0.d

The advantage of character level model stems from the fact that two words can be very similar, for example "walk" and "walked". With a word level model we don't leverage this similarity and ignore this information which may lead to poorer performance. Character level model can infer that the character sequence "walk" and the character sequence "walked" are similar. If we treat words with capital letters as different words, e.g. "Walk" and "walk" then it's even more confusing to word level model. The disadvantage of character level model is that it increases the length of the sequence. If we have a document of 100 words, and the average length of a word is for example 4, then with a word level model the sequence length is 100 and with a character level model the length is 400. Plus, with character level model there's an added difficulty - the model needs to learn what sequences of characters constitute as legitimate words. This is not a problem with word level model that get this information "for free".

0.e

By using basic power laws, we can rewrite the given expression as follows:

$$
\begin{aligned}
&2^{-\frac{1}{M}\sum log_2 p(s_i|s_1,...,s_{i-1})} = \\
&= (2^{\sum log_2 p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\
&= (\prod 2^{log_2 p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\
&= (\prod p(s_i|s_1,...,s_{i-1}))^{-\frac{1}{M}} = \\
&= (\prod e^{ln p(s_i|s_1,...,s_{i-1})})^{-\frac{1}{M}} = \\
&= e^{-\frac{1}{M}\sum ln p(s_i|s_1,...,s_{i-1})}
\end{aligned}
$$

References