CS4248 Project

Machine Comprehension

Question-Answering
on
SQUAD 2.0

Group G09:

Dolly Agarwal (A0228490B)
Lim Wei Quan Ernest (A0201835M)
Rashi Sharma (A0228492X)
Rohit Jain (A0228500R)
Simran Aggarwal (A0228520M)

Problem Statement

Passage (*P*) + Question (*Q*) ⟶ Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q What city is Alyssa in? *A* Miami

Background

- Machine Learning Approaches –
  - Sliding Window (Richardson et. Al, 2013)
    - Compute the unigram/bigram overlap between the sentence containing the candidate answer and the question.
    - Use TF-IDF based similarity to select the best candidate answer
  - Logistic Regression (Rajpurkar et. Al, 2013)
    - Extract several types of features for each candidate answer
    - Features
      - Matching Word Frequencies
      - Matching Bigram Frequencies
      - Lengths
      - Span POS Tags

Background

- Deep Learning Approaches –

  - RNN + Attention (2016)
    - Bi-directional Attention Flow (BiDAF) network
      - query-aware context representation

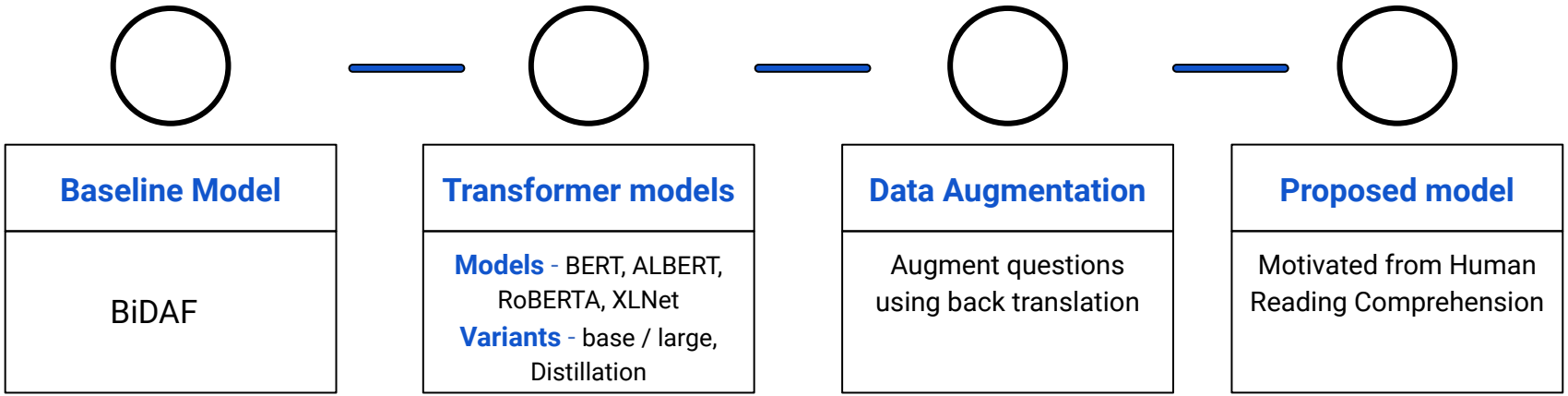
  - Transformer Models (2019 - Present)
    - pretrained weights on a large-scale language modeling dataset
    - Dominates SQUAD 2.0 leaderboard
    - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
    - BERT, ALBERT, RoBERTa etc.

