NLP Generalization

Rahul Kulkarni, Kevin Hanna, Justin Stanley
W210 Capstone, MIDS, Summer 2020
UC Berkeley School of Information

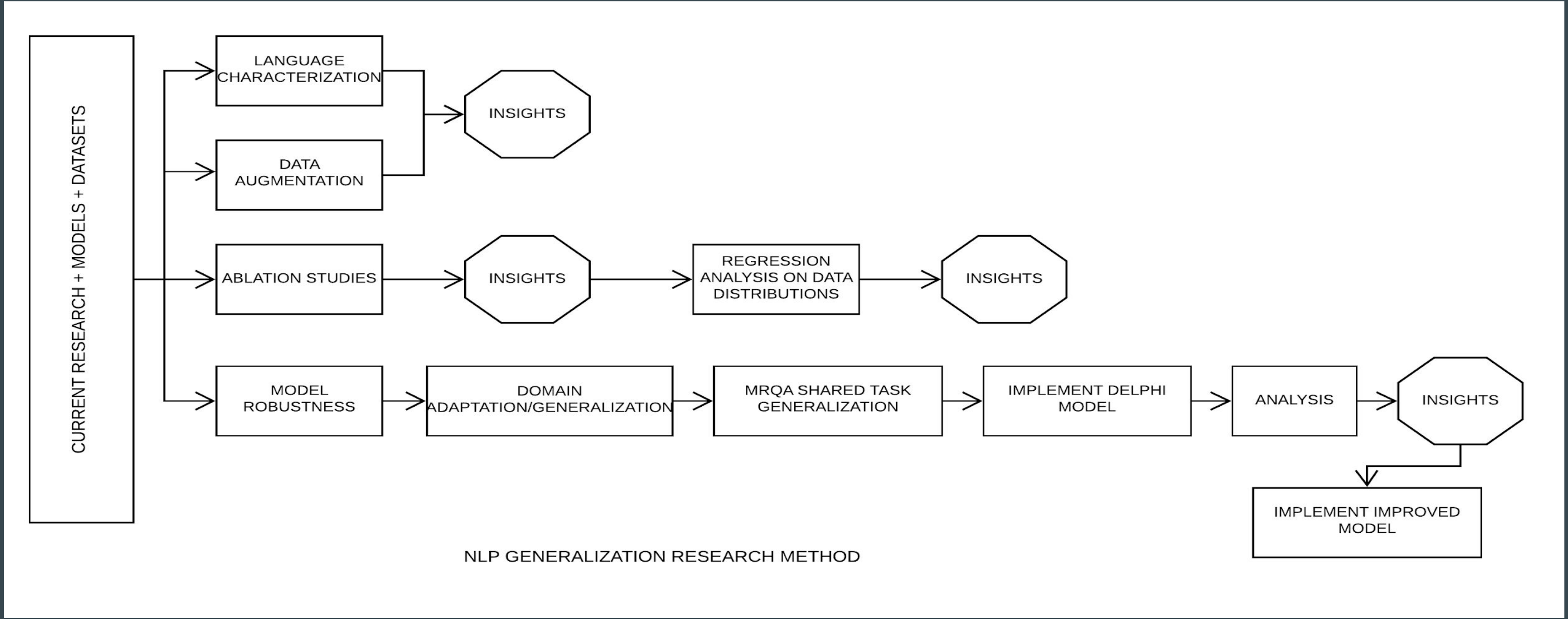
Background

- Prior work from Roelofs, Schmidt et al. points out the shortcomings of current classification algorithms and their sensitivity to small shifts in the underlying data distribution
- Key highlights of the Research:
    - Distribution shifts in data hinders the algorithm to generalize, leading to incorrect decisions and severe consequences
    - Inability to interpret why a ML algorithm arrived at a certain decision, driven by empirical progress with little guidance from theory
    - Majority of published papers justify improved performance using key benchmarks with little explanation
- Propose to empirically examine key theoretical pillars of ML to build algorithms that are reliable and robust by building accurate measurements for generalization

