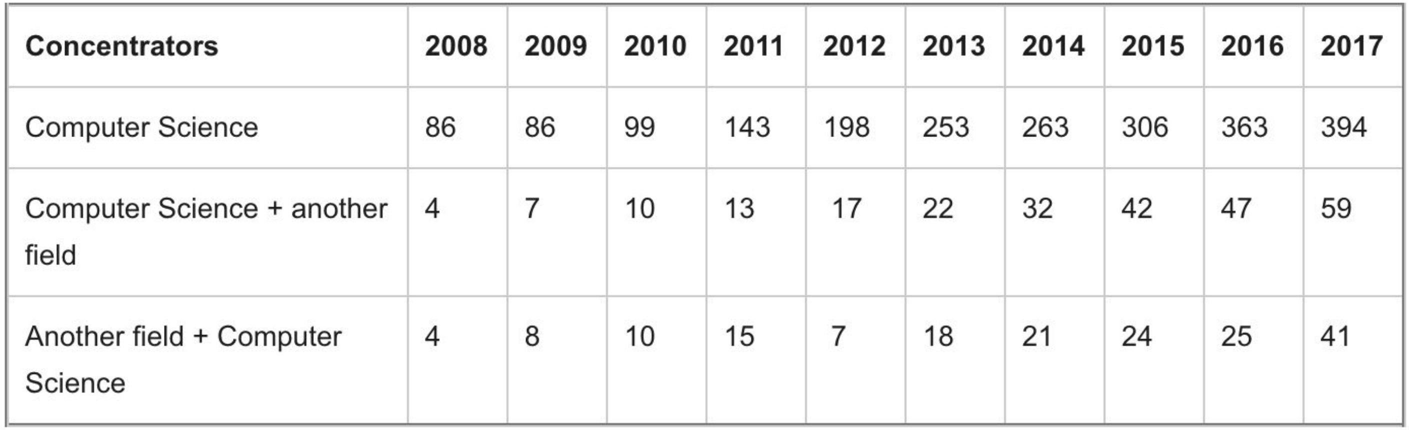
Long Story Short: Using BERT for abstractive text summarization on a small, curated corpus

Limor Gultchin, Pydata Berlin Meetup,
Oct 28, 2020
@gultchin

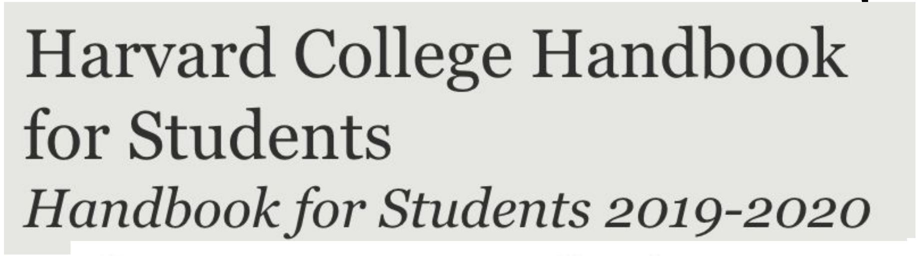
Problem definition

- Summer internship project for industry research team
- Clients have many PDF documents. Many of them contain tables. While software can extract tables from documents, devoid of context they make far less sense



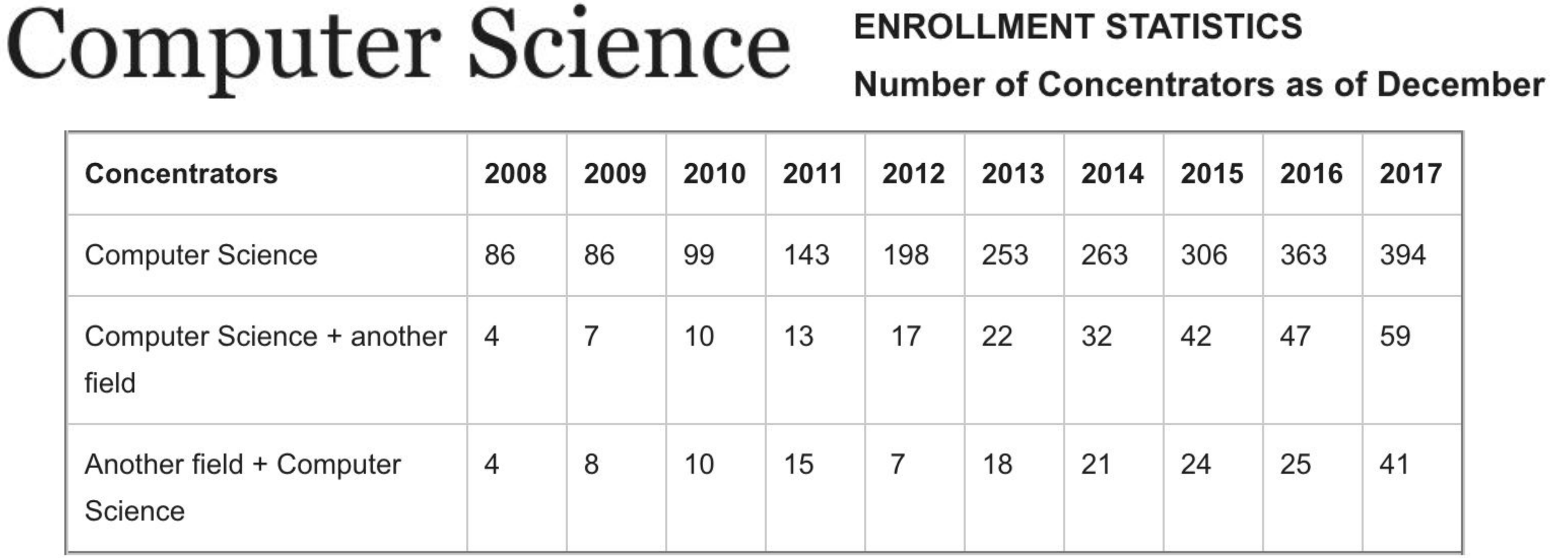
Concentrators	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Computer Science	86	86	99	143	198	253	263	306	363	394
Computer Science + another field	4	7	10	13	17	22	32	42	47	59
Another field + Computer Science	4	8	10	15	7	18	21	24	25	41