Dataset: Stanford Question Answering Dataset (SQuAD) V2.0

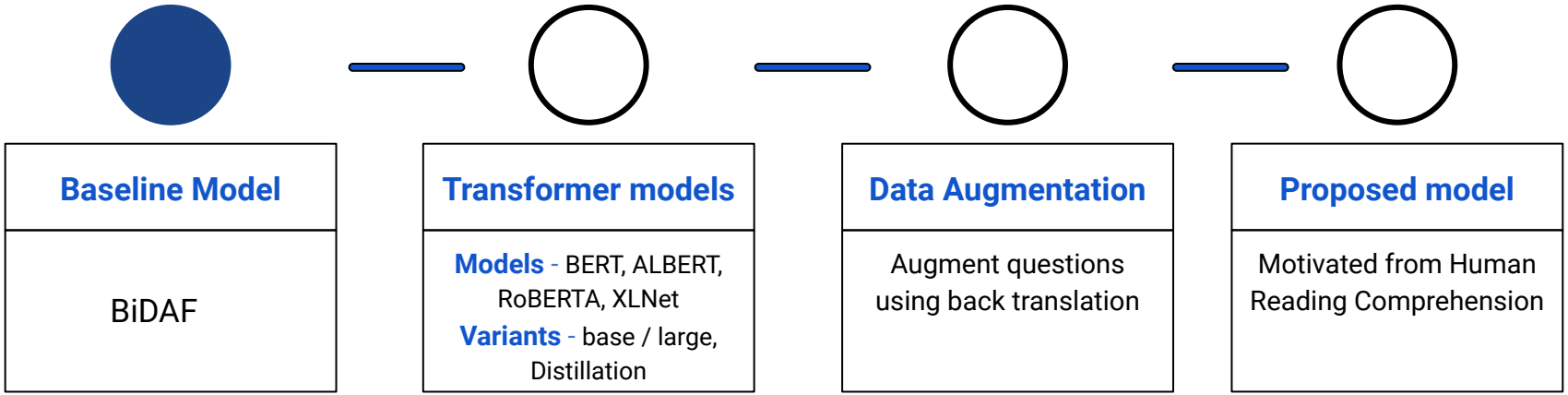
- Reading Comprehension Dataset
- SQUAD v1.0 - All questions answerable
- High-quality wikipedia articles
- Covers a diverse range of topics across a variety of domains
- Answer to question is a segment of text, or span
- SQUAD v2.0 - Question might be unanswerable

Type	Number of Questions	Number of Questions with No Answer
Train	130319	43498 (33%)
Dev	11873	5945 (50%)