Problem Statement

- NLP models exhibit a generalization gap when evaluated on un-seen or out-of-domain datasets
- Miller, et. al. demonstrated on SQuAD models, robustness to some natural distribution shifts, though they still suffer substantial performance degradation on others such as the New York Times, Reddit and Amazon Datasets and the degradation being attributed to changes in the source text.
- Miller, et. al. raises the following questions:
  - What are some metrics to compare Q&A datasets to explain performance differences
  - Why do models perform better on certain Q&A datasets vs others
  - What is the interplay of training models with additional data and model robustness

Research Method


CURRENT RESEARCH + MODELS + DATASETS
LANGUAGE CHARACTERIZATION
DATA AUGMENTATION
INSIGHTS
ABLATION STUDIES
INSIGHTS
REGRESSION ANALYSIS ON DATA DISTRIBUTIONS
INSIGHTS
MODEL ROBUSTNESS
DOMAIN ADAPTATION/GENERALIZATION
MRQA SHARED TASK GENERALIZATION
IMPLEMENT DELPHI MODEL
ANALYSIS
INSIGHTS
IMPLEMENT IMPROVED MODEL
NLP GENERALIZATION RESEARCH METHOD

Datasets: SQuAD, MRQA, NYT, REDDIT, AMAZON, WIKIPEDIA
Models: BERT, XLNET, DELPHI

Libraries: PyTorch, HuggingFace, AllenNLP, WordsAPI, PyHypen
Platforms: CodaLab, AWS

Extractive QA Example

CONTEXT

“The building, which originally had 17 rentals and has been in the same hands since the 1970s, benefits from proximity to private schools, said Nikki Field, an associate broker with Sotheby’s International Realty, which is marketing the property. Another appeal is that the area, part of Carnegie Hill and Lenox Hill, could be considered club land. ”

QUESTIONS

The building has had the same owner since when?	“1970s” , “the 1970s”
Carnegie Hill and Lenox Hill could be considered what?	“club land”

Extractive QA Model Evaluation

- Normalized Predictions/Correct Answers
  - Remove Common Articles (a, an, the)
  - Remove Consecutive Whitespace
  - Remove Punctuation
  - Convert to Lowercase

- Exact Match
  - Normalized prediction exactly matches a possible answer.
- F1 Score
  - Best F1 score between the normalized, predicted answer and each possible answer.
  - Calculated comparing space-delimited tokens

F1 Example

Answer: “UC Berkeley School of Information”

Prediction: “School of Information”

Answer Tokens: [uc, berkeley, school, of, information]

Prediction Tokens: [school, of, information]

Matched Tokens: [school, of, information]

Precision = # Matched Tokens / # Predicted Tokens = 3/3 = 1.0

Recall = # Matched Tokens / # Answer Tokens = 3/5 = 0.6

F1 = (2 * Precision * Recall) / (Precision + Recall) = (2 * 1.0 * 0.6) / (1.0 + 0.6) = .75

Distribution Shift Example

Wikipedia:

“The Canon EOS 5D Mark III is a professional-grade 22.3 megapixels full-frame digital single-lens reflex (DSLR) camera made by Canon.

