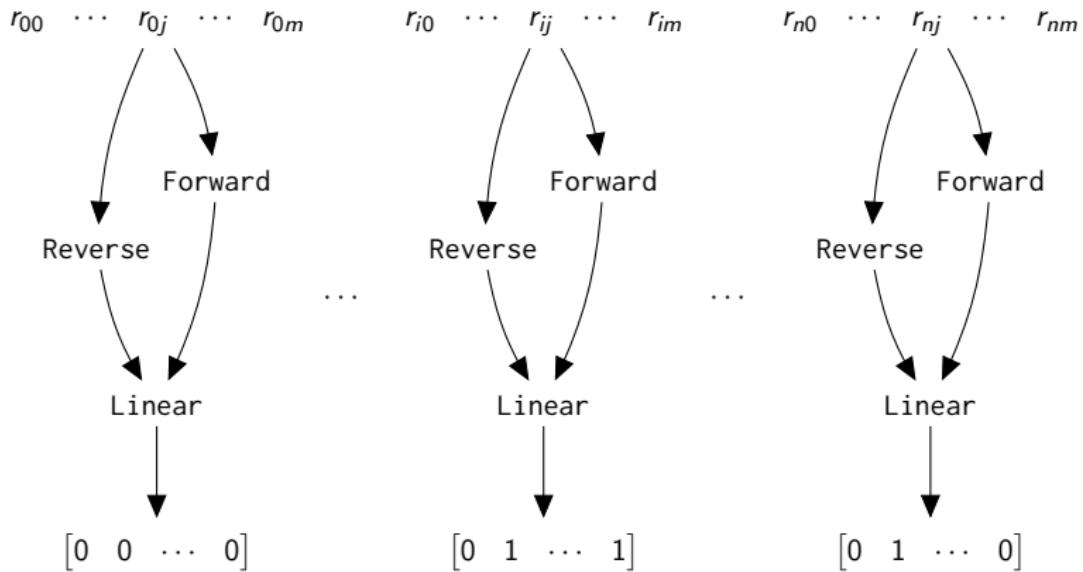
	Precision	Recall	F1-Score	Support
Label: 0	0.99	0.96	0.98	146,327
Label: 1	0.80	0.96	0.88	24,764
accuracy			0.96	171,091
macro avg	0.90	0.96	0.93	171,091
weighted avg	0.97	0.96	0.96	171,091

Independent Row: Sample Results (Validation)



	Precision	Recall	F1-Score	Support
Label: 0	0.97	0.94	0.95	21,948
Label: 1	0.71	0.85	0.77	3,921
accuracy			0.92	25,869
macro avg	0.84	0.89	0.86	25,869
weighted avg	0.93	0.92	0.93	25,869

How can we improve on this?



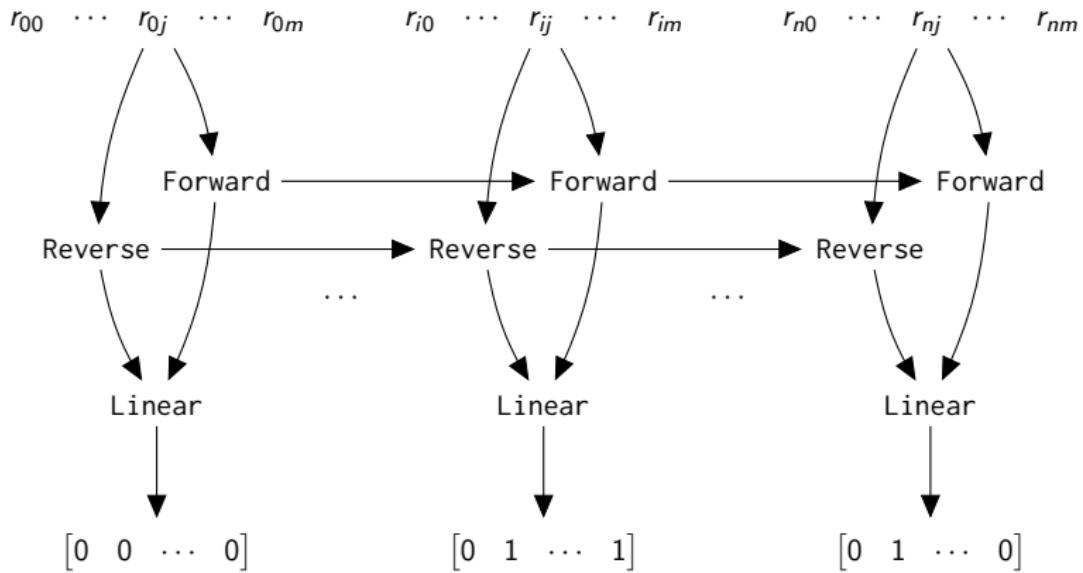
Joint Training on a Table

- ▶ We saw that we could regenerate a sequence from its hidden state alone.
- ▶ Why not use this information to link rows within the same table?
 - ▶ Bidirectional for each row
 - ▶ Propagate hidden state to next row
 - ▶ Linear layers for prediction