Experiments

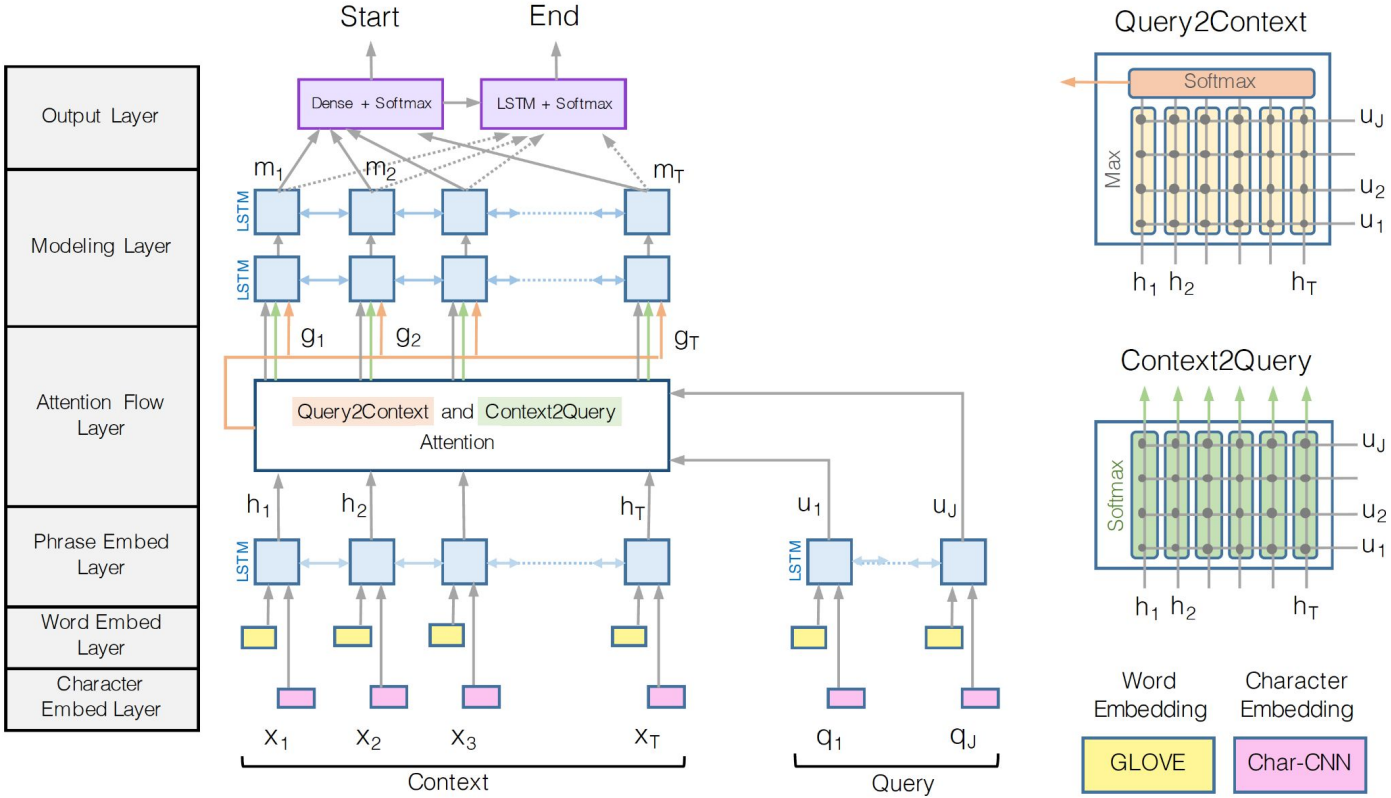


Experiments

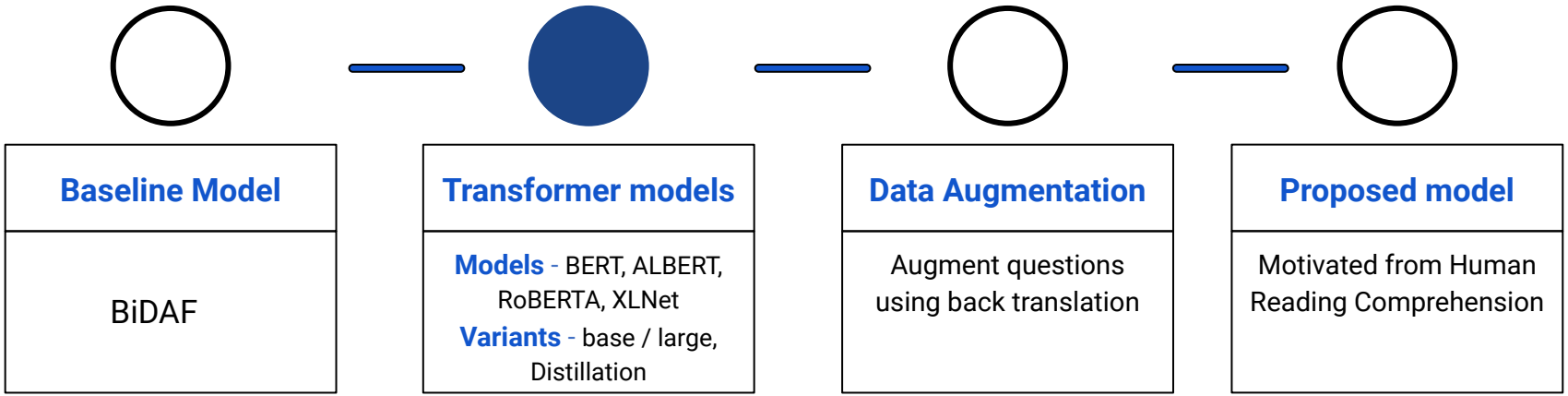


Baseline Model

1) Bi-directional Attention Flow (BiDAF)



Experiments



Transformer based models

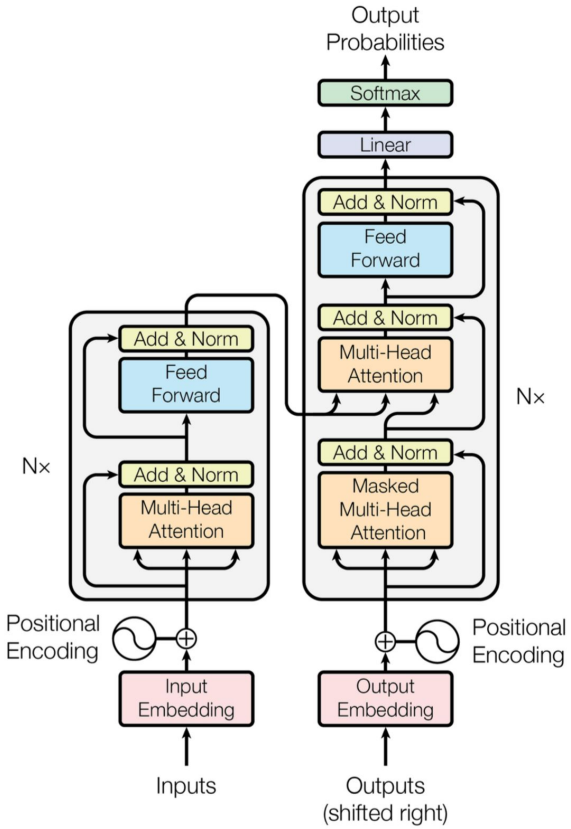


Figure 1: The Transformer - model architecture.