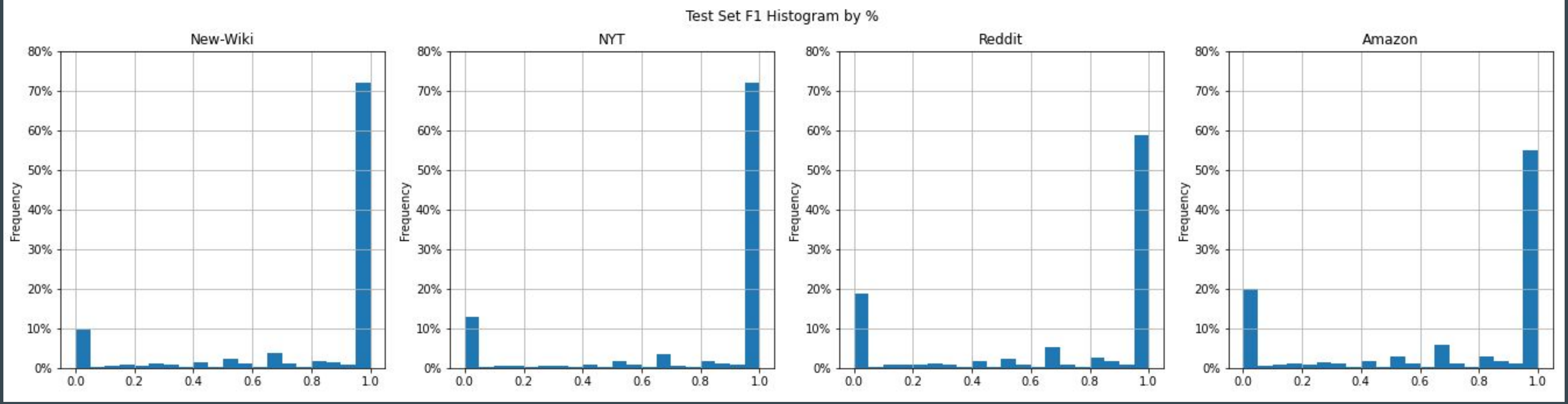
Succeeding the EOS 5D Mark II, it was announced on 2 March 2012, the 25th anniversary of the announcement of the first camera in the EOS line, the EOS 650, and was also Canon's 75th anniversary. The Mark III went on sale later in March with a retail price of $3,499 in the US, £2999 in the UK, and €3569 in the Eurozone.”

Amazon:

“I was relieved that Canon DIDN'T try to stuff 36 megapixels into the Mk III. They kept it roughly the same at 22mp. Way to go, Canon! It's been proven time and time again that more megapixels doesn't make for a sharper image, only larger file sizes. "More megapixels equals better image quality is what's known as "the megapixel myth" Cramming in more megapixels means a lower signal-to-noise ratio and less full well capacity for each photo site. At some point you don't get more detail with a higher pixel count; you just spread the detail around on more pixels. I hardly ever need 21mp as it is...”

EDA

Objective - Develop quantitative method for measuring the distribution shift between training and test datasets.



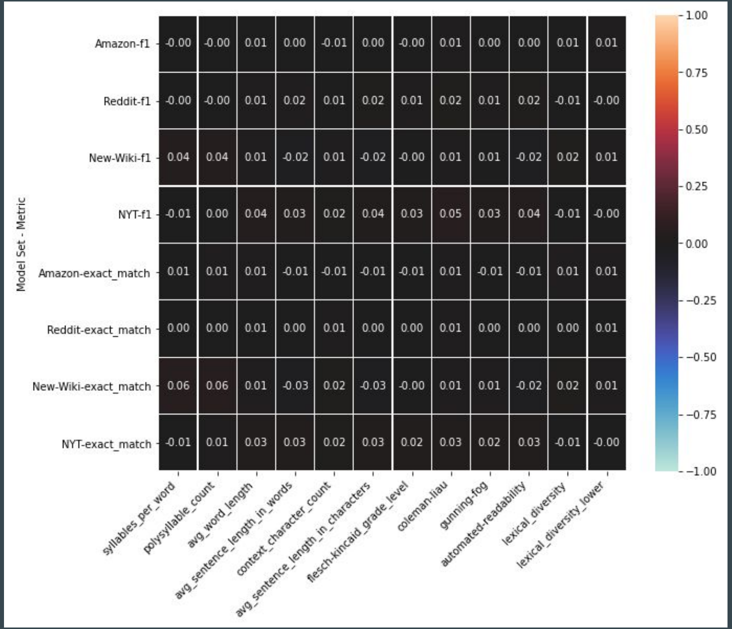
Language Complexity and Readability Metrics

Examination of relationship between various linguistic complexity metrics and model scores.

