Question Answering: SQuAD

Presented by: Minhah Saleem
ID: 2023120344

Question Answering

- Task of answering questions - SQuAD dataset
- Project pipeline:

- Implementation of Baseline DistilBERT:
  - Reference: https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/question_answering.ipynb#scrollTo=_trMUv7zfVce
- Data Augmentation:
  - Using Back-Translation (BT) and Synonym Replacement (SR)
  - Reference: Tell Me How to Ask Again: Question Data Augmentation with Controllable Rewriting in Continuous Space (Liu et al., EMNLP 2020)
- Data Sampling:
  - Active Learning: Selects challenging examples for data augmentation based on F1 scores, prioritizing diversity and usefulness over uniform sampling.
  - Reference: An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering (Longpre et al., 2019)
- Project Repository: Github

SQuAD v1 Dataset

```
{'train': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 117287
 }),
 'validation': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 13032
 }),
 'test': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 130319
 })}
```

- The answers are indicated by their start position in the text and their full text which is a substring of the context.

```
{'id': '5733be284776f41900661182',
 'title': 'University_of_Notre_Dame',
 'context': 'Architecturally, the school has a Catholic character. Atop the Main Building\'s gold dome is a golden statue of th
e Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with th
e legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica i
s the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary r
eputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects throu
gh 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.',
 'question': 'To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?',
 'answers': {'text': ['Saint Bernadette Soubirous'], 'answer_start': [515]}}
```

- Preprocessing is done by a Transformers Tokenizer which will tokenize the inputs.

```
tokenizer("To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?", "Saint Bernadette Soubirous")
```

```
{'input_ids': [101, 2000, 3183, 2106, 1996, 6261, 2984, 9382, 3711, 1999, 8517, 1999, 10223, 26371, 2605, 1029, 102, 3002, 1659
5, 9648, 4674, 2061, 12083, 9711, 2271, 102], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1]}
```

SQuAD v2 Dataset

```
{'train': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 117287
}),
'validation': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 13032
}),
'test': Dataset({
    features: ['id', 'title', 'context', 'question', 'answers'],
    num_rows: 11873
})}
```

- The answers are indicated by their start position in the text and their full text which is a substring of the context.

```
{'id': '56be85543aeaaa14008c9063',
 'title': 'Beyoncé',
 'context': 'Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny\'s Child. Managed by her father, Mathew Knowles, the group became one of the world\'s best-selling girl groups of all time. Their hiatus saw the release of Beyoncé\'s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".',
 'question': 'When did Beyonce start becoming popular?',
 'answers': {'text': ['in the late 1990s'], 'answer_start': [269]}}
```

- Preprocessing is done by a Transformers Tokenizer which will tokenize the inputs.

```
tokenizer("When did Beyonce start becoming popular?", "in the late 1990s")
```

```
{'input_ids': [101, 2043, 2106, 20773, 2707, 3352, 2759, 1029, 102, 1999, 1996, 2397, 4134, 102], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]}
```

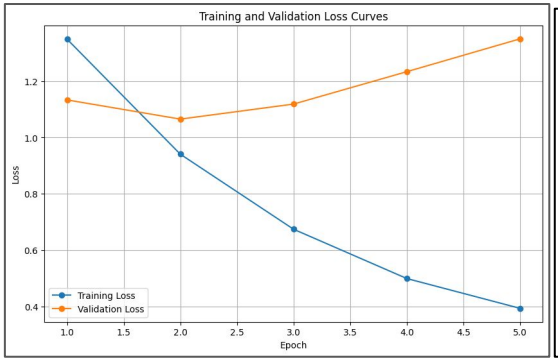
Implementation of Baseline (DistilBERT)

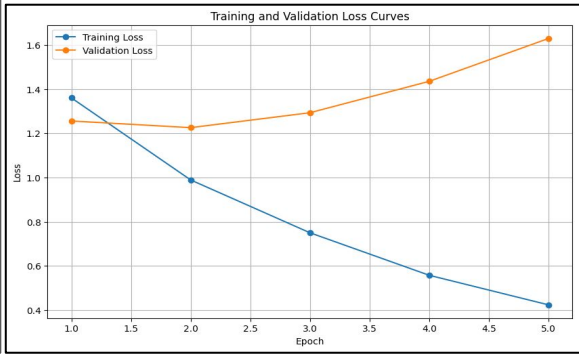
- Implementation of Baseline DistilBERT:
  - Reference: https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/question_answering.ipynb#scrollTo=_trMUv7zfVce
- Training Parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Evaluation

