1. N gram language model

(a) implementation

(b) We applied a grid search between 0 to 0.9 for both yielding the following results:

(0.4, 0.5, 0.1) 51.056755

(0.3, 0.6, 0.1) 51.076132

(0.3, 0.5, 0.2) 51.271615

(0.4, 0.4, 0.2) 51.411267

(0.5, 0.4, 0.1) 51.752504

(0.2, 0.7, 0.1) 51.850421

(0.2, 0.6, 0.2) 51.972257

(0.5, 0.3, 0.2) 52.407144

(0.3, 0.4, 0.3) 52.887356

(0.4, 0.3, 0.3) 53.305675

(0.6, 0.3, 0.1) 53.319851

(0.2, 0.5, 0.3) 53.474881

(0.1, 0.8, 0.1) 53.719830

(0.1, 0.7, 0.2) 53.844768

(0.6, 0.2, 0.2) 54.580978

(0.5, 0.2, 0.3) 54.885269

(0.1, 0.6, 0.3) 55.376331

(0.3, 0.3, 0.4) 55.554704

(0.2, 0.4, 0.4) 55.934771

(0.7, 0.2, 0.1) 56.248181

(0.4, 0.2, 0.4) 56.499031

(0.1, 0.5, 0.4) 57.858234

(0.6, 0.1, 0.3) 58.488045

(0.7, 0.1, 0.2) 59.059968

(0.0, 0.9, 0.1) 59.293246

(0.5, 0.1, 0.4) 59.390019

(0.2, 0.3, 0.5) 59.421234

(0.3, 0.2, 0.5) 59.467058

(0.0, 0.8, 0.2) 59.627504

(0.1, 0.4, 0.5) 61.318553

(0.0, 0.7, 0.3) 61.563957

(0.4, 0.1, 0.5) 61.614176

(0.8, 0.1, 0.1) 62.137834

(0.2, 0.2, 0.6) 64.312655

(0.0, 0.6, 0.4) 64.620749

(0.3, 0.1, 0.6) 65.419878

(0.1, 0.3, 0.6) 66.030838

(0.0, 0.5, 0.5) 68.859828

(0.6, 0.0, 0.4) 71.140803

(0.2, 0.1, 0.7) 71.625713

(0.5, 0.0, 0.5) 71.732161

(0.1, 0.2, 0.7) 72.622319

(0.7, 0.0, 0.3) 72.652028

(0.4, 0.0, 0.6) 74.226364

(0.0, 0.4, 0.6) 74.629917

(0.8, 0.0, 0.2) 77.439629

(0.3, 0.0, 0.7) 79.029286

(0.1, 0.1, 0.8) 82.634682

(0.0, 0.3, 0.7) 82.687762

(0.2, 0.0, 0.8) 87.508771

(0.9, 0.0, 0.1) 90.453036

(0.0, 0.2, 0.8) 94.729557

(0.1, 0.0, 0.9) 104.210141

(0.0, 0.1, 0.9) 115.664846

(0.0, 0.0, 1.0) 189.624432

2. Neural language model

(a)

Given the following definitions:

We will derive the gradient with regard of the inputs of the SoftMax function with the cross-entropy loss.

Knowing that is a one hot vector, and w.l.o.g assume that and we have that

We will substitute with the definition of the SoftMax and have

Because we are deriving in respect of the result will be a vector.

Notice that is a vector with 1 only for the k’th element and 0 otherwise (which is the definition of y)

Using that and applying the chain rule for the second term we have:

As for the term, we have an element-wise derivation for which is exactly in the j’th index in the vector .

We will denote the vector as and have