Automatic Question Generation

June 22nd, 2020

Authors
Sasi Kiran Gaddipati
Desiana Dien Nurchalifah

Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Introduction

Motivation

Input: Her fourth studio album 4 was released on **June 28, 2011** in the us. 4 sold 310,000 copies in its first week and debuted atop the billboard 200 chart, giving Beyoncé her fourth consecutive number - one album in the US. the album was

....

....

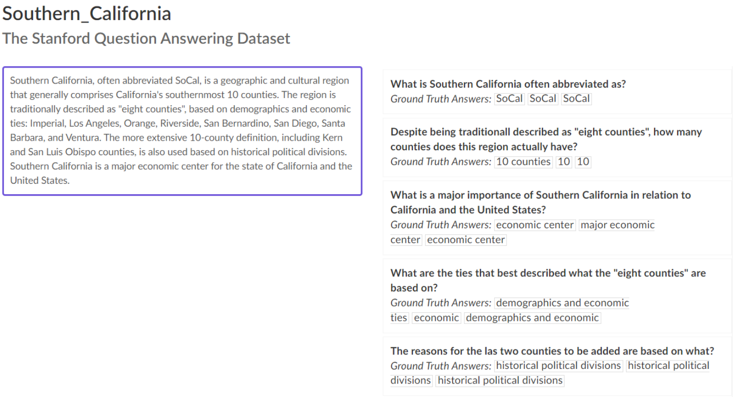
....

she took the stage at New York's Roseland ballroom for four nights of special performances: the 4 intimate nights with Beyoncé concerts saw the performance of her 4 albums to a standing room only.

Target: When was Beyoncé's fourth album released?

Dataset

Stanford Question Answering Dataset (SQuAD 2.0)



Southern_California

The Stanford Question Answering Dataset

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal | SoCal | SoCal

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties | 10 | 10

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center | major economic center | economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties | economic | demographics and economic

The reasons for the las two counties to be added are based on what?
Ground Truth Answers: historical political divisions | historical political divisions | historical political divisions

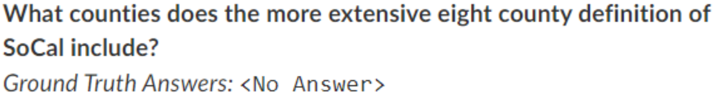
Figure 1: An excerpt from SQuAD 2.0[1]

[1] https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/Southern_California.html

Stanford Question Answering Dataset (SQuAD 2.0)

- Comprises comprehensions, questions and answers [1]

- Contains unanswerable questions



What counties does the more extensive eight county definition of SoCal include?

Ground Truth Answers: `<No Answer>`

Figure 2: An example for the question having no answer in the dataset[2]

- Training and validation data are publicly available

- Around 500+ articles and 100,000+ questions in training data

[2] https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/Southern_California.html

Experimentation

Preprocessing

- Case normalization and case detection
- Tokenization
- Named entity recognition (NER)
- POS-Tagging
- IOB-Tagging