SQuAD_v1 | SQuAD_v2

```
{'exact_match': 75.89403973509934, 'f1': 84.62013581517354}
```

```
{'exact': 62.94112692664028,
 'f1': 66.68131681284449,
 'total': 11873,
 'HasAns_exact': 66.48110661268556,
 'HasAns_f1': 73.97221230075927,
 'HasAns_total': 5928,
 'NoAns_exact': 59.411269974768715,
 'NoAns_f1': 59.411269974768715,
 'NoAns_total': 5945,
 'best_exact': 62.94112692664028,
 'best_exact_thresh': 0.0,
 'best_f1': 66.68131681284478,
 'best_f1_thresh': 0.0}
```





SQuAD_v1 | SQuAD_v2

Data Augmentation

- Paraphrasing the questions:
  - Reference: Tell Me How to Ask Again: Question Data Augmentation with Controllable Rewriting in Continuous Space (Liu et al., EMNLP 2020)
- For paraphrasing:
  - Back Translation (BT):
    - English to French and then back to English translation using Google Translate API
    - Reference: https://github.com/mouuff/mtranslate
  - Synonym Replacement (SR):
    - Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random. Where n is 10% of total no. of words in the question.
    - Reference: https://github.com/jasonwei20/eda_nlp/tree/master

Original	Back Translation	Synonym Replacement
When did Beyonce start becoming popular?	When did Beyoncé start becoming popular?	when did beyonce start comme il faut popular?
In the Enuma Elish, what is creation seen as the union of?	In the Enuma Elish, what is creation considered to be the union of?	in the enuma elish what is creation look as the union of?
To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?	To whom did the Virgin Mary appear in 1858 in Lourdes in France?	to whom did the virgin virgin mary allegedly appear in in lourdes france?
How many bells are contained in the World War I Memorial Carillon?	How many bells does the First World War Memorial Carillon have?	how many bells are contain in the world war i memorial carillon?

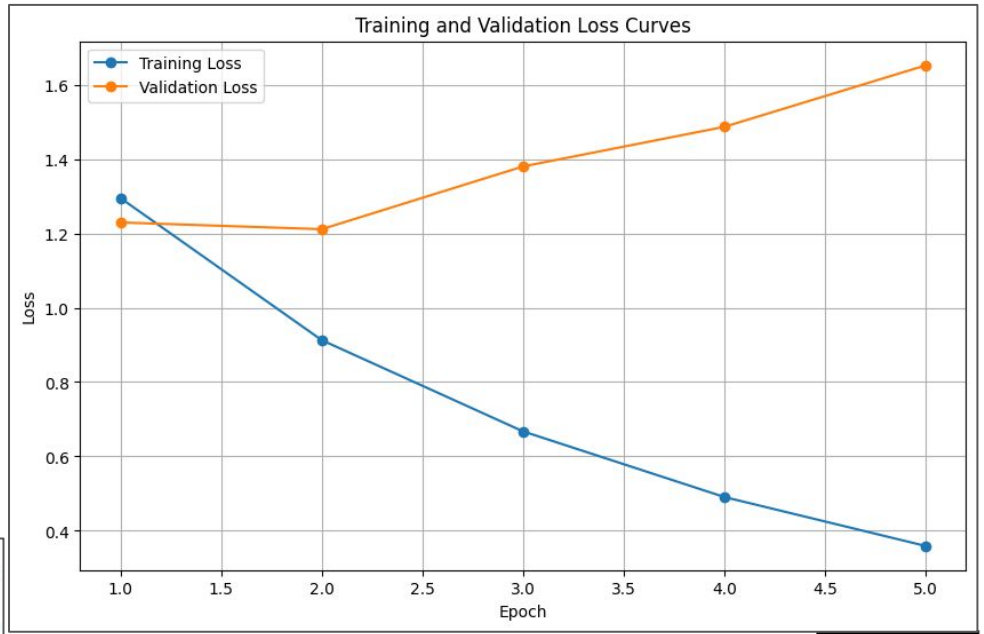
BT data augmentation (uniform sampling)