Models:

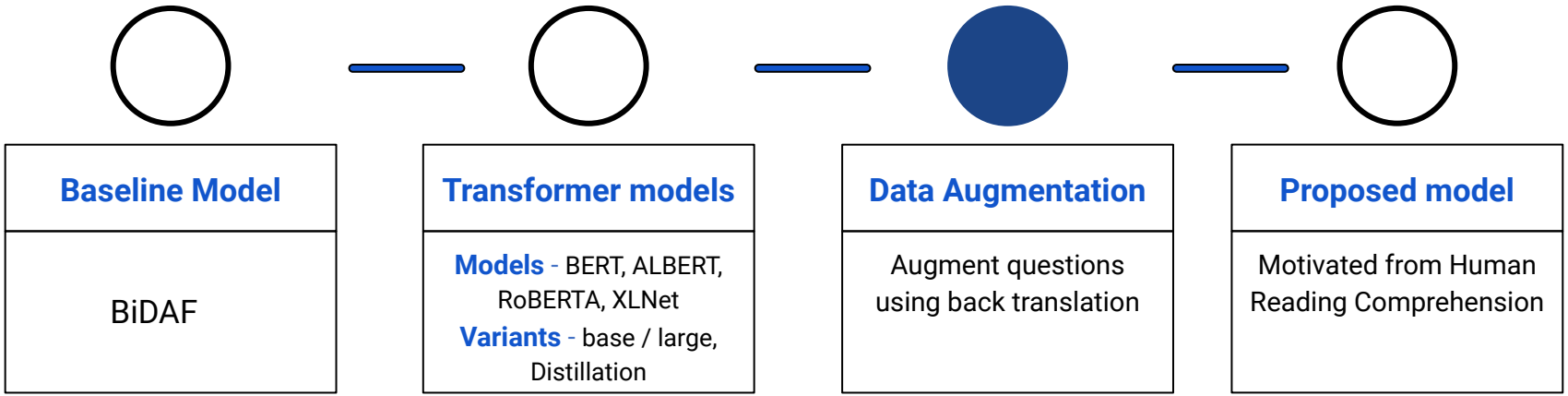
- BERT
- ALBERT
- RoBERTA
- XLNet

Variants:

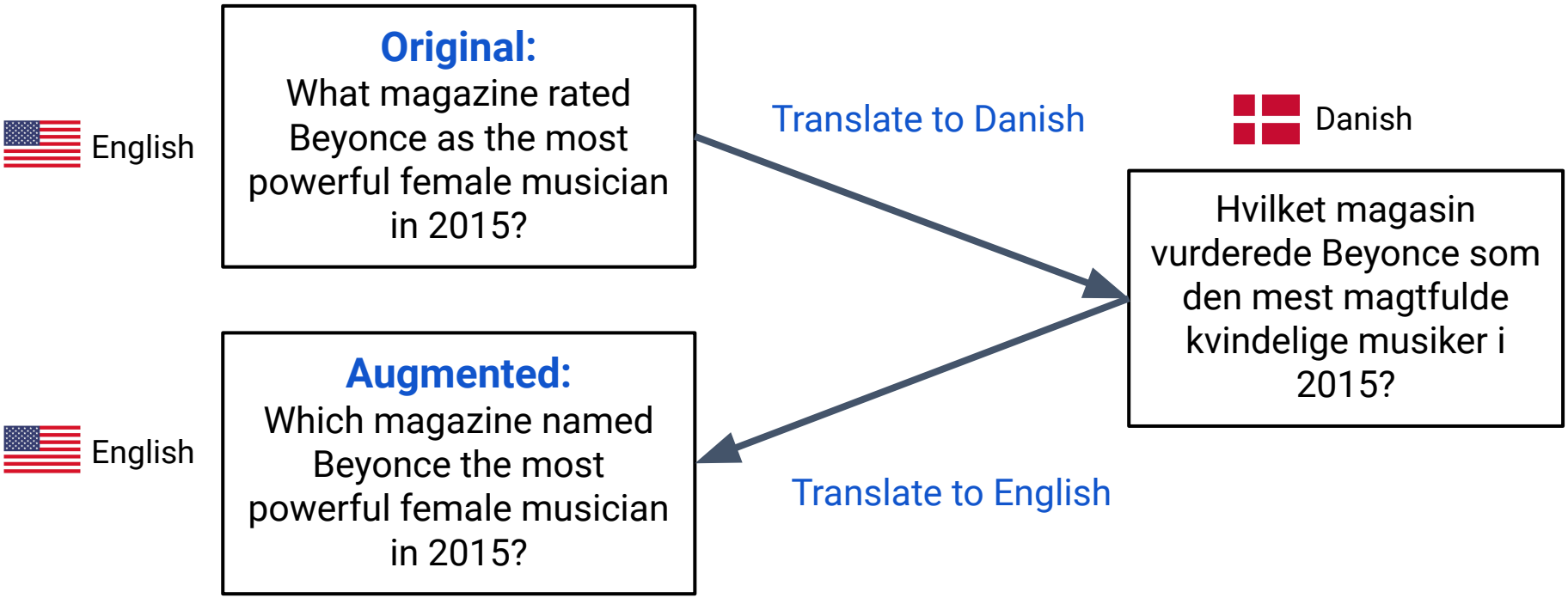
- Base
- Large
- Distillation

https://huggingface.co/models

Experiments

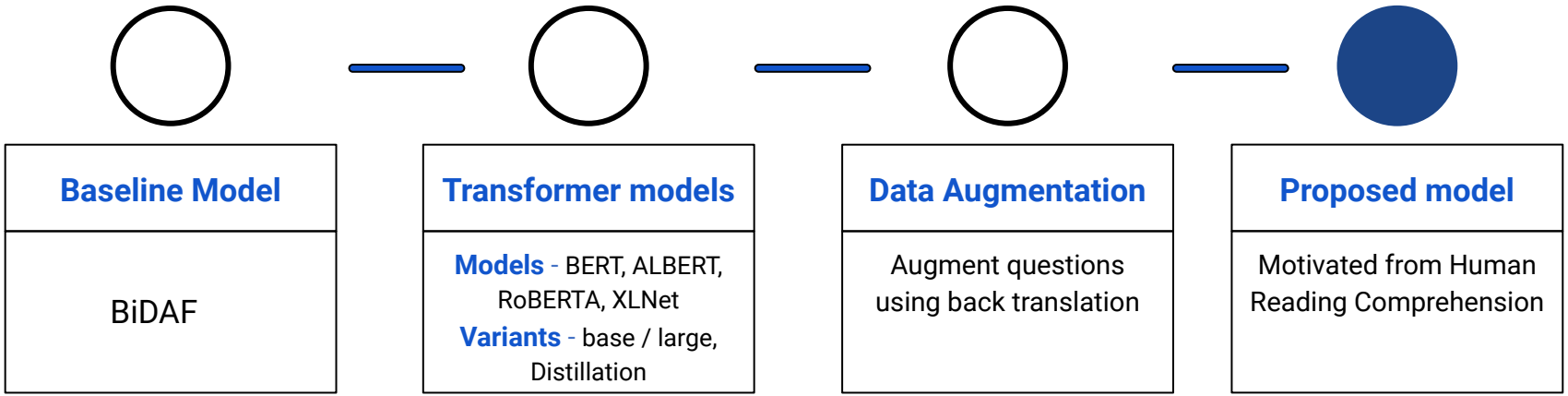