Intuitive solution: table + page titles



Harvard College Handbook for Students

Handbook for Students 2019-2020



Computer Science

ENROLLMENT STATISTICS

Number of Concentrators as of December

Concentrators	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Computer Science	86	86	99	143	198	253	263	306	363	394
Computer Science + another field	4	7	10	13	17	22	32	42	47	59
Another field + Computer Science	4	8	10	15	7	18	21	24	25	41

Case study: Scientific papers as table-title-generation prototype

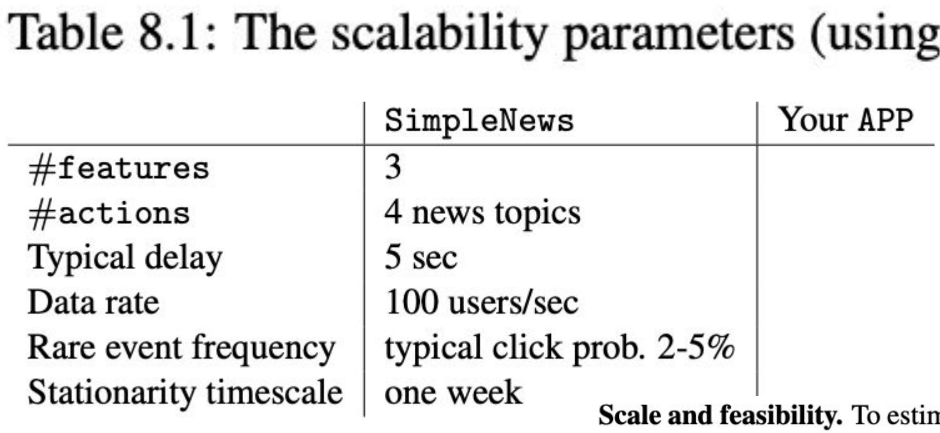
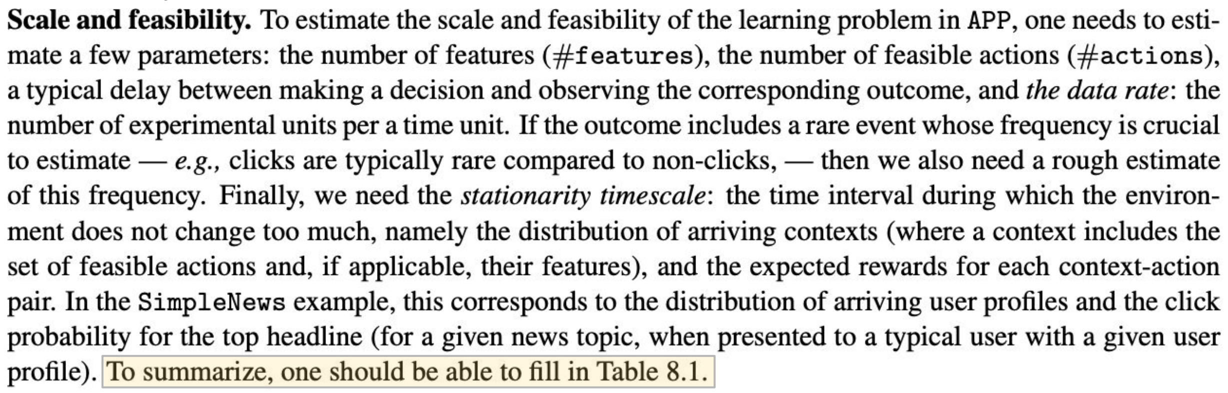


Table 8.1: The scalability parameters (using SimpleNews as an example)

	SimpleNews	Your APP
#features	3	
#actions	4 news topics	
Typical delay	5 sec	
Data rate	100 users/sec	
Rare event frequency	typical click prob. 2-5%	
Stationarity timescale	one week	

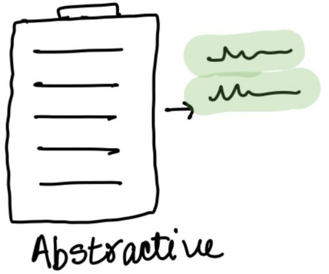


Scale and feasibility. To estimate the scale and feasibility of the learning problem in APP, one needs to estimate a few parameters: the number of features (#features), the number of feasible actions (#actions), a typical delay between making a decision and observing the corresponding outcome, and *the data rate*: the number of experimental units per a time unit. If the outcome includes a rare event whose frequency is crucial to estimate — *e.g.,* clicks are typically rare compared to non-clicks, — then we also need a rough estimate of this frequency. Finally, we need the *stationarity timescale*: the time interval during which the environment does not change too much, namely the distribution of arriving contexts (where a context includes the set of feasible actions and, if applicable, their features), and the expected rewards for each context-action pair. In the SimpleNews example, this corresponds to the distribution of arriving user profiles and the click probability for the top headline (for a given news topic, when presented to a typical user with a given user profile). To summarize, one should be able to fill in Table 8.1.