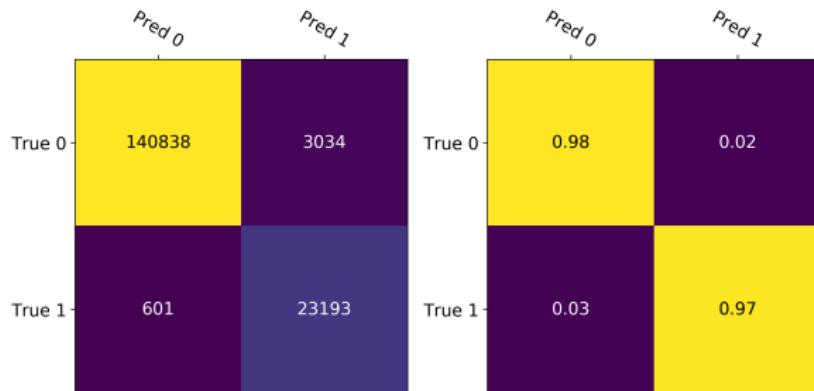
```
TableClassifier(  
    (embedding): Embedding(21, 128)  
    (rnn): GRU(128, 128, batch_first=True, bidirectional=True)  
    (ff): Linear(in_features=256, out_features=1, bias=True)  
)
```

The model has 201,089 trainable parameters

Joint Table Training Visual

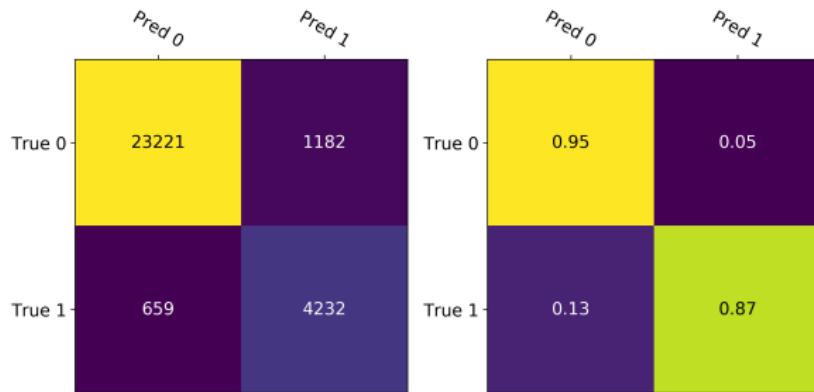


Joint Table Training: Sample Results (Training)



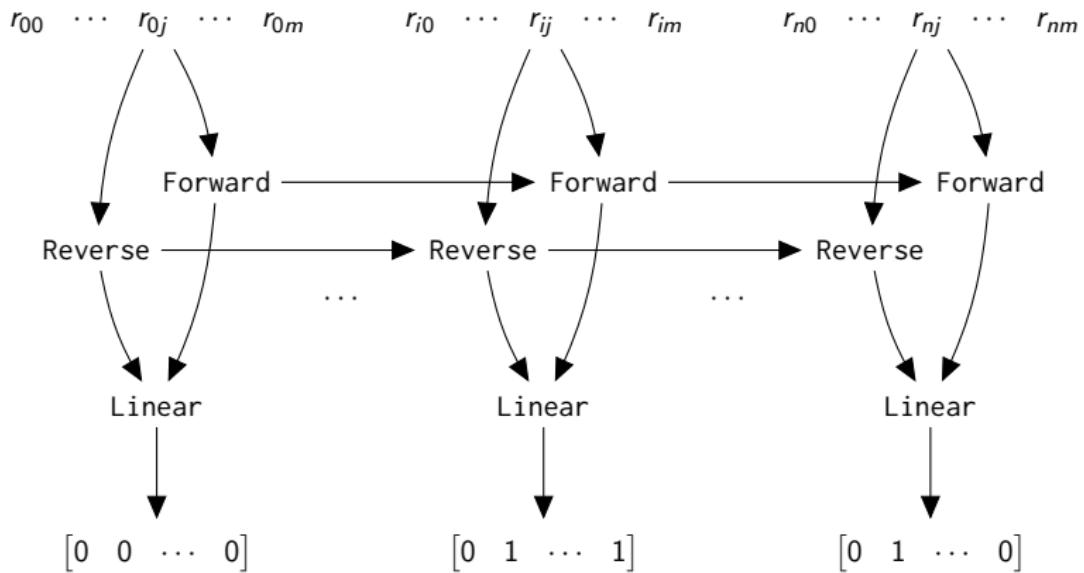
	Precision	Recall	F1-Score	Support
Label: 0	1.00	0.98	0.99	143,872
Label: 1	0.88	0.97	0.93	23,794
accuracy			0.98	167,666
macro avg	0.94	0.98	0.96	167,666
weighted avg	0.98	0.98	0.98	167,666