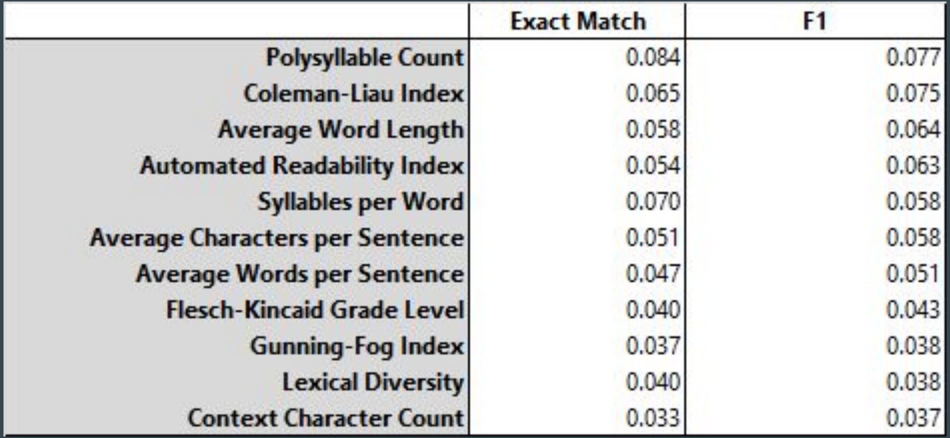
- Flesch-Kincaid Grade Level (word and syllable counts)
- Coleman-Liau Index (character and and sentence counts)
- Gunning Fog Index (word, sentence, and polysyllable word counts)
- Automated Readability Index (alphanumeric character, word, and sentence counts)
- Lexical Diversity (unique words vs. total words)

Metric Correlation Results

No significant correlation found between F1 scores and any examined metrics at either the test set or model level.



	syllables_per_word	polysyllable_count	avg_word_length	avg_sentence_length_in_words	context_character_count	avg_sentence_length_in_characters	flesch-kincaid_grade_level	coleman-liau	gunning-fog	automated-readability	lexical_diversity	lexical_diversity_lower
Amazon-f1	-0.00	-0.00	0.01	0.00	-0.01	0.00	-0.00	0.01	0.00	0.00	0.01	0.01
Reddit-f1	-0.00	-0.00	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.02	-0.01	-0.00
New-Wiki-f1	0.04	0.04	0.01	-0.02	0.01	-0.02	-0.00	0.01	0.01	-0.02	0.02	0.01
NYT-f1	-0.01	0.00	0.04	0.03	0.02	0.04	0.03	0.05	0.03	0.04	-0.01	-0.00
Amazon-exact_match	0.01	0.01	0.01	-0.01	-0.01	-0.01	-0.01	0.01	-0.01	-0.01	0.01	0.01
Reddit-exact_match	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01
New-Wiki-exact_match	0.06	0.06	0.01	-0.03	0.02	-0.03	-0.00	0.01	0.01	-0.02	0.02	0.01
NYT-exact_match	-0.01	0.01	0.03	0.03	0.02	0.03	0.02	0.03	0.02	0.03	-0.01	-0.00



	Exact Match	F1
Polysyllable Count	0.084	0.077
Coleman-Liau Index	0.065	0.075
Average Word Length	0.058	0.064
Automated Readability Index	0.054	0.063
Syllables per Word	0.070	0.058
Average Characters per Sentence	0.051	0.058
Average Words per Sentence	0.047	0.051
Flesch-Kincaid Grade Level	0.040	0.043
Gunning-Fog Index	0.037	0.038
Lexical Diversity	0.040	0.038
Context Character Count	0.033	0.037

Data Augmentation

- Two Data Augmentation experiments
  - Reduce the distribution gap between NLP corpora.
- Masked tokens and predicted replacements
  - Bert Large Cased.
  - 4 question answering test sets.

An improvement in F1 Score from predictions using Bert Large Cased would suggest the augmentation was narrowing the distribution gap. Providing a pathway for further experimentation and analysis.

Data Augmentation: Experiment 1 - Parts of Speech

- Masked tokens based on part of speech.
  - verb
  - adverb
  - verb, adverb
  - adjective
  - adjective, verb
- Best performing part of speech augmentation was ‘verb’ with an increase in mean F1 of 2.4% (80.0 to 82.3).