Titles are conceptually similar to summarization

- Text summarization is an area of much interest within the study of NLP
- Text summarization falls into two groups:
  - **Extractive** summarization
  - **Abstractive** summarization

Extractive



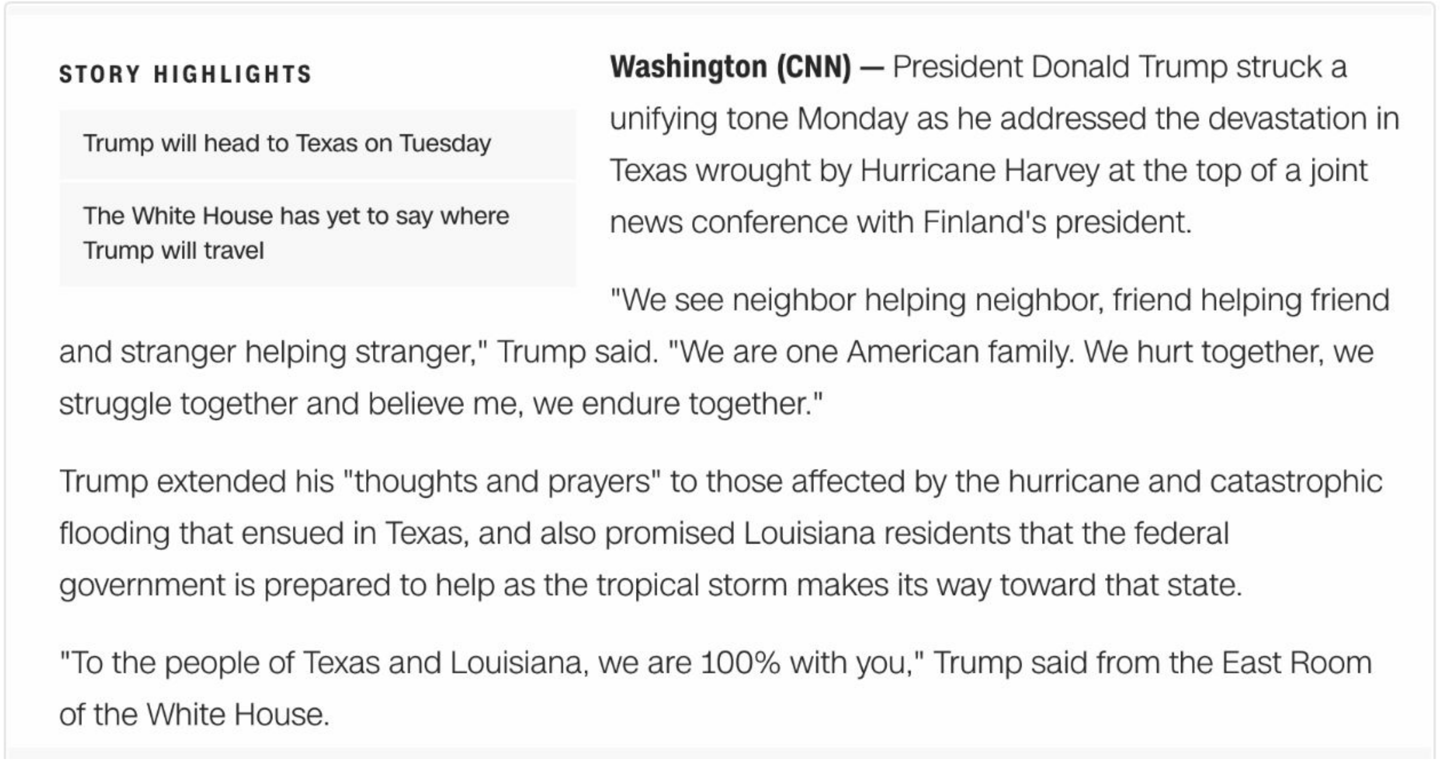
NLP approaches to summarization

Area of much research, **usually tested on corpuses of short news articles** with labeled **three-sentence summaries**.

- **Extractive:** Text-rank, Latent Semantic Analysis (LSA w\ SVD decomposition), Bayesian approaches, etc… Most based on frequency of appearance of word in sentence/document level.

- **Abstractive:** Seq-to-Seq, Pointer-Generator networks, Transformers (can be used for extractive as well, depends on optimization objective)

What usual text summarization datasets look like



STORY HIGHLIGHTS

Trump will head to Texas on Tuesday

The White House has yet to say where Trump will travel

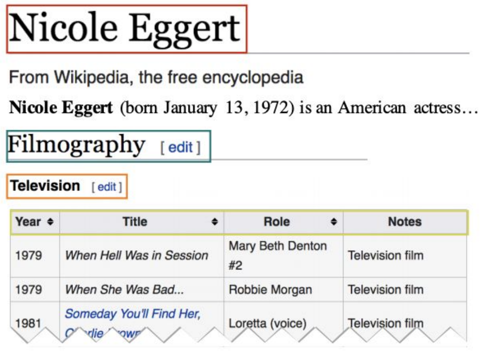
Washington (CNN) — President Donald Trump struck a unifying tone Monday as he addressed the devastation in Texas wrought by Hurricane Harvey at the top of a joint news conference with Finland's president.

"We see neighbor helping neighbor, friend helping friend and stranger helping stranger," Trump said. "We are one American family. We hurt together, we struggle together and believe me, we endure together."

Trump extended his "thoughts and prayers" to those affected by the hurricane and catastrophic flooding that ensued in Texas, and also promised Louisiana residents that the federal government is prepared to help as the tropical storm makes its way toward that state.

"To the people of Texas and Louisiana, we are 100% with you," Trump said from the East Room of the White House.