Joint Table Training: Sample Results (Validation)

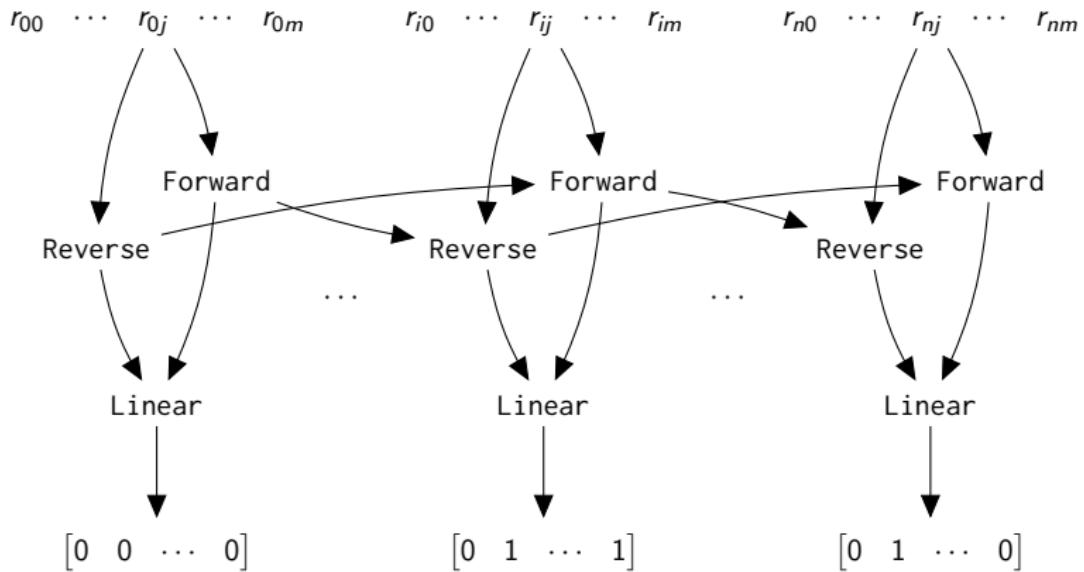


	Precision	Recall	F1-Score	Support
Label: 0	0.97	0.95	0.96	24,403
Label: 1	0.78	0.87	0.82	4,891
accuracy			0.94	29,294
macro avg	0.88	0.91	0.89	29,294
weighted avg	0.94	0.94	0.94	29,294

Is this a correct way to link rows?



Joint Table Training with Swapping Visual

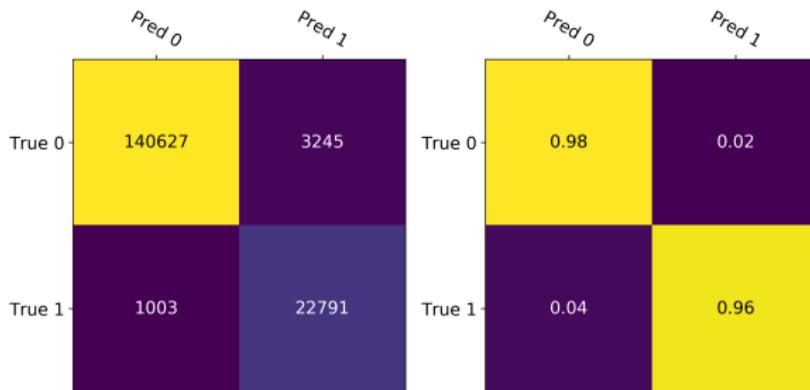


Trick for Swapping in Joint Table Training

- ▶ Note that the hidden state is biased towards the items closest to it
- ▶ So the beginning of a new row has context from the end of the last row
 - ▶ Is this a good choice?
 - ▶ What if we flip the bidirectional hidden states every iteration
 - ▶ AKA the beginning of a new row has the hidden state info from the beginning of the last row
 - ▶ And vice-versa