- DistilBERT for question answering as implemented in the HuggingFace transformers library.
- Data Split:

```
DatasetDict({
    train: Dataset({
        features: ['context', 'question', 'answers', 'id', 'title'],
        num_rows: 122287
    })
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 13032
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

- Training parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Results:
  - EM: 63.73283921502569
  - F1: 67.77227203680067

```
'total': 11873,
'HasAns_exact': 66.44736842105263,
'HasAns_f1': 74.53781813308572,
'HasAns_total': 5928,
'NoAns_exact': 61.026072329688816,
'NoAns_f1': 61.026072329688816,
'NoAns_total': 5945,
'best_exact': 63.73283921502569,
'best_exact_thresh': 0.0,
'best_f1': 67.77227203680091,
'best_f1_thresh': 0.0}
```



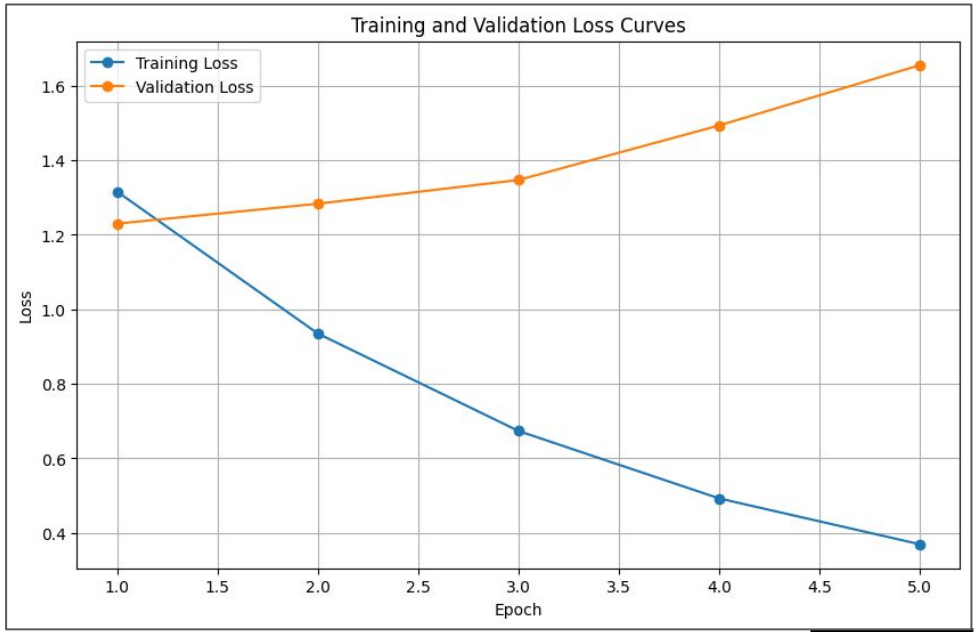
SR data augmentation (uniform sampling)

- DistilBERT for question answering as implemented in the HuggingFace transformers library.
- Data Split:

```
DatasetDict({
    train: Dataset({
        features: ['context', 'question', 'answers', 'id', 'title'],
        num_rows: 122287
    })
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 13032
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

- Training parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Results:
  - EM: 63.9686684073107
  - F1: 67.88932112106961

```
'total': 11873,
'HasAns_exact': 66.26180836707152,
'HasAns_f1': 74.1143572318584,
'HasAns_total': 5928,
'NoAns_exact': 61.682085786375104,
'NoAns_f1': 61.682085786375104,
'NoAns_total': 5945,
'best_exact': 63.9686684073107,
'best_exact_thresh': 0.0,
'best_f1': 67.88932112106988,
'best_f1_thresh': 0.0}
```



Data Sampling

- Reference: [An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering (Longpre et al., 2019)]
- The paper proposes an Active Learning data augmentation Technique:
  - Sampling more challenging examples for augmentation.
  - Metric for difficulty of an example : 1-F1 score
  - For each example, its probability weighting is derived from its F1 score. This weight replaces the uniform probability previously used to draw samples for query and context augmentations.
  - Three weighting strategies are devised to experiment with different distributions referred to as the hard, moderate and soft distributions.
  - Each distribution employs its own scoring function Sx, which is normalized across all examples to determine the probability of drawing that sample.

$$S(x) = \begin{cases} 1 - F1(x) + \epsilon & \text{Hard Score} \\ 2 - F1(x) & \text{Moderate Score} \\ 3 - F1(x) & \text{Soft Score} \end{cases}$$

$$P(x) = \frac{S(x)}{\Sigma_{i=1..n} S(i)}$$

epsilon = 0.01

- Using the sampling technique as described in the paper, each sampled example will be paraphrased.