Generating Titles for Web Tables (Hancock et al. 2019)



Nicole Eggert

From Wikipedia, the free encyclopedia

Nicole Eggert (born January 13, 1972) is an American actress…

Filmography [edit]

Television [edit]

Year	Title	Role	Notes
1979	*When Hell Was in Session*	Mary Beth Denton #2	Television film
1979	*When She Was Bad...*	Robbie Morgan	Television film
1981	*Someday You'll Find Her, Charlie Brown*	Loretta (voice)	Television film

Flattened Input:

#page_title __nicole__ __eggert__ - wikipedia
#section_headings __nicole__ __eggert__
#section_headings filmography [edit]
#section_headings television [edit]
#column_headers year
#column_headers title
#column_headers role
#column_headers notes

Generated Output:

nicole eggert television roles

$$p_{gen} = \sigma(W_c c_t + W_h h_t + W_x x_t)$$

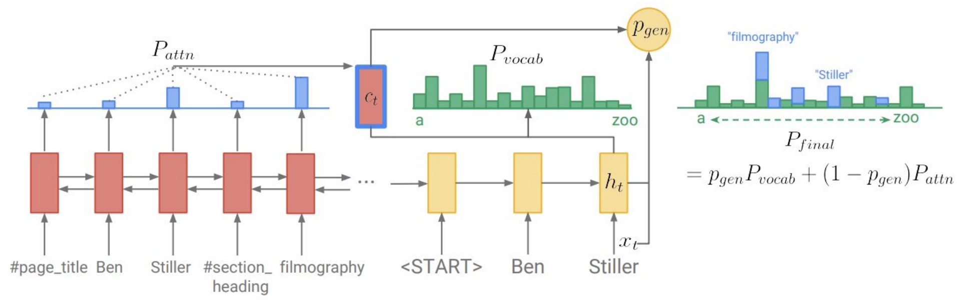
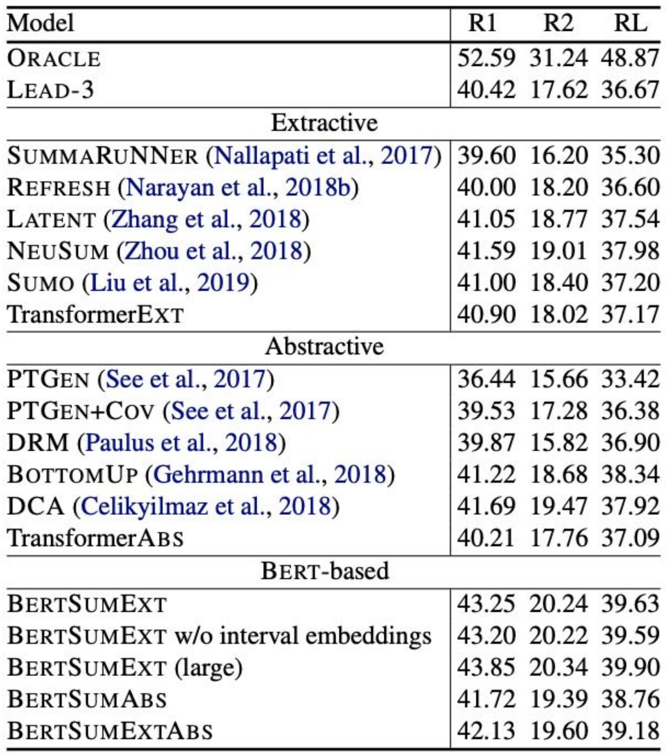


Figure 6: At each decode step, the pointer-generator network calculates the scalar value p_{gen}, which determines the relative contributions of the *vocabulary distribution* (tokens that can be generated) and the *attention distribution* (tokens that can be copied from the input sequence). Beam search is performed over the *final distribution*.

Massive dataset: original 14.1 tables scraped from google (e.g. Wikipedia) + user study 10K tables, trained from scratch.

What our dataset looked like



Model	R1	R2	RL
ORACLE	52.59	31.24	48.87
LEAD-3	40.42	17.62	36.67
Extractive			
SUMMARUNNER (Nallapati et al., 2017)	39.60	16.20	35.30
REFRESH (Narayan et al., 2018b)	40.00	18.20	36.60
LATENT (Zhang et al., 2018)	41.05	18.77	37.54
NEUSUM (Zhou et al., 2018)	41.59	19.01	37.98
SUMO (Liu et al., 2019)	41.00	18.40	37.20
TransformerEXT	40.90	18.02	37.17
Abstractive			
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
DRM (Paulus et al., 2018)	39.87	15.82	36.90
BOTTOMUP (Gehrmann et al., 2018)	41.22	18.68	38.34
DCA (Celikyilmaz et al., 2018)	41.69	19.47	37.92
TransformerABS	40.21	17.76	37.09
BERT-based			
BERTSUMEXT	43.25	20.24	39.63
BERTSUMEXT w/o interval embeddings	43.20	20.22	39.59
BERTSUMEXT (large)	43.85	20.34	39.90
BERTSUMABS	41.72	19.39	38.76
BERTSUMEXTABS	42.13	19.60	39.18

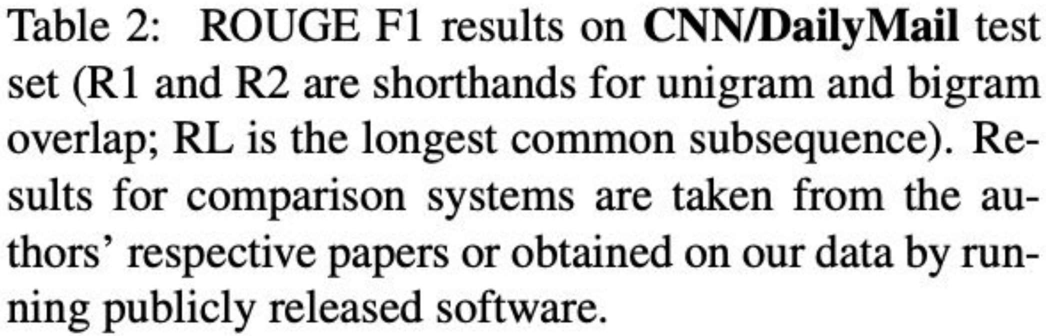


Table 2: ROUGE F1 results on **CNN/DailyMail** test set (R1 and R2 are shorthands for unigram and bigram overlap; RL is the longest common subsequence). Results for comparison systems are taken from the authors' respective papers or obtained on our data by running publicly released software.