Preprocessing

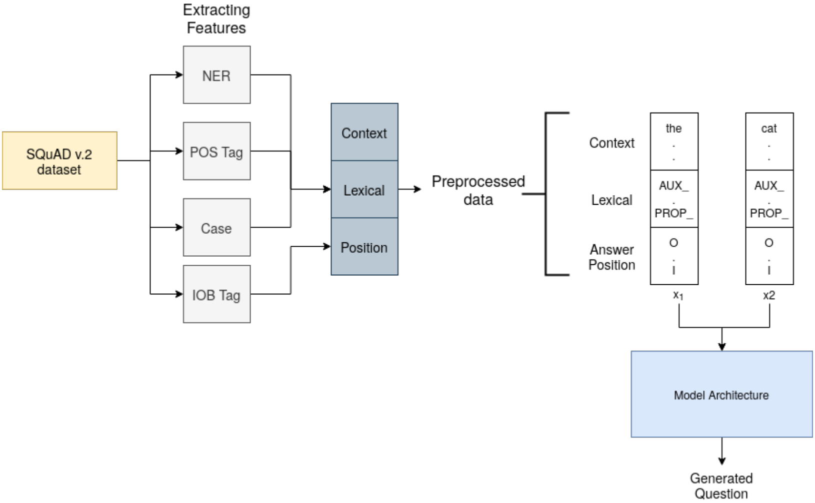


Figure 3: Preprocessing

Model Architecture

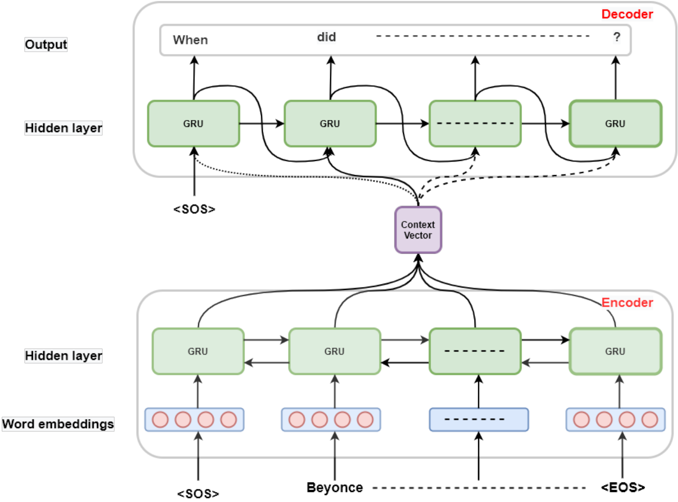


Figure 4: Seq2Seq attention architecture

Training

- Trained on 130,319 input and output pairs
- Vocabulary: 70,859
- Teacher forcing ratio: 0.5
- **Hyper-parameters:**
  - Epochs: 100
  - Batch size: 128
  - Optimizer: Adam
  - Learning rate: 0.001
  - Embedding size: 300
  - Hidden size : 512

Validation and Testing

- Split 'dev' data into 70% testing and 30% validation data

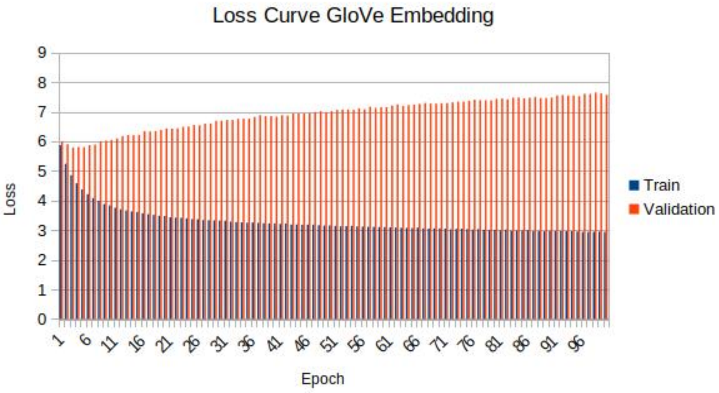
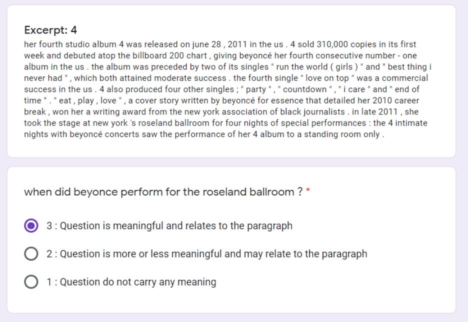


Figure 5: Loss Curve GloVe

Metrics

Human Evaluation

- Randomly generated 25 outputs are shown to humans for evaluation [2]



Excerpt: 4

her fourth studio album 4 was released on june 28 , 2011 in the us . 4 sold 310,000 copies in its first week and debuted atop the billboard 200 chart , giving beyoncé her fourth consecutive number - one album in the us . the album was preceded by two of its singles " run the world (girls) " and " best thing i never had " , which both attained moderate success . the fourth single " love on top " was a commercial success in the us . 4 also produced four other singles ; " party " , " countdown " , " i care " and " end of time " . " eat , play , love " , a cover story written by beyoncé for essence that detailed her 2010 career break , won her a writing award from the new york association of black journalists . in late 2011 , she took the stage at new york 's roseland ballroom for four nights of special performances : the 4 intimate nights with beyoncé concerts saw the performance of her 4 album to a standing room only .

