NLP Lab 2

Moritz Bohm, 140164022.

Results Table:

	Target Words:										
Language Model	Whether	Through	Piece	Court	Allowed	Check	Hear	Cereal	Chews	Sell	Accuracy:
Unigram											6/10
Bigram								NCA	NCA		7/10
Bigram add-1											9/10

Where a green cell signifies a correct choice, a red cell signifies a wrong choice and a purple signifies that both choices had 0 probabilities and as such there was no correct answer.

Discussion:

Unigram Language Model

As expected, the unigram language model performs lowest of the models in terms of accuracy. This is due to the fact that it does not consider any context for each candidate word and simply selects the one with the highest probability of appearing in the corpus. This means that the results of using a unigram language model are directly affected by the corpus used to train the data, suggesting that, if a corpus was used which was specifically tailored to selecting the correct word for each of the questions, the unigram language model could be more successful.

Bigram Language Model

The bigram language model performs, as expected, better than the unigram language model - due to the fact that it does consider the context of the target word. However, it is also prone to producing 0 probabilities, as some of the context bigrams are not present in the corpus. This is evidenced by the results of NCA for two of the words, where none of the bigrams are present in the corpus.

Bigram add-1 Language Model

Once again, the results for the bigram add-1 model meet their expectations, as this is the most successful of the three languages models, achieving an almost perfect accuracy. This is due to the fact that it eliminates the zero probabilities which are produced by the bigram model without smoothing, not only eliminating the two NCA results produced by the bigram model, but also resulting in the correct word.