Week 1 Assignment

Generative AI

Saarland University – Winter Semester 2024/25

Martínez
7057573
cama00005@stud.uni-saarland.de

1 Exercise Assignment: E1

We can express an upper bound on the number of parameters required for an n-gram model as follows: the number of possible 2-grams is V^2, and the number of possible 4-grams is V^4 [1]. Thus, the number of parameters required for an n-gram model is V^n. This is because the model needs to store the probabilities of all possible n-grams, and the number of possible n-grams is V^n, where V is the size of the vocabulary.

2 Exercise Assignment: E2

Following the above, for a 1-grams with $V = 50{,}000$, the number of parameters required is $V^n = 50{,}000^1 = 50{,}000$. Similarly, for a 3-grams, the number of parameters required is $50{,}000^3 = 125 \times 10^{12}$.

3 Exercise Assignment: I5

Regardless of n, the generated sentences training with the corpus `processed_berkeley_restaurant.txt` contain less punctuation marks (e.g., `,`, `.`, `!`, `;`, etc.) than the sentences generated by training with the corpus `processed_shakespeare.txt`. This is probably due to the fact that the latter by Shakespeare uses more punctuation marks than the former, and we are training and generating the sentences based on that respective corpus. For instance, the following is a sentence from `processed_shakespeare.txt` generated with $n = 1$:

```
<s> the.  ,, gracious of </s>
```

whereas the following is a sentence from `processed_berkeley_restaurant.txt` with $n = 3$:

```
<s> it should not cost more than six dollars each meal </s>
```

On the other hand, regardless of the corpus, the generated sentences with $n = 3$ are longer than those with $n = 1$, and the ones with $n = 3$ seem to be more coherent than those with $n = 1$. The previous example also exemplifies this behavior. The better *coherence* is naturally due to the fact that the 3-gram model has more context to generate the next word, namely, the 2 previous words, which makes the generated sentences in general more prone to "sound" coherent. Note the quotation marks around the word *sound*, as most of the generated sentences are *readable* and *human-like*, but they do not necessarily make sense semantically.

Acknowledgements

Chapter 3 from the book by Jurafsky and Martin [1], provided as this week's reading assignment.

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2024. Online manuscript released August 20, 2024; https://web.stanford.edu/~jurafsky/slp3/.