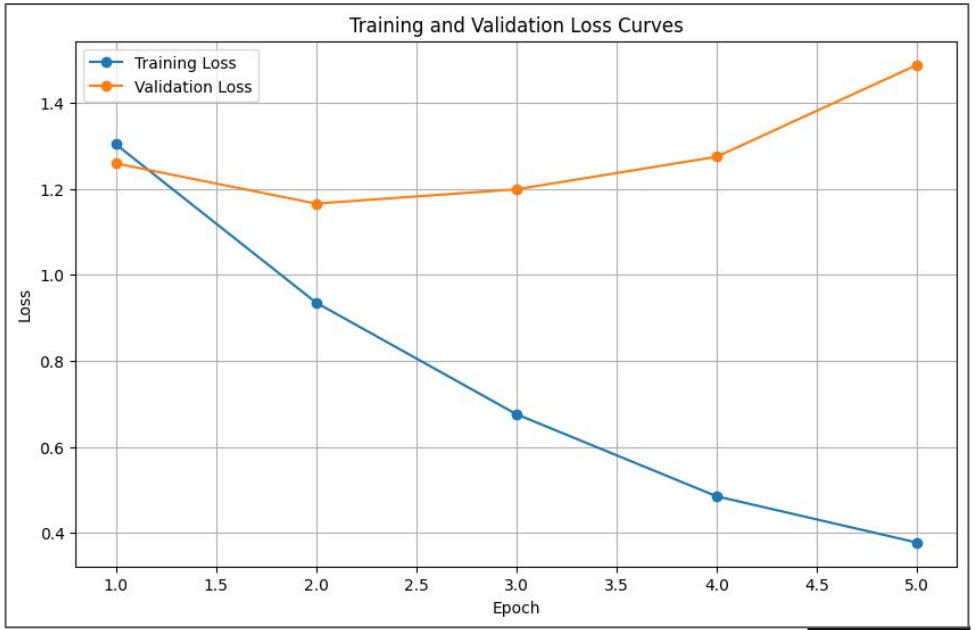
SR data augmentation (soft sampling)

- DistilBERT for question answering as implemented in the HuggingFace transformers library.
- Data Split:

```
DatasetDict({
    train: Dataset({
        features: ['context', 'question', 'answers', 'id', 'title'],
        num_rows: 122182
    })
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 13032
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

- Training parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Results:
  - EM: 63.65703697464836
  - F1: 67.592173589653

```
'total': 11873,
'HasAns_exact': 67.49325236167341,
'HasAns_f1': 75.37481056510593,
'HasAns_total': 5928,
'NoAns_exact': 59.83179142136249,
'NoAns_f1': 59.83179142136249,
'NoAns_total': 5945,
'best_exact': 63.65703697464836,
'best_exact_thresh': 0.0,
'best_f1': 67.59217358965324,
'best_f1_thresh': 0.0}
```



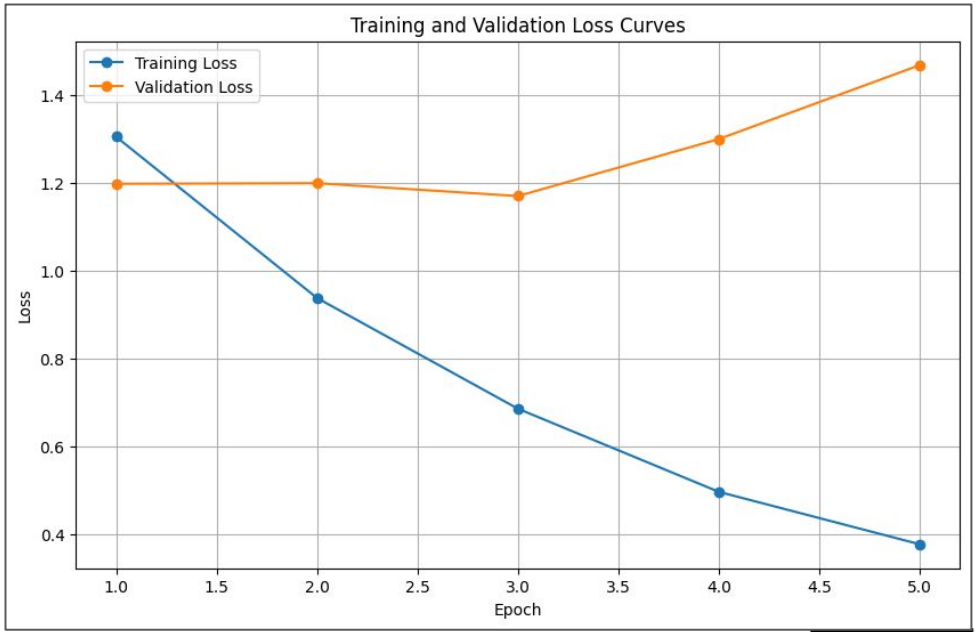
SR data augmentation (moderate sampling)

- DistilBERT for question answering as implemented in the HuggingFace transformers library.
- Data Split:

```
DatasetDict({
    train: Dataset({
        features: ['context', 'question', 'answers', 'id', 'title'],
        num_rows: 122194
    })
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 13032
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

- Training parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Results:
  - EM: 63.9686684073107
  - F1: 67.85471660720431

```
'total': 11873,
'HasAns_exact': 65.9919028340081,
'HasAns_f1': 73.77514343409833,
'HasAns_total': 5928,
'NoAns_exact': 61.951219512195124,
'NoAns_f1': 61.951219512195124,
'NoAns_total': 5945,
'best_exact': 63.9686684073107,
'best_exact_thresh': 0.0,
'best_f1': 67.85471660720454,
'best_f1_thresh': 0.0}
```



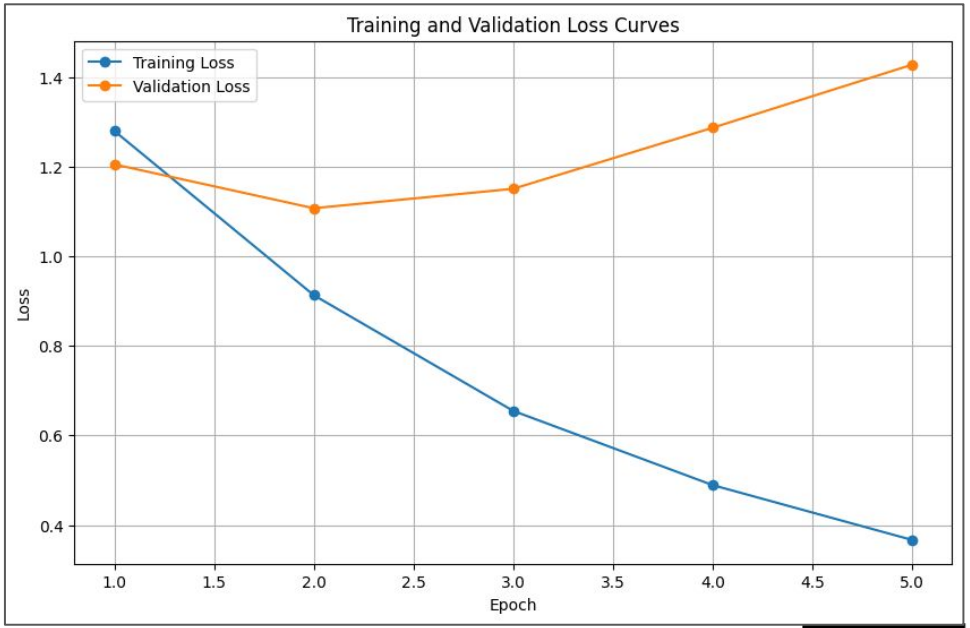
SR data augmentation (hard sampling)

- DistilBERT for question answering as implemented in the HuggingFace transformers library.
- Data Split:

```
DatasetDict({
    train: Dataset({
        features: ['context', 'question', 'answers', 'id', 'title'],
        num_rows: 122056
    })
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 13032
    })
    test: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

- Training parameters:
  - batch_size = 32
  - learning_rate=5e-5
  - num_train_epochs=5
- Results:
  - EM: 64.3139897245852
  - F1: 68.1786121049203