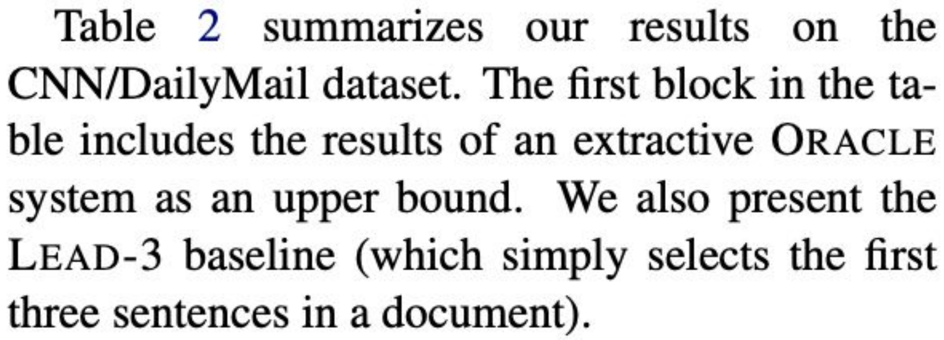
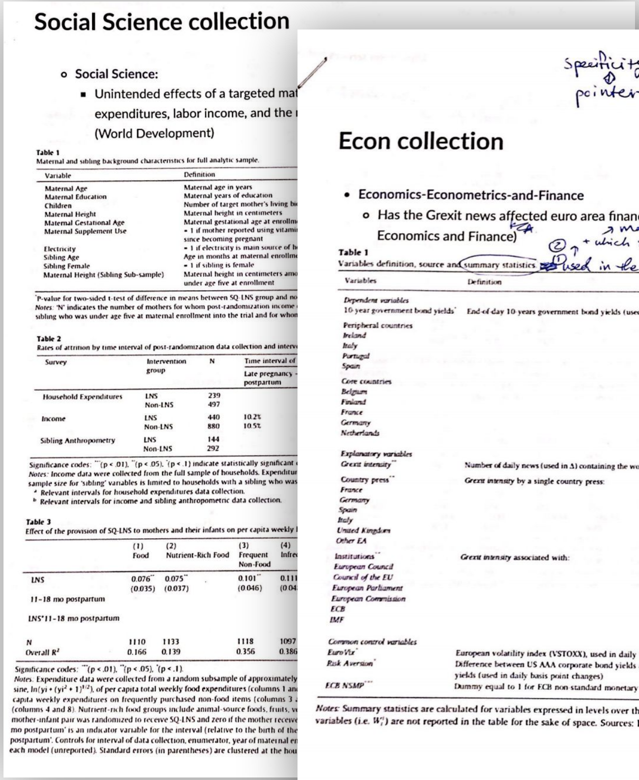


Table 2 summarizes our results on the CNN/DailyMail dataset. The first block in the table includes the results of an extractive ORACLE system as an upper bound. We also present the LEAD-3 baseline (which simply selects the first three sentences in a document).

Mixture of patterns and topics in our dataset



Social Science collection

- Social Science:
  - Unintended effects of a targeted mat
    expenditures, labor income, and the
    (World Development)

Table 1
Maternal and sibling background characteristics for full analytic sample.

Variable	Definition
Maternal Age	Maternal age in years
Maternal Education	Maternal years of education
Children	Number of target mother's living bi
Maternal Height	Maternal height in centimeters
Maternal Gestational Age	Maternal gestational age at enrollm
Maternal Supplement Use	= 1 if mother reported using vitami since becoming pregnant
Electricity	= 1 if electricity is main source of h
Sibling Age	Age in months at maternal enrollme
Sibling Female	= 1 if sibling is female
Maternal Height (Sibling Sub-sample)	Maternal height in centimeters amo under age five at enrollment

*P-value for two-sided t-test of difference in means between SQ-LNS group and no
Notes: 'N' indicates the number of mothers for whom post-randomization income
sibling who was under age five at maternal enrollment into the trial and for whon

Table 2
Rates of attrition by time interval of post-randomization data collection and interv

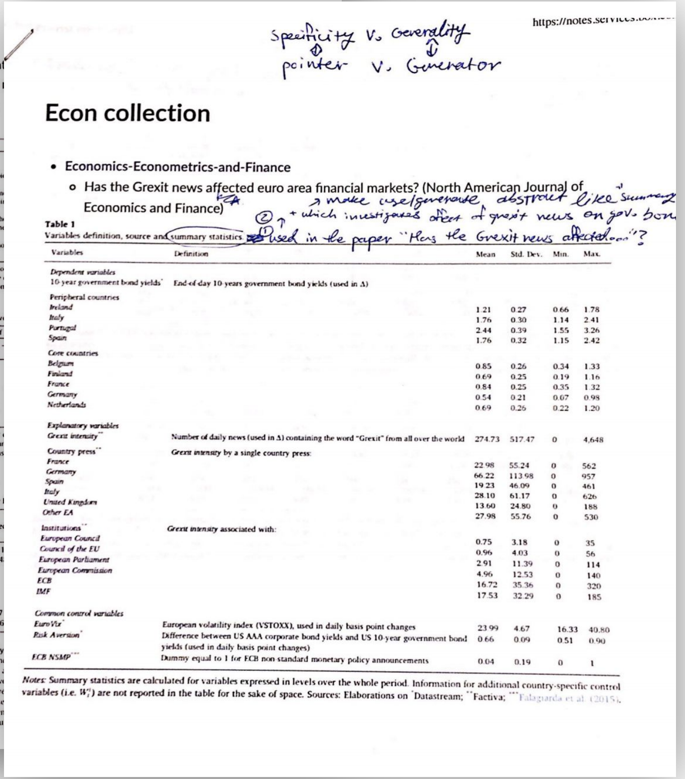
Survey	Intervention group	N	Time interval of
			Late pregnancy - postpartum
Household Expenditures	LNS	239	
	Non-LNS	497	
Income	LNS	440	10.2%
	Non-LNS	880	10.5%
Sibling Anthropometry	LNS	144	
	Non-LNS	292	

