**FNLP text answers**

# **Q2.3**

model tags = [('``', ' . '), ('My', 'DET'), ('taste', 'NOUN') , ('is', 'VERB'), ('gaudy', 'ADV')\*,

(' . ', ' . ')]

correct tags = = [('``', ' . '), ('My', 'DET'), ('taste', 'NOUN') , ('is', 'VERB'), ('gaudy', 'ADJ')\*,

(' . ', ' . ')]

Here we see that the HMM assumes the transition probability is only dependent on the previous word, and therefore “gaudy” is seen as independent of “taste” when in fact that is not the case. This could be because the model sees the verb ‘is’ followed by an adverb in the training data more frequently. Hence, the transition probability from ‘VERB’ to ‘ADV’ would be higher, causing the HMM incorrectly choose the label ‘ADV’.

Word count: 73 words

# **Q3.2**

1) In the labelled data, there are no occurrences of the pronoun ‘he’, but it’s probability won’t be zero due to Lidstone smoothing. However, other pronouns appear in the training set , and "he" appears in the same contexts as these pronouns within the unlabelled data. Therefore, by including the additional unlabelled data, the model observes a higher expected count from verb to ‘he’, assigning a higher transition probability from ‘VERB’ to ‘PRON’ instead of ‘NUM’.

2) In T0, ‘them’ is correctly labelled as ‘PRON’, but in T3 it is incorrectly labelled as ‘NOUN’. This could be because the general tendency in the data is ‘ADP’ followed by ‘NOUN’ instead of ‘PRON’, and due to pseudo-labelling of the unlabelled data, each time the data is tagged, more examples are tagged as ‘NOUN’. This results in data pollution, affecting hard EM’s ability to correctly select the most possible observation.

Word count: 148 words

# **Q4.1**

We would expect translation from the low-resource language to English to be better. When translating from English to a low-resource language, the scarcity of training data can impact the model’s accuracy and ability to understand the target language's morphology and syntax. English's relatively simpler morphology, with less inflection and a simpler verb conjugation system, implies that translating into English often requires fewer adjustments for grammatical agreement, thus making translations from low-resource languages to English better. Additionally, machine translation models have more extensive training data in high-resource languages like English, enabling them to understand and generate English more accurately.

Word count: 98 words

# **Q4.2**

Using synonym words in a Naïve Bayes classifier can lead to issues with semantic understanding and model training. This is because of the Naïve Bayes assumption, which states that features are conditionally independent given the class. As a result, synonyms add more complexity by artificially inflating the feature space with highly correlated features, hindering the classifier’s ability to accurately associate these words with the correct classification or making its predictions overconfident.

Given a set of synonym words S, a modification to enhance handling of synonyms in Naive Bayes could involve synonym consolidation: by mapping the set of synonyms S to a single representative term before training, the model can reduce the feature space complexity and improve semantic understanding. However, this approach could be problematic, as it may oversimplify language nuances and lose the context-specific meanings of words.

Logistic regression addresses the issue of synonyms through feature weighting: it can assign different weights to synonyms based on their context and usage in the training data. This allows for a more nuanced understanding of language, by inherently adjusting to the importance of words in relation to the output variable, which can mitigate some of the synonym-related issues present in Naive Bayes.

Word count: 199 words

# **Q4.3**

One method to detect text generated by a language model (LM) with the potential to generate human-like text could be the use of statistical analysis. This technique exploits patterns in the text generated by language models, by considering specific word frequencies, sentence lengths, and syntactic structures. These patterns can drastically differ from human writing: for instance, language models often struggle with word repetition and/ or word rarity when generating text on certain topics. An example of word repetition could be if the model was asked to explain ‘global warming’: in this scenario, language models tend to repeatedly use the phrase "global warming" when describing different causes and effects of this phenomenon, usually beyond what we would consider as ‘typical human writing’. For the same example of ‘global warming’, the model could alternatively use the following construct:

‘The Earth experiences warming globally in a manner that is unprecedentedly rapid.’

While the sentence is grammatical, it has a higher complexity and is more convoluted, and hence would be less likely to be considered human writing.

Another method that could be used for detecting text generated by language models is stylometric analysis. This method focuses on a specific author’s style of writing, including their use of vocabulary, punctuation, and grammatical patterns. A key characteristic of language models is that they lack a consistent, personalised style of writing, making their text much easier to differentiate from human generated text. The stylometric analysis technique can be particularly powerful when comparing human generated text with suspected language model generated text, as certain deviations in style may reveal insights about the suspected text’s origin.

Lastly, the technique of semantic coherence evaluation could also be used for detecting LM generated text. This method explores the coherence and logical progression over long passages of text. A common drawback of language models is that they struggle to preserve the coherence of certain naratives over long stretches of text by introducing irrelevant or false information. An example of this could be when discussing the impact of technology on education. A human would start by focusing on the importance of technology in modren education, ellaborate on existing tools that can help enhance learning, and conclude by assessing future trends and potential challenges. A language model, however, might initially start in a similar way, but then go on a tangent about the impact of technology on the environment without a logical connection.

Word count: 399 words