Data Augmentation using back translation



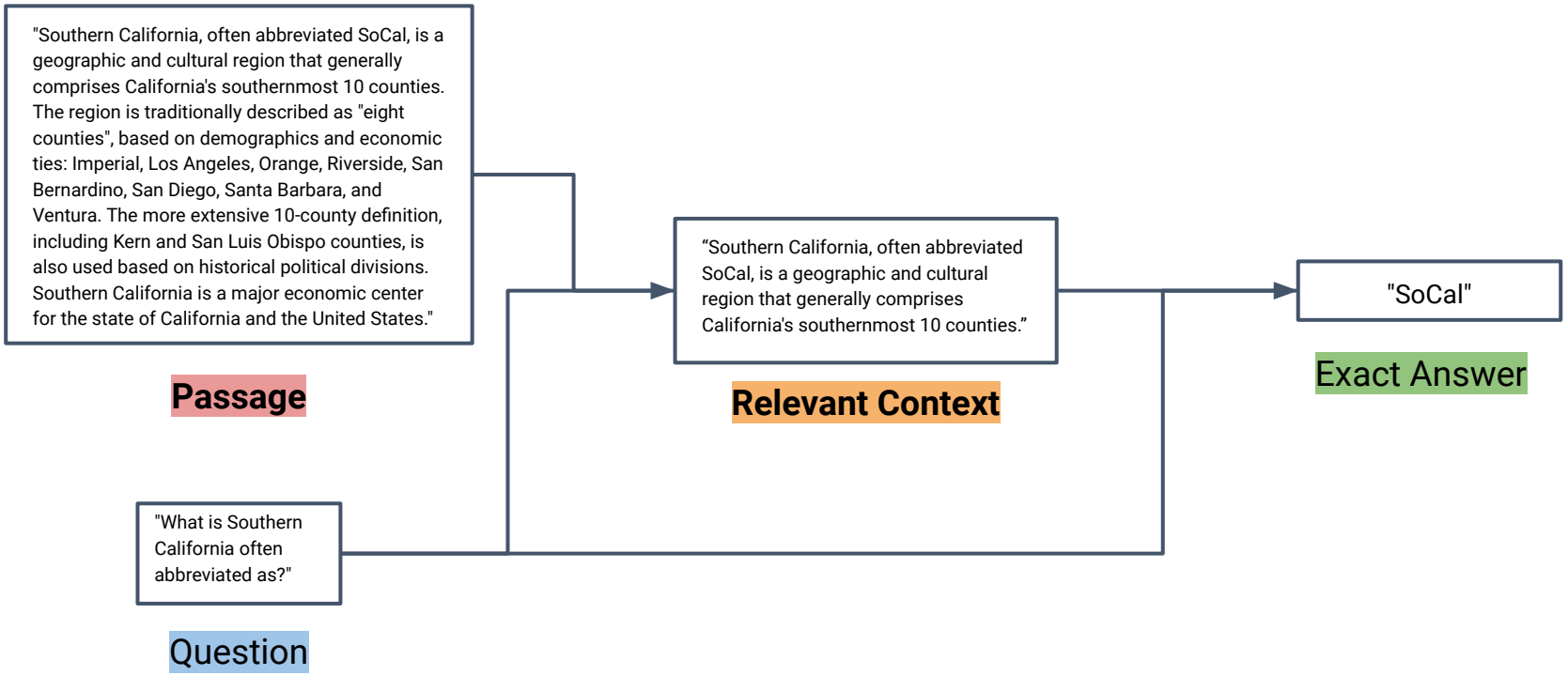
Key idea: help the network generalize to the syntactic variance in the question