Data Augmentation: Experiment 2 - Word Frequency

- Masked tokens based on their rank in the context based on frequency the token was used in Wikipedia
    - Lowest 10, 20, 30 and 50 percentiles.
- Best performing word frequency augmentation was 10-percentile with an increase in mean F1 of 0.048% (79.99 to 80.03).

Data Augmentation - Results

- Augmentations were able to make slight increases in F1 scores.
  - Increase in F1 scores was minor, significantly smaller than real world distributions.
  - Changes in mean F1 scores were not systematic, resulted from a large number changes in question accuracy in both directions
    - E.g. : 2,604 context question pairs increased in accuracy and 1,770 decreased.
- Results were not significant for the purpose of our analysis
  - Resulting changes in mean F1 scores were too minor.
  - Resulting metrics included a great deal of noise.

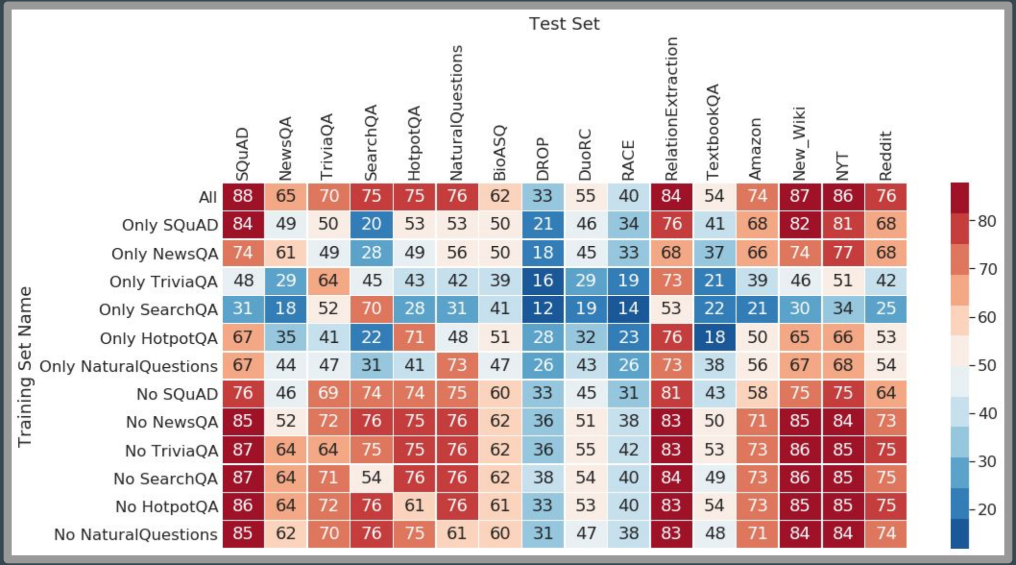
Ablation Study

- Using 6 MRQA training sets, we fine tuned Bert Base Cased with 43 permutations.
- Predicted answers and scored results on 16 test sets.

	Single Training Set		Training Set Excluded	
Model	Mean F1	F1 Delta	Mean F1	F1 Delta
Baseline	68.77		68.77	
SQuAD	54.71	**-14.05**	61.30	**-7.47**
NewsQA	53.40	-15.36	66.81	-1.95
TriviaQA	40.40	-28.36	68.25	**-0.51**
SearchQA	31.30	**-37.47**	67.08	-1.68
HotpotQA	46.73	-22.03	67.30	-1.46
NaturalQuestions	49.87	-18.90	65.54	-3.22

Ablation Study - Findings

- SQuAD was the most influential training set for generalizing models.
- The 13 models included in the heatmap highlight distribution gaps, higher in-domain scores and lower out-of-domain F1 scores.