Significance codes: ***(p < .01), **(p < .05), *(p < .1) indicate statistically significant
Notes: Income data were collected from the full sample of households. Expenditur
sample size for 'sibling' variables is limited to households with a sibling who was
[a] Relevant intervals for household expenditures data collection.
[b] Relevant intervals for income and sibling anthropometric data collection.

Table 3
Effect of the provision of SQ-LNS to mothers and their infants on per capita weekly

	(1) Food	(2) Nutrient-Rich Food	(3) Frequent Non-Food	(4) Infre
LNS	0.076**	0.075**	0.101**	0.111
	(0.035)	(0.037)	(0.046)	(0.04
11–18 mo postpartum				
LNS*11–18 mo postpartum				
N	1110	1133	1118	1097
Overall R^2	0.166	0.139	0.356	0.386

Significance codes: ***(p < .01), **(p < .05), *(p < .1).
Notes: Expenditure data were collected from a random subsample of approximately
sine, $\ln(yi + (yi^2 + 1)^{1/2})$, of per capita total weekly food expenditures (columns 1 an
capita weekly expenditures on frequently purchased non-food items (columns 3
(columns 4 and 8). Nutrient-rich food groups include animal-source foods, fruits, v
mother-infant pair was randomized to receive SQ-LNS and zero if the mother receive
mo postpartum' is an indicator variable for the interval (relative to the birth of the
postpartum'. Controls for interval of data collection, enumerator, year of maternal en
each model (unreported). Standard errors (in parentheses) are clustered at the hou



https://notes.services.

Specificity v. Generality
⇕ ⇕
pointer v. Generator

Econ collection

- Economics-Econometrics-and-Finance
  - Has the Grexit news affected euro area financial markets? (North American Journal of Economics and Finance)

→ make use/generate abstract like summary
② ↑ + which investigates effect of grexit news on gov. bon
① used in the paper "Has the Grexit news affected..."?

Table 1
Variables definition, source and summary statistics.

Variables	Definition	Mean	Std. Dev.	Min.	Max.
Dependent variables					
10-year government bond yields*	End-of-day 10-years government bond yields (used in Δ)				
Peripheral countries					
Ireland		1.21	0.27	0.66	1.78
Italy		1.76	0.30	1.14	2.41
Portugal		2.44	0.39	1.55	3.26
Spain		1.76	0.32	1.15	2.42
Core countries					
Belgium		0.85	0.26	0.34	1.33
Finland		0.69	0.25	0.19	1.16
France		0.84	0.25	0.35	1.32
Germany		0.54	0.21	0.07	0.98
Netherlands		0.69	0.26	0.22	1.20
Explanatory variables					
*Grexit intensity***	Number of daily news (used in Δ) containing the word "Grexit" from all over the world	274.73	517.47	0	4,648
Country press**	*Grexit intensity* by a single country press:				
France		22.98	55.24	0	562
Germany		66.22	113.98	0	957
Spain		19.23	46.09	0	461
Italy		28.10	61.17	0	626
United Kingdom		13.60	24.80	0	188
Other EA		27.98	55.76	0	530
Institutions**	*Grexit intensity* associated with:				
European Council		0.75	3.18	0	35
Council of the EU		0.96	4.03	0	56
European Parliament		2.91	11.39	0	114
European Commission		4.96	12.53	0	140
ECB		16.72	35.36	0	320
IMF		17.53	32.29	0	185
Common control variables					
*EuroVix**	European volatility index (VSTOXX), used in daily basis point changes	23.99	4.67	16.33	40.80
*Risk Aversion**	Difference between US AAA corporate bond yields and US 10-year government bond yields (used in daily basis point changes)	0.66	0.09	0.51	0.90
*ECB NSMP****	Dummy equal to 1 for ECB non-standard monetary policy announcements	0.04	0.19	0	1

Notes: Summary statistics are calculated for variables expressed in levels over the whole period. Information for additional country-specific control variables (i.e. W_i^t) are not reported in the table for the sake of space. Sources: Elaborations on *Datastream; **Factiva; ***Falagiarda et al. (2015).

Pattern 1: $\underbrace{\text{Table X}}_{\text{Table name}}$ $\underbrace{\text{presents/shows/explains}}_{\text{Verb}}$ $\underbrace{\text{\_\_}}_{\text{Desc.}}$

Pattern 2: As we $\underbrace{\text{show}}_{\text{Verb}}$ in $\underbrace{\text{Table Y}}_{\text{Table name}}$, $\underbrace{\text{\_\_\_\_\_\_}}_{\text{Desc.}}$

Pattern 3: $\underbrace{\text{\_\_\_\_\_\_}}_{\text{Desc.}}$, $\underbrace{\text{see}}_{\text{Verb}}$ $\underbrace{\text{Table Z}}_{\text{Table name}}$

....

