Recep Oğuz Araz
Prof. Engin Erzin
ELEC 390
19.01.2021

Report IX: Language Model Integration

Our Network outputs a probability matrix, containing the probabilities of observing each character during each time step. We decode this matrix and obtain a transcription.

Decoding

1. **Max Decoding**

   We simply take the character with the highest probability at each time step.

   It takes 22 seconds in total to decode the complete METUbet test set. Avg_CER = 0.227

2. **Beam Search Decoding**

   We keep track of N paths with the highest probability at each decoding step, and in the last step choose the path with max probability. (Usually very close probabilities)

   For a beam width of 5, it takes ≈ 500 seconds in total to decode the complete METUbet test set. Avg_CER = 0.227

Although the CER is near satisfactory, we observe a large WER. To correct the spelling mistakes of our Speech2Text model, we integrate a Language Model (LM) following the Neural Network. We could integrate an N-gram LM, but due to the unavailability for Turkish, we try to make use of the publicly available BERT LM.

BERTurk

Pre-trained transformer with vocabulary size 32k and 128k.
128.000 tokens = 111.000 words, 15000 suffixes, emojis, letters, punctuations, numbers.

- "henüz", "henuz" both included.
- "belirlenmedi" included, "belirlenmesi" *not* included
- some Non-Turkish roots: "air"

We compare these tokens with 3 different datasets, Mozilla CommonVoice, METUbet and their union.

Dataset Vocabulary	#Intersecting Words	% Words in DS2 Training Set
32k-uncased ⋂ METUbet	4036	51%
32k-uncased ⋂ CommonVoice	5012	56%
32k-uncased ⋂ Union	7267	48%
128k-uncased ⋂ METUbet	6143	77%
128k-uncased ⋂ CommonVoice	7331	82%
128k-uncased ⋂ Union	11500	77 %

When 128k-uncased is used, the coverage increases but the probability of observing each token decreases.

Recep Oğuz Araz
Prof. Engin Erzin
ELEC 390
19.01.2021

Spell Checking with Language Models

We can integrate a LM in one of two ways,

1) During the decoding process,
2) After the decoding is done.

1. Prefix Beam Search

We use the beam search algorithm as before, but whenever we decide to output a space character, we enforce a LM on the decoded words until now.

Problems:

- Python implementation is slow.
- Requires n-gram LM with huge vocabulary.
    - BERT is not a next word predictor.
- Frequent underflow. *(Conversion to Log Space is challenging)*

2. NLMs for Spell Checking

We search for ways to implement the new champion of NLP, BERT for improving the WER.

2 proposed algorithms for using BERT as a spell checker.

Algorithm1:

Following the words from left to right, for each word, we try to replace the word with the closest Levenshtein token between N contextually closest tokens.

```
Original sentence:
Hükümetin değişmesi halinde müzakereler ve katılım süreci ne yönde etkilenir?

Masked Sentence:
Hükümetin değişmesi halinde müzakereler [MASK] katılım süreci ne yönde etkilenir?

Input Tokens:
['[CLS]', 'huk', '##umet', '##in', 'degis', '##mesi', 'halinde', 'muz', '##aker', '##eler', '[MASK]', 'katılım', 'surec', '##
i', 'ne', 'yon', '##de', 'etkilenir', '?', '[SEP]']

Tokens, Corresponding Errors and Relative Percentage Errors:

ve                  (0.0, 2) |        0.0
,                   (2.0, 2) |      100.0
ile                 (2.0, 2) |      100.0
-                   (2.0, 2) |      100.0
de                  (1.0, 2) |       50.0
/                   (2.0, 2) |      100.0
veya                (2.0, 2) |      100.0
konusunda           (9.0, 2) |      450.0
dahil               (5.0, 2) |      250.0
icin                (4.0, 2) |      200.0
sonrası             (7.0, 2) |      350.0
yani                (4.0, 2) |      200.0
sonrasında          (10.0, 2) |      500.0
sonucu              (6.0, 2) |      300.0
dolayısıyla         (11.0, 2) |      550.0
baslayan            (8.0, 2) |      400.0
yoluyla             (7.0, 2) |      350.0
acısından           (9.0, 2) |      450.0
+                   (2.0, 2) |      100.0
sonucunda           (9.0, 2) |      450.0
```

Recep Oğuz Araz
Prof. Engin Erzin
ELEC 390
19.01.2021

Algorithm2:

Following the words from left to right, for each word, we find N closest tokens w.r.t the Levenshtein distance and try to replace it with the closest contextual representation.

```
Original Sentence:
Ancak Gruevski iki öneriyi geri çevirdi.

Misspelled Sentence:
Abcad Gruevskl ioö öneriyv gericçevirhi.
----------------------------------------
Output Sentence:
ancak Gruevskl ioö öneriyv gericçevirhi.

Output Sentence:
ancak gruevski ioö öneriyv gericçevirhi.

Output Sentence:
ancak gruevski ilk öneriyv gericçevirhi.

Output Sentence:
ancak gruevski ilk öneriye gericçevirhi.
```

```
Original Sentence:
Calasan yetenekli çocuk bursu da aldı.

Misspelled Sentence:
Calasan yatanekli çocuk barau da alaı.
----------------------------------------
Output Sentence:
dalaman yatanekli çocuk barau da alaı.

Output Sentence:
dalaman yetenekli çocuk barau da alaı.

Output Sentence:
dalaman yetenekli çocuk bursu da alaı.

Output Sentence:
dalaman yetenekli çocuk bursu da aynı
```

Challenges:

i) How to identify misspelled words?
- Vocabulary is not perfect.
  - Turkish vs ASCII characters:
    - Both versions are in the vocabulary
    - Increases Levenshtein distance.
  - Proper nouns

ii) How to perform true correction?
- When to decide correction is desirable?

iii) Enormous run time.

Test Performance:

Unsatisfactory.

```
Target: manipülasyon lafları sayesinde ergun kurtuldu.
Predicted: mayıplasyon dafleri sayisnde erdun kurtuldu.
LM Corrected: mayıplasyon defteri sayesinde evden kurtuldu..

Target: çocuk sen kimseyi değil kendini kandırıyorsun.
Predicted: şocokksan kimseği değiz kendini kandbırıyorsun.
LM Corrected: şocokksan kimse degil kendini kandbırıyorsun. ..

Target: onları mutlu olduğu dönemden iyi hatırlıyordu.
Predicted: onları muluolduğun dönemden iyi atırbiy ordu.
LM Corrected: onları muluolduğun dönemden iyi terbiy ordu..
------------------------------------------------------------
Batch time: 24.18minutes
Total time: 462.27minutes
------------------------------------------------------------
Target: havada dönen bir sıçramayla çarka oturabilirdi.
Predicted: avoda doren birsıçramayla çarkolturabilirdı.
LM Corrected: oda dort birsıçramayla çarkolturabilirdı..

Target: kalaylı alaşımdan yapılan eşyalar güzel oluyor.
Predicted: önce dış çizglri şi sona ileidol..
LM Corrected: önce dış çizgisi şi sona ileidol...

Target: burnunu parmaklarının arasında ustaca sümkürdü.
Predicted: bulbunu parmakları nlarasında l ustuadu sünkürdü.
LM Corrected: bulun parmakları arasında l ustunde söndürdü.
```