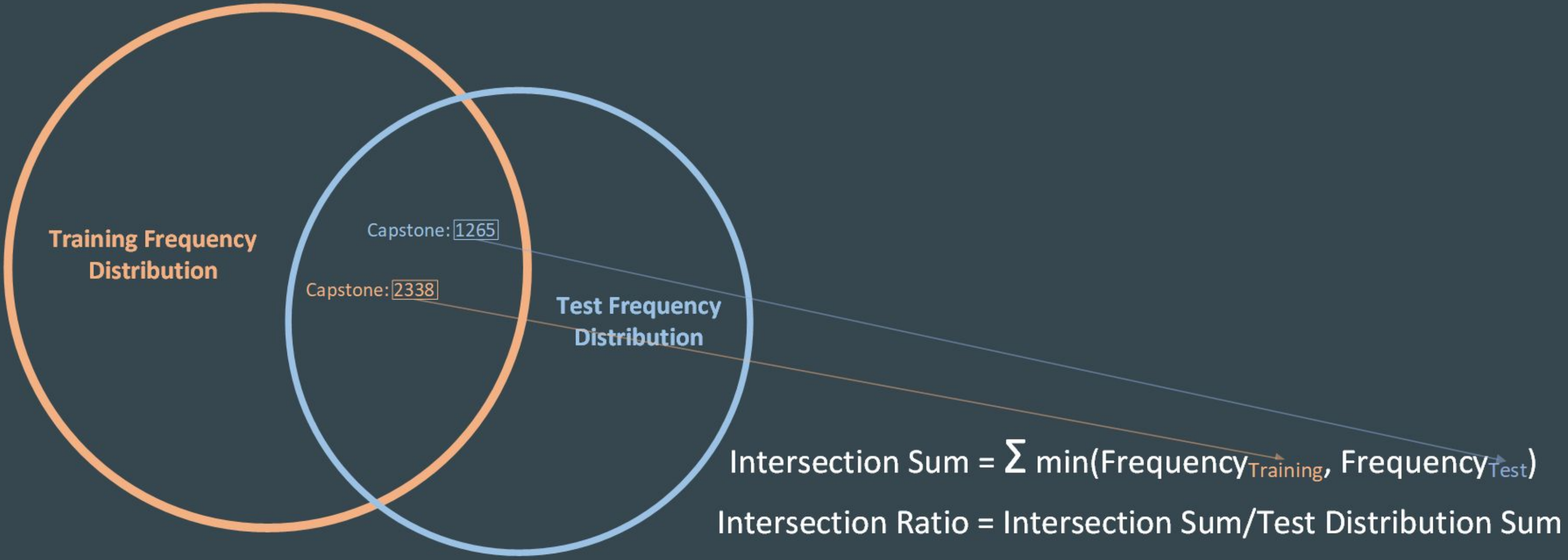


Test Set

Training Set Name	SQuAD	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQuestions	BioASQ	DROP	DuoRC	RACE	RelationExtraction	TextbookQA	Amazon	New_Wiki	NYT	Reddit
All	88	65	70	75	75	76	62	33	55	40	84	54	74	87	86	76
Only SQuAD	84	49	50	20	53	53	50	21	46	34	76	41	68	82	81	68
Only NewsQA	74	61	49	28	49	56	50	18	45	33	68	37	66	74	77	68
Only TriviaQA	48	29	64	45	43	42	39	16	29	19	73	21	39	46	51	42
Only SearchQA	31	18	52	70	28	31	41	12	19	14	53	22	21	30	34	25
Only HotpotQA	67	35	41	22	71	48	51	28	32	23	76	18	50	65	66	53
Only NaturalQuestions	67	44	47	31	41	73	47	26	43	26	73	38	56	67	68	54
No SQuAD	76	46	69	74	74	75	60	33	45	31	81	43	58	75	75	64
No NewsQA	85	52	72	76	75	76	62	36	51	38	83	50	71	85	84	73
No TriviaQA	87	64	64	75	75	76	62	36	55	42	83	53	73	86	85	75
No SearchQA	87	64	71	54	76	76	62	38	54	40	84	49	73	86	85	75
No HotpotQA	86	64	72	76	61	76	61	33	53	40	83	54	73	85	85	75
No NaturalQuestions	85	62	70	76	75	61	60	31	47	38	83	48	71	84	84	74

N-Gram Frequency Distributions

- Created n-gram frequency distributions for each of the 6 MRQA training sets.
- Developed an “Intersection Ratio” to quantify the overlap in distributions
  - Sum of common n-grams in the intersection and divided by the sum of n-grams in the test distribution.