BERT for text summarization (Liu et al. 2019)

Text Summarization with Pre-trained Encoders

Yang Liu, Mirella Lapata (EMNLP 2019)

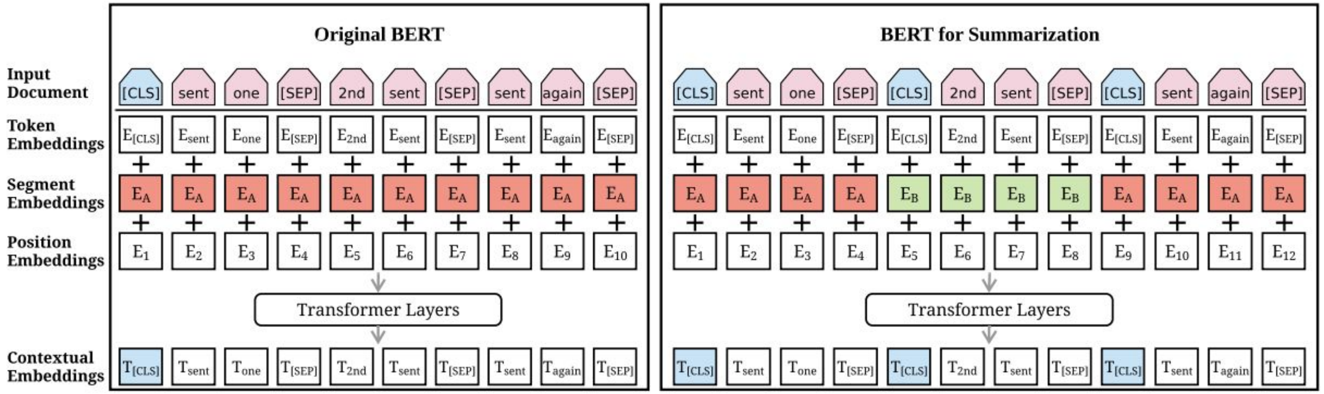
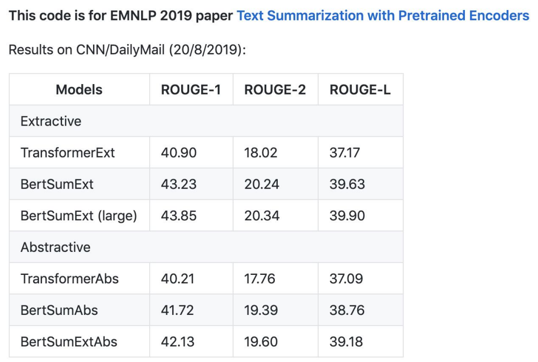


Figure 1: Architecture of the original BERT model (left) and BERTSUM (right). The sequence on top is the input document, followed by the summation of three kinds of embeddings for each token. The summed vectors are used as input embeddings to several bidirectional Transformer layers, generating contextual vectors for each token. BERTSUM extends BERT by inserting multiple [CLS] symbols to learn sentence representations and using interval segmentation embeddings (illustrated in red and green color) to distinguish multiple sentences.



This code is for EMNLP 2019 paper Text Summarization with Pretrained Encoders

Results on CNN/DailyMail (20/8/2019):

Models	ROUGE-1	ROUGE-2	ROUGE-L
Extractive			
TransformerExt	40.90	18.02	37.17
BertSumExt	43.23	20.24	39.63
BertSumExt (large)	43.85	20.34	39.90
Abstractive			
TransformerAbs	40.21	17.76	37.09
BertSumAbs	41.72	19.39	38.76
BertSumExtAbs	42.13	19.60	39.18