Experiments

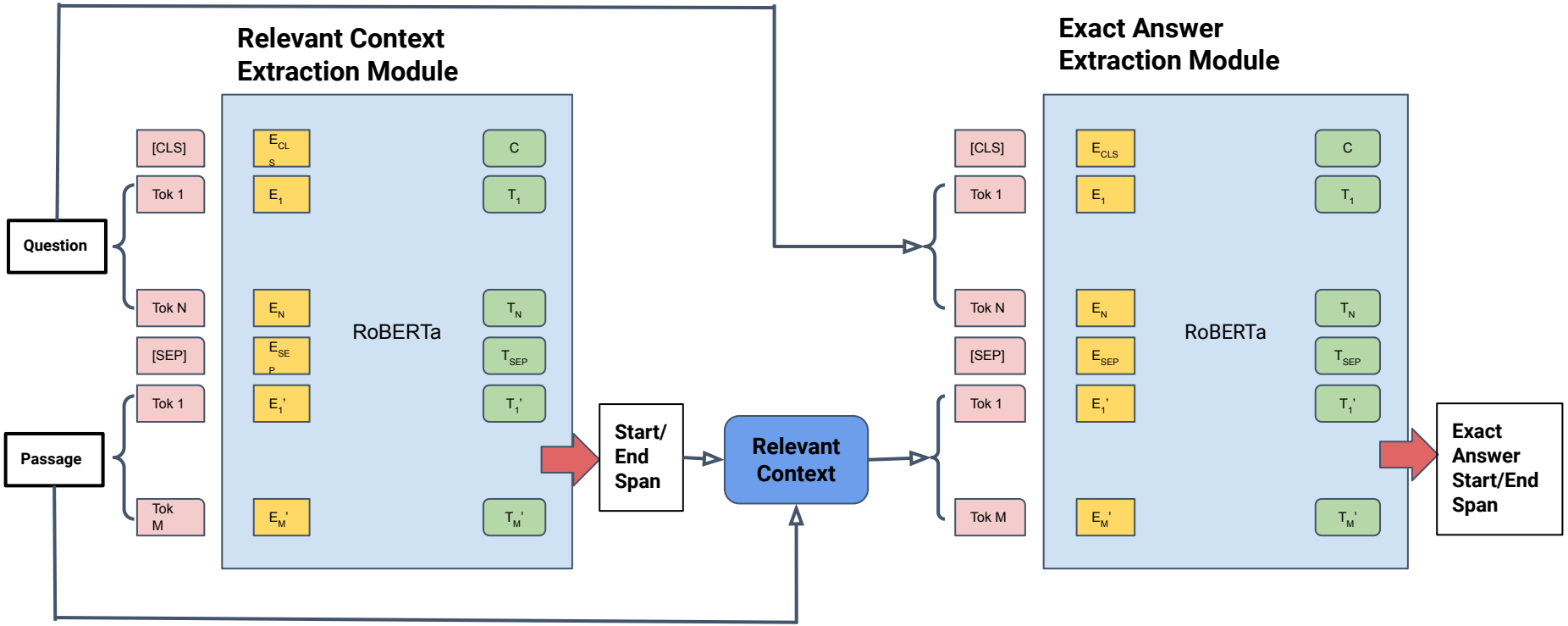


Proposed Model

Motivation - Human Reading Comprehension



Proposed Model Architecture



Evaluation Metrics

- **Exact Match (strict)**
  If the *characters* of the model's prediction exactly match the characters of (one of) the True Answer(s), EM = 1, otherwise EM = 0.

- **F1 Score (lenient)**

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Question: Where on Earth is free oxygen found?

True Answers: ['water', "in solution in the world's water bodies", "the world's water bodies"]

Prediction: water bodies

EM	F1
0	0.8

Overall EM and F1 scores are computed for a model by averaging over the individual example scores.

Results

Model	Variant	F1	EM	No. of parameters
BiDAF	-	60.86	57.44	< 1M
BERT	base large	80.3 83.9	77.3 80.8	108M 340M
ALBERT	base large	79.1 82.1	76.1 79.0	11M 17M
ROBERTA	base **Large** augmented	84.7 **87.1** 86.7	81.2 **83.9** 83.1	125M **355M** 355M
XLNet	base large	84.1 86.7	80.8 83.4	110M 340M
DistilBERT	base	77.4	76.2	66M
Our Model	RoBERTa base	79.6	76.3	250M

Pipeline Model Analysis

Model	F1	EM
Relevant Context Extraction	80.9	80
Exact Answer Extraction	95.6	91.9
End to End Pipeline	79.6	76.3

Details of the best performing model

Fine-tuned roberta large

'F1': 87.18

'Exact': 83.90

'HasAns_f1': 87.37

'HasAns_exact': 80.80

'NoAns_f1': 86.99

'NoAns_exact': 86.99

Human performance

'F1' : 89.45

'Exact' : 86.83

Model hyperparameters setting

- 24-layer, 1024-hidden, 16-heads, 355M parameters
- Training time: > 1 day (batch_size=8)
- GPU: **NVIDIA A100** (xgpg0 cluster node)

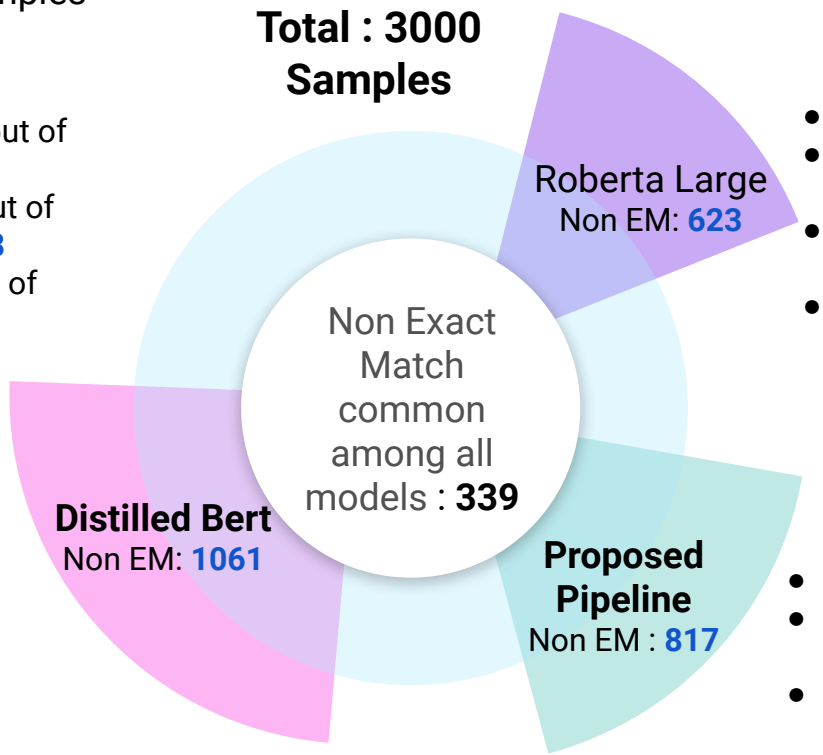

```
{
  "architectures": [
    "RobertaForQuestionAnswering"
  ],

  "num_epochs" :3,
  "learning_rate" : 5e-5,
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "eos_token_id": 2,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 1024,
  "initializer_range": 0.02,
  "intermediate_size": 4096,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "roberta",
  "num_attention_heads": 16,
  "num_hidden_layers": 24,
  "pad_token_id": 1,
  "type_vocab_size": 1,
  "vocab_size": 50265
}
```