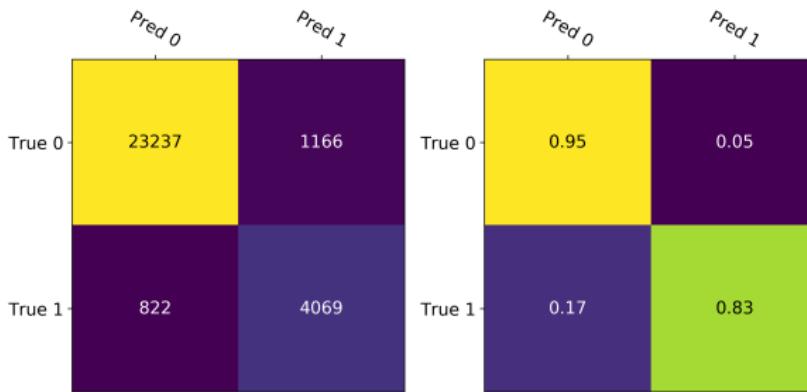
```
TableClassifierSwap(  
    (embedding): Embedding(21, 128)  
    (rnn): GRU(128, 128, batch_first=True, bidirectional=True)  
    (ff): Linear(in_features=256, out_features=1, bias=True)  
)  
The model has 201,089 trainable parameters
```

Joint Table Training with Swapping: Sample Results (Training)



	Precision	Recall	F1-Score	Support
Label: 0	0.99	0.98	0.99	143,872
Label: 1	0.88	0.96	0.91	23,794
accuracy			0.97	167,666
macro avg	0.93	0.97	0.95	167,666
weighted avg	0.98	0.97	0.98	167,666

Joint Table Training with Swapping: Sample Results (Validation)



	Precision	Recall	F1-Score	Support
Label: 0	0.97	0.95	0.96	24,403
Label: 1	0.78	0.83	0.80	4,891
accuracy			0.93	29,294
macro avg	0.87	0.89	0.88	29,294
weighted avg	0.93	0.93	0.93	29,294

Summary of Methods (Validation)

		Independent	Joint	Joint Swapped
Label: 0	Precision	0.97	0.97	0.97
	Recall	0.94	0.95	0.95
	F1-Score	0.95	0.96	0.96
Label: 1	Precision	0.71	0.78	0.78
	Recall	0.85	0.87	0.83
	F1-Score	0.77	0.82	0.80
Accuracy		0.92	0.94	0.93

What have we learned?

- ▶ We can learn text segments purely from the sequence of Part-of-Speech tags
- ▶ It is useful to jointly train/predict on the table instead of assuming independence
- ▶ Is there any benefit to swapping hidden states during propagation?
- ▶ **Note:**
 - ▶ We use the concept of *deep transition* in Seq2Seq for joint training
 - ▶ Where our training is dependent on the hidden state encoding useful information

Future Plans

- ▶ Features
 - ▶ Use word location information
 - ▶ Tesseract gives bounding boxes
- ▶ Encoder Structure
 - ▶ Convolution
 - ▶ Transformer

Other Applications For Sequence Modeling

Recall: Encoders, Decoders, and Seq2Seq Models

- ▶ **Encoders** given a sequence meaning
- ▶ **Decoders** generate a new sequence
- ▶ **Seq2Seq** generate sequences conditioned on another sequence

What to use for which application?

- ▶ **Encoders**
 - ▶ POS Tagging
 - ▶ Sentence Embeddings
 - ▶ Anything where you are given the sentence at test time
- ▶ **Decoders**
 - ▶ Text Generation
 - ▶ Language Modeling
 - ▶ Anything where you need to create a sequence at test time
- ▶ **Seq2Seq**
 - ▶ Translation
 - ▶ Speech Recognition
 - ▶ Summarization
 - ▶ Question/Answering
 - ▶ Anything where you convert a sequence into another sequence

Tools, References, and Further Reading

Acknowledgment

- ▶ Pieces adapted from my college Machine Translation Course by Philipp Koehn
- ▶ Additional acknowledgment:
 - ▶ Adam Lopez
 - ▶ Matt Post
 - ▶ Chris Callison-Burch
 - ▶ Philipp Koehn

Papers

- ▶ Sutskever et al., Sequence to Sequence Learning with Neural Networks
- ▶ Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
- ▶ Sennrich et al., Neural Machine Translation of Rare Words with Subword Units
- ▶ Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation
- ▶ Koehn, Neural Machine Translation
- ▶ Koehn, Six Challenges for Neural Machine Translation
- ▶ Curated Machine Translation Reading List

Tutorials

- ▶ Pytorch
 - ▶ Official PyTorch Seq2Seq Tutorial
 - ▶ PyTorch Seq2Seq with Torchtext
 - ▶ Ben Trevett Seq2Seq Tutorial
- ▶ Tensorflow
 - ▶ NMT with Attention

Libraries

- ▶ Facebook: fairseq (PyTorch)
- ▶ Open NMT (PyTorch)
- ▶ Open NMT (Tensorflow)

Thank You!