when did beyonce perform for the roseland ballroom ? *

● 3 : Question is meaningful and relates to the paragraph

○ 2 : Question is more or less meaningful and may relate to the paragraph

○ 1 : Question do not carry any meaning

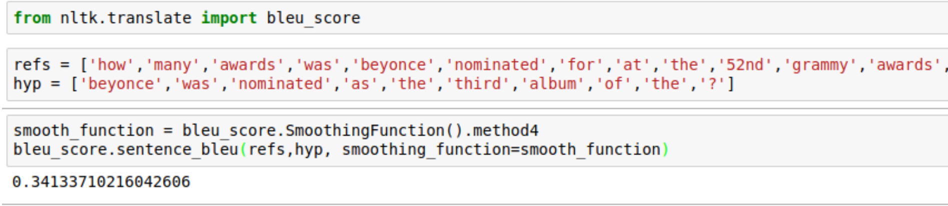
Figure 6: An example of human evaluation for single question[3]

- Calculate the mean of allotted scores by different individuals
- Compute correspondability between the evaluators with the Fleiss' kappa score [3]

[3] https://forms.gle/TFLSjGU843XSTac39

BLEU Score

- N-gram based matching evaluation [4]
- Fraction n-gram of generated hypothesis matches n-gram in the reference hypothesis
- Final score: Geometric mean of n-gram precision obtained from the parameter n in the range of 1 to N (N is typically 4)



```
from nltk.translate import bleu_score
```

```
refs = ['how','many','awards','was','beyonce','nominated','for','at','the','52nd','grammy','awards',
hyp = ['beyonce','was','nominated','as','the','third','album','of','the','?']
```

```
smooth_function = bleu_score.SmoothingFunction().method4
bleu_score.sentence_bleu(refs,hyp, smoothing_function=smooth_function)
```

```
0.34133710216042606
```

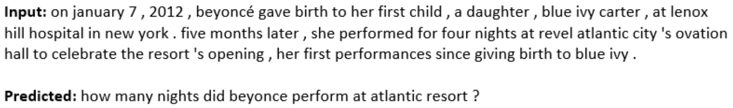
Figure 7: BLEU Example Calculation

Meteor Score

- Metric for Evaluation of Translation with Explicit ORdering[5].
- Fraction of hypothesis that matches the reference (precision) and computes the fraction of reference that matches hypothesis (recall) by matches as follows:
  - Stemmed words
  - Synonyms
  - Paraphrases
  - Different weight values for function and content words
- Final score: Harmonic mean of precision and recall from all the matches

Results

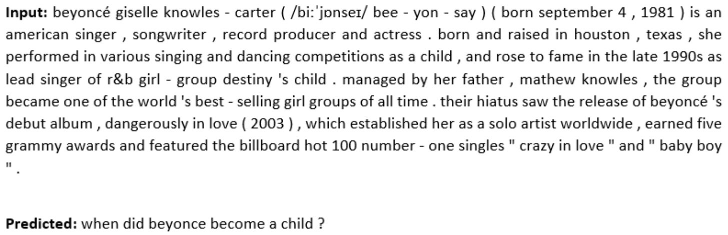
Qualitative Results I



Input: on january 7 , 2012 , beyoncé gave birth to her first child , a daughter , blue ivy carter , at lenox hill hospital in new york . five months later , she performed for four nights at revel atlantic city 's ovation hall to celebrate the resort 's opening , her first performances since giving birth to blue ivy .

Predicted: how many nights did beyonce perform at atlantic resort ?