```
'total': 11873,
'HasAns_exact': 66.6497975708502,
'HasAns_f1': 74.39012508800883,
'HasAns_total': 5928,
'NoAns_exact': 61.98486122792262,
'NoAns_f1': 61.98486122792262,
'NoAns_total': 5945,
'best_exact': 64.3139897245852,
'best_exact_thresh': 0.0,
'best_f1': 68.17861210492057,
'best_f1_thresh': 0.0}
```



Results

Observations:

- The performance of BT and SR is comparable as:
  - Back Translation, in some cases, fails to introduce variation:
    - Eg. ‘When did Beyonce start becoming popular?’ Remains ‘When did Beyoncé start becoming popular?’
  - Synonym Replacement replaces the words without any consideration for context.
    - Eg. ‘In the Enuma Elish, what is creation seen as the union of?’ Becomes ‘in the enuma elish what is creation look as the union of?’
- The sampling techniques help improve the performance.

Technique	EM	F1
Baseline	62.94112693	66.68131681
BT	63.73283922	67.77227204
SR	63.96866841	67.88932112
SR+soft sampling	63.65703697	67.59217359
SR+moderate sampling	63.96866841	67.85471661
SR+hard sampling	**64.31398972**	**68.1786121**

Results


```
{'id': '56ddde6b9a695914005b962c',
 'title': 'Normans',
 'context': 'The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centur
ies gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders an
d pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Fr
ancia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants
would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Nor
mans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.',
 'question': 'What century did the Normans first gain their separate identity?',
 'answers': {'text': ['10th century',
   'the first half of the 10th century',
   '10th',
   '10th'],
  'answer_start': [671, 649, 671, 671]}}
```


Baseline: '10th'

BT: '10th'

SR: '10th'

SR+Soft: '10th'

SR+Moderate: '10th'

SR+Hard: '10th'

Results


```
{'id': '5729f60caf94a219006aa6f0',
 'title': 'Economic_inequality',
 'context': 'Effects of inequality researchers have found include higher rates of health and social problems, and lower rates o
f social goods, a lower level of economic utility in society from resources devoted on high-end consumption, and even a lower l
evel of economic growth when human capital is neglected for high-end consumption. For the top 21 industrialised countries, coun
ting each person equally, life expectancy is lower in more unequal countries (r = -.907). A similar relationship exists among U
S states (r = -.620).',
 'question': 'Why does a lower level of economic growth occur due to high-end consumption?',
 'answers': {'text': ['human capital is neglected',
   'a lower level of economic utility in society',
   'human capital is neglected'],
  'answer_start': [270, 130, 270]}}
```


Baseline: 'human capital is neglected for high-end consumption'

BT: 'human capital'

SR: 'human capital is neglected for high-end consumption'

SR+soft: 'human capital is neglected for high-end consumption'

SR+Moderate:'human capital is neglected'

SR+Hard: 'human capital is neglected'

Results


```
{'id': '570d2af0fed7b91900d45c96',
 'title': 'Victoria_(Australia)',
 'context': "Victoria (abbreviated as Vic) is a state in the south-east of Australia. Victoria is Australia's most densely popu
lated state and its second-most populous state overall. Most of its population is concentrated in the area surrounding Port Phi
llip Bay, which includes the metropolitan area of its capital and largest city, Melbourne, which is Australia's second-largest
city. Geographically the smallest state on the Australian mainland, Victoria is bordered by Bass Strait and Tasmania to the sou
th,[note 1] New South Wales to the north, the Tasman Sea to the east, and South Australia to the west.",
 'question': 'What city is the capital of Victoria?',
 'answers': {'text': ['Melbourne', 'Melbourne', 'Melbourne'],
  'answer_start': [321, 321, 321]}}
```


Baseline: 'Port Phillip Bay'

BT: 'Melbourne'

SR: 'Melbourne'

SR+soft: 'Melbourne'

SR+Moderate:'Melbourne'

SR+Hard:'Melbourne'

Results