BERT vs. other pre-trainings

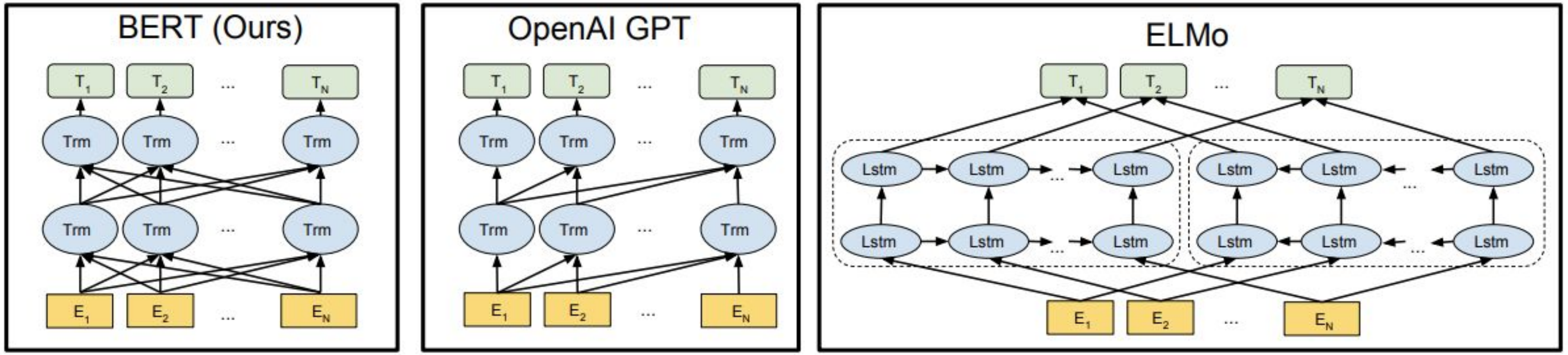


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

BERT fine-tuning original tasks

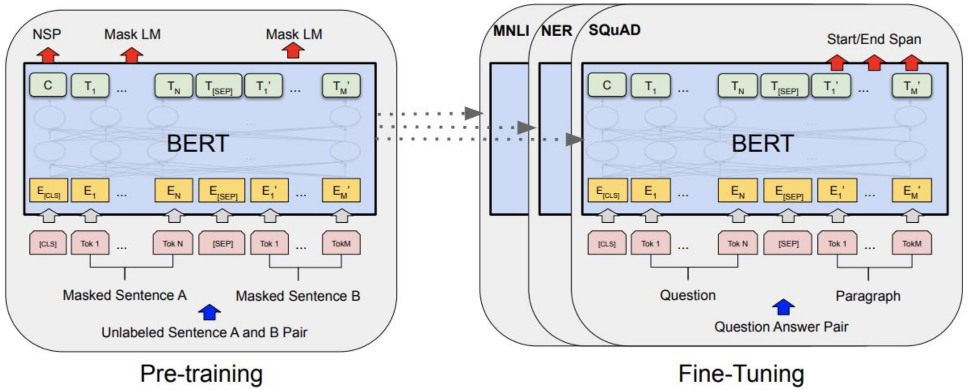
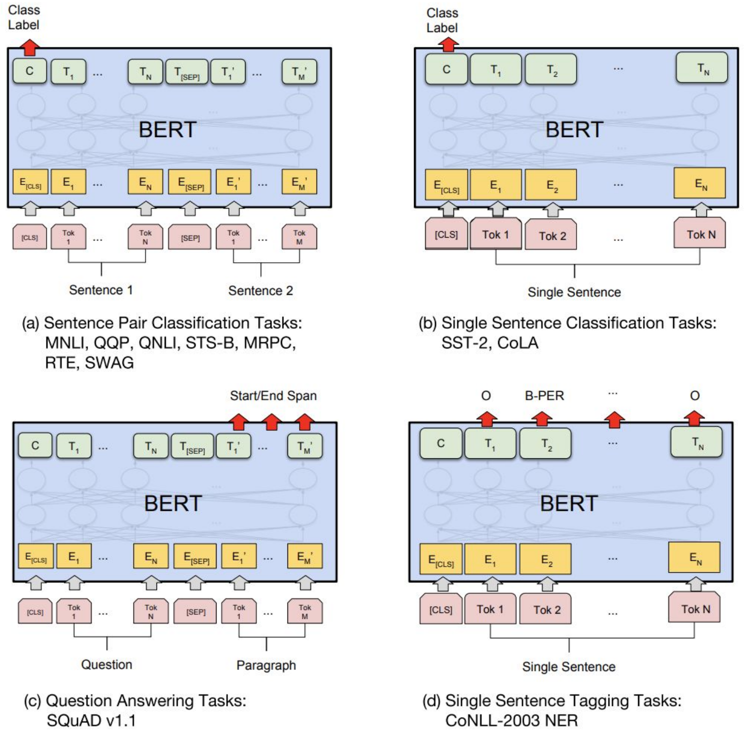


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).


Class Label
BERT
Sentence 1
Sentence 2
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
Class Label
BERT
Single Sentence
(b) Single Sentence Classification Tasks: SST-2, CoLA
Start/End Span
BERT
Question
Paragraph
(c) Question Answering Tasks: SQuAD v1.1
O
B-PER
O
BERT
Single Sentence
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Liu et al., BERT fine-tuning for text summaries

Transformers

build passing | license Apache-2.0 | website online | release v2.3.0

State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch

Extractive

$$\tilde{h}^l = \text{Layer Norm}\,(h^{l-1} + \text{Multi-Head att.}\,(h^{l-1}))$$

$$h^l = \text{Layer Norm}\,(\tilde{h}^l + \text{Feed-Forward}\,(\tilde{h}^l))$$

Output

$$\hat{y}_i = \sigma(W_o h_i^L + b_o)$$

Loss: Binary classification entropy

$$L_i = H_b(\hat{y}_i)$$

Abstractive

Continue training from Extractive

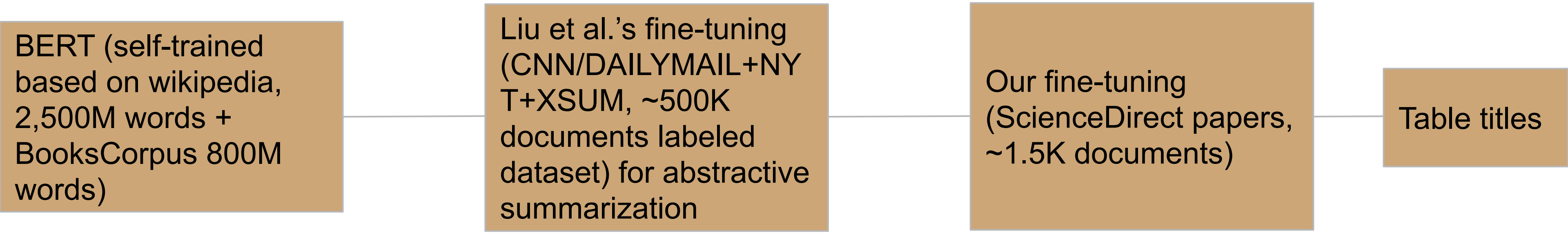
Decoder: 6-layered Transformer initialized randomly + label smoothing + allow higher slack in KL divergence

Optimize encoder/decoder separately for Stability

Subword tokenizer handles OOV

BERT transfer learning to the rescue

With a little bit of trial (and error) it turns out continuing to train Liu et al.'s model with our dataset of 1500k samples was surprisingly successful!


BERT (self-trained based on wikipedia, 2,500M words + BooksCorpus 800M words)
Liu et al.'s fine-tuning (CNN/DAILYMAIL+NYT+XSUM, ~500K documents labeled dataset) for abstractive summarization
Our fine-tuning (ScienceDirect papers, ~1.5K documents)
Table titles

Experience and Insights

- Deep nets are excellent at dealing with noise. Do not spend too much time cleaning your data! Try it first.

- A small dataset can be used to fine tune a model that was first trained on a big, established corpus. It might perform better than you think.

- Be careful not to overfit your dataset when moving from bigger dataset to smaller one. Early stopping is important.

- Summarization metrics (ROUGE, BLEU) are not as relevant for abstractive summaries compared to extractive, but can still be a useful ballpark.