Figure 8: An example of model predicting correct output

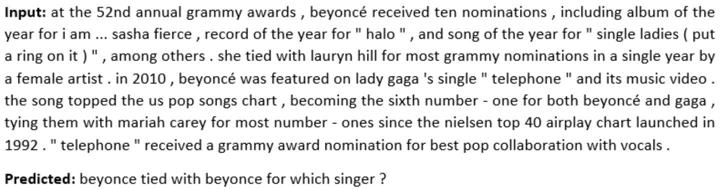


Input: beyoncé giselle knowles - carter (/bi:ˈjɒnseɪ/ bee - yon - say) (born september 4 , 1981) is an american singer , songwriter , record producer and actress . born and raised in houston , texas , she performed in various singing and dancing competitions as a child , and rose to fame in the late 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 " .

Predicted: when did beyonce become a child ?

Figure 9: An example of model predicting partially correct output

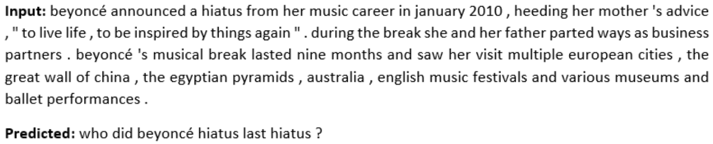
Qualitative Results II



Input: at the 52nd annual grammy awards , beyoncé received ten nominations , including album of the year for i am ... sasha fierce , record of the year for " halo " , and song of the year for " single ladies (put a ring on it) " , among others . she tied with lauryn hill for most grammy nominations in a single year by a female artist . in 2010 , beyoncé was featured on lady gaga 's single " telephone " and its music video . the song topped the us pop songs chart , becoming the sixth number - one for both beyoncé and gaga , tying them with mariah carey for most number - ones since the nielsen top 40 airplay chart launched in 1992 . " telephone " received a grammy award nomination for best pop collaboration with vocals .

Predicted: beyonce tied with beyonce for which singer ?

Figure 10: An example of model predicting wrong output



Input: beyoncé announced a hiatus from her music career in january 2010 , heeding her mother 's advice , " to live life , to be inspired by things again " . during the break she and her father parted ways as business partners . beyoncé 's musical break lasted nine months and saw her visit multiple european cities , the great wall of china , the egyptian pyramids , australia , english music festivals and various museums and ballet performances .

Predicted: who did beyoncé hiatus last hiatus ?

Figure 11: An example of model predicting wrong output

Quantitative Results

- **Human Evaluation:**
  - Mean score: 1.750 (Greater than [2])
  - Fleiss' Kappa Score: 0.238 (Fair agreement)
- **Automatic Evaluation**

Evaluation	GloVe (%)	ConceptNet-Numberbatch (%)
BLEU	2.16	1.02
METEOR	18.19	17.92

Conclusion

Key Takeaways

- Preprocessing improves the model's ability to predict related and correct outputs

- Removal of stopwords does not help the performance of question generation

- Automatic evaluation metrics (BLEU and METEOR) are sub-optimal for language modeling tasks

- With the current parameter, model has already started to over-fit training from 10th epoch

- Gradient clipping does not impact much on model performance

- , "... teacher forcing can result in problems in generation as small prediction error compound in the conditioning context." [6], proven in this model that one wrong prediction leads to misguided result

Future Work

- Extended approach for multiple questions
- Implement beam search procedure to select overall best next predicted sequence
- Improving the predicting ability of the language model
- Evaluation on domain-specific scenarios
- Compatible to real time scenario

Questions??

References I

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD". In: arXiv preprint arXiv:1806.03822 (2018).

[2] Qingyu Zhou et al. "Neural question generation from text: A preliminary study". In: National CCF Conference on Natural Language Processing and Chinese Computing. Springer. 2017, pp. 662–671.

[3] Joseph L Fleiss. "Measuring nominal scale agreement among many raters.". In: Psychological bulletin 76.5 (1971), p. 378.

[4] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002, pp. 311–318.

[5] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005, pp. 65–72.

[6] Alex Lamb et al. Professor Forcing: A New Algorithm for Training Recurrent Networks. 2016. arXiv: 1610.09038 [stat.ML].