Error Analysis

Common Non Matching Samples

- Dependency Distance: **7.02**
- Proportion of “What” questions out of 339 samples: **0.63**
- Proportion of “Why” questions out of total “Why” samples in data: **0.38**
- Proportion of DESC answers out of 339 samples : **0.76**



Roberta Large

- Dependency Distance: **6.33**
- Proportion of “What” questions not matching : **0.21**
- Proportion of “Why” questions not matching : **0.51**
- Proportion of DESC answers not matching : **0.256**

Distilled Bert

- Dependency Distance: **6.5**
- Proportion of “What” questions not matching : **0.37**
- Proportion of “Why” questions not matching : **0.66**
- Proportion of DESC answers not matching : **0.42**

Proposed Pipeline

- Dependency Distance: **6.36**
- Proportion of “What” questions not matching : **0.28**
- Proportion of “Why” questions not matching : **0.55**
- Proportion of DESC answers not matching : **0.32**

Error Analysis Example

Context

Historically, Victoria has been the base for the manufacturing plants of the major car brands Ford, Toyota and Holden; however, closure announcements by all three companies in the 21st century will mean that Australia will no longer be a base for the global car industry, with Toyota's statement in February 2014 outlining a closure year of 2017. Holden's announcement occurred in May 2013, followed by Ford's decision in December of the same year (Ford's Victorian plants—in Broadmeadows and Geelong—will close in October 2016).

What type of manufacturing plant is Victoria soon losing?

Question

Groundtruth

- Major car brands
- Car

Roberta: base

Distilled Bert:

Proposed Pipeline:

Predictions

Conclusion

- Transformer based models outperform the baseline BiDAF model.
- RoBERTa and XLNet produce comparable results, with highest F1 and EM score for their large variants.
- No improvement with Data Augmentation
- Our pipeline model which is based on ‘RoBERTa base’ produces comparable results to the ALBERT model, but is not able to outperform the original RoBERTa model.
- The F1 score of **Exact Answer Extraction Module** tested independently using ground-truth dev set is 95.6. This highlights potential in the pipeline architecture if first model can be improved.

Future Work

- Fine tune to improve the Relevant Context Extraction module of the pipeline architecture
- Experiment with Linguistic post-processing
- Ensemble different models
- Hyperparameter tuning

Challenges

- Relevant Context Extraction
  - Given the answer start, extracting relevant sentences from the context (tried nltk and spacy sentence tokenizer but it fails in some edge cases which were handled separately)

- Data Augmentation
  - Back translation: Limited resources, paid conversion tools - Google translation API

- Training time
  - CUDA out of memory

References

- Bidirectional attention flow for machine comprehension, Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi.
- Retrospective Reader for Machine Reading Comprehension, Zhuosheng Zhang and Junjie Yang and Hai Zhao, 2020
- Attention is all you need}, 2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,Lukasz Kaiser, and Illia Polosukhin.
- Ensemble ALBERT on SQuAD 2.0, https://arxiv.org/pdf/2110.09665.pdf}{https://arxiv.org/pdf/2110.09665.pdf
- Know What You Don't Know: Unanswerable Questions for SQuAD, https://arxiv.org/abs/1806.03822}{https://arxiv.org/abs/1806.03822
- Ensemble BERT with Data Augmentation and Linguistic Knowledge on SQuAD 2.0 https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15845024.pdf