```
{'id': '5ad3ed26604f3c001a3ff79c',
 'title': 'Normans',
 'context': "Robert Guiscard, an other Norman adventurer previously elevated to the dignity of count of Apulia as the result of
his military successes, ultimately drove the Byzantines out of southern Italy. Having obtained the consent of pope Gregory VII
and acting as his vassal, Robert continued his campaign conquering the Balkan peninsula as a foothold for western feudal lords
and the Catholic Church. After allying himself with Croatia and the Catholic cities of Dalmatia, in 1081 he led an army of 30,0
00 men in 300 ships landing on the southern shores of Albania, capturing Valona, Kanina, Jericho (Orikumi), and reaching Butrin
t after numerous pillages. They joined the fleet that had previously conquered Corfu and attacked Dyrrachium from land and sea,
devastating everything along the way. Under these harsh circumstances, the locals accepted the call of emperor Alexius I Comnen
us to join forces with the Byzantines against the Normans. The Albanian forces could not take part in the ensuing battle becaus
e it had started before their arrival. Immediately before the battle, the Venetian fleet had secured a victory in the coast sur
rounding the city. Forced to retreat, Alexius ceded the command to a high Albanian official named Comiscortes in the service of
Byzantium. The city's garrison resisted until February 1082, when Dyrrachium was betrayed to the Normans by the Venetian and Am
alfitan merchants who had settled there. The Normans were now free to penetrate into the hinterland; they took Ioannina and som
e minor cities in southwestern Macedonia and Thessaly before appearing at the gates of Thessalonica. Dissension among the high
ranks coerced the Normans to retreat to Italy. They lost Dyrrachium, Valona, and Butrint in 1085, after the death of Robert.",
 'question': 'How many men did Roberts army face?',
 'answers': {'text': [], 'answer_start': []}}
```


Baseline: '30,000'

BT: '30,000'

SR: '30,000'

SR+soft: '30,000'

SR+Moderate: '30,000'

SR+Hard: ''

Results


```
{'id': '5ad3e96b604f3c001a3ff689',
 'title': 'Normans',
 'context': 'Some Normans joined Turkish forces to aid in the destruction of the Armenians vassal-states of Sassoun and Taron i
n far eastern Anatolia. Later, many took up service with the Armenian state further south in Cilicia and the Taurus Mountains.
A Norman named Oursel led a force of "Franks" into the upper Euphrates valley in northern Syria. From 1073 to 1074, 8,000 of th
e 20,000 troops of the Armenian general Philaretus Brachamius were Normans—formerly of Oursel—led by Raimbaud. They even lent t
heir ethnicity to the name of their castle: Afranji, meaning "Franks." The known trade between Amalfi and Antioch and between B
ari and Tarsus may be related to the presence of Italo-Normans in those cities while Amalfi and Bari were under Norman rule in
Italy.',
 'question': 'Who joined Norman forces in the destruction of the Armenians?',
 'answers': {'text': [], 'answer_start': []}}
```


Baseline: 'Some Normans'

BT: 'Some Normans'

SR: 'Oursel'

SR+soft: 'Oursel'

SR+Moderate: 'Some Normans'

SR+Hard: 'Some Normans'

Results


```
{'id': '571ce5055efbb31900334e2d',
 'title': 'Oxygen',
 'context': "Breathing pure O\n2 in space applications, such as in some modern space suits, or in early spacecraft such as Apol
lo, causes no damage due to the low total pressures used. In the case of spacesuits, the O\n2 partial pressure in the breathing
gas is, in general, about 30 kPa (1.4 times normal), and the resulting O\n2 partial pressure in the astronaut's arterial blood
is only marginally more than normal sea-level O\n2 partial pressure (for more information on this, see space suit and arterial
blood gas).",
 'question': 'What si the comparison to sea level with the oxygen level in space suits?',
 'answers': {'text': ['only marginally more',
   'marginally more than normal',
   'marginally more',
   'marginally more',
   'only marginally more'],
  'answer_start': [369, 374, 374, 374, 369]}}
```


Baseline: "

BT: "

SR: "

SR+soft: "

SR+Moderate:"

SR+Hard: "

Conclusion and Future Work

- While Back-Translation is relatively slower, Synonym Replacement is faster and offers comparable performance. This also allows us to have a greater sampling size.
- The improvement, as we go from soft to hard score sampling, proves that emphasizing on difficult examples during data augmentation is more beneficial.
- Future Work:
  - More data size of augmentation
  - Question Data Augmentation with Controllable Rewriting in Continuous Space
  - Use of Chat GPT
  - Experimentation with other sampling techniques mentioned in An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering (Longpre et al., 2019)

Thanks!