Intersection Sum = $\Sigma \min(\text{Frequency}_{\text{Training}}, \text{Frequency}_{\text{Test}})$

Intersection Ratio = Intersection Sum/Test Distribution Sum

N-Gram Frequency Distribution Analysis

- Used regression analysis to find the correlation between the intersection ratios F1 scores excluding in-domain models.
    - Found a positive correlation between the ratio and resulting F1 scores.
        - R-Squared: 11.6%
        - Coefficient: 13.1 (SE: 3.7, 95% CI: 5.8-20.4)
        - Mean Ratio: 53.2%
        - For every percent increase in 1-gram intersection ratio, there is a 1.3% increase in F1.

Regression Insights

- F1 score is negatively correlated with the test set mean context token length.
- F1 score is positively correlated with the number of training examples.
- F1 score is negatively correlated with training set mean context length.
  - For every additional 100 tokens in the mean F1 score drops by 1.8%.
  - There are some confounding factors such as a relationship between context source and context length.
  - Requires a specific experiment equalizing the mean context tokens for each training set.
- F1 score is inversely correlated with the number of unique n-grams in the frequency distributions.

Model Robustness Background

- MRQA Shared Task Primarily focuses on Generalization for extractive Question and Answering
    - Uses 6 Training Sets and 6 Dev/Test Sets from Multiple Domains
- Delphi MRQA model was the most robust model in John et al. generalization tests and second on leaderboard on MRQA benchmarks
    - Domain-agnostic question answering model for the Machine Reading Question Answering (Longpre et al., 2019)*
    - Established high generalizability by fine-tuning on multiple MRQA datasets (multi-domain)
    - Negative sampling (including No-answer data in the training set)

Model Robustness Implementation

- Replicated Delphi Model evaluation on codalab on 4 test sets here
- Model training code for Delphi NOT available
- Implemented algorithm for negative sampling techniques to train BERT using no-answer segments:
  - Traditional extractive QA does not have a notion of negative class
  - All (Qn,Context) pairs are guaranteed to have an answer
  - When long contexts are broken into multiple segments the presence of an answer is not guaranteed within each segment
  - At Inference time model computes the most probable answer span for each segment independently (Sum of start and end span probs)
  - At training time NA segments are discarded
  - Include naturally occuring “Negative” segments during training
  - For each negative segment set the indices for start and end span to point to [CLS] token
  - Train the model to *abstain* from selecting a span in a given segment
  - At Inference time select the highest probability answer across all segments excluding the No Answer option

Ongoing Work

- Establish benchmarks on our model aligned with Delphi and MRQA results
- Collaborate with John and Ludwig on further analysis and characterization
- Validate our model code from Delphi Team (Apple)
- Contribute back to Hugging Face Transformers codebase
- Propose an improved model for extractive QA with improved robustness

Conclusion

- Language Characterization using multiple descriptive measures of context complexity and F1 scores showed no correlation
- We saw some evidence of lexical complexity of corpora correlate with F1 scores
- Data Augmentation experiments reduced the distribution gap by a small fraction
- Our Ablation studies showed in-domain out performed out-of-domain and SQuAD was the most influential training set
  - Expertly written text does better?
- Regression Analysis showed smaller contexts did better than larger contexts
  - Needs more experimentation
- No Answer Trained MRQA models are most robust for extractive QA, we continue our exploits in this area

Acknowledgments

- Alberto Todeschini and Puya Vahabi for guidance and introducing us to the problem of Model Robustness and Generalization
- John Miller and Ludwig Schmidt from Berkeley CS dept. for their support by providing us feedback and sharing their source code for their paper
- Apple Delphi Team for answering all our questions on the Delphi model implementation