1.1 Q: What is it? How much did using a better smoothing method improve it?
Our perplexity of new model is 18.8688.
(42.4421 - 18.8688) * 100% / 42.4421 = 55.54% (roughly)
Our new method improve perplexity by approximately 55.54% from the old method.

1.2 Q: What is it? Did it improve? Why do you think this is?
Our perplexity of trigram model with Kneser-Ney smoothing is 19.3117. Our trigram model does not have an improved perplexity because we have a small testing set and sentences in the testing set use phrases from the training set and thus bigram model with proper smoothing has an incredible performance and trigram cannot achieve as good performance as bigram model. It has high demand on how many words matched.

2. Simple inference in an HMM [Written answers only]

a. ***Given this graphical model, factor the joint probability p(w1,w2,w3,t1,t2,t3,t4) into seven terms by using the conditional independence relationships specified by the model. Write this answer.***

   P(t1) * P(w1|t1) * P(t2|t1) * P(w2|t2) * P(t3|t2) * P(w3|t3) * P(t4|t3)

b. ***What is the probability of w2=man given that t2=N, i.e., p(w2=man|t2=N)?***

   P(t1 = N|start) * P(t2 = N|t1 = N) * P(man|N) + P(t1 = A|start) * P(t2 = N|t1 = A) * P(man|N)

   = 0.8 * 0.3 * 0.4 + 0.2 * 0.8 * 0.4

   = 0.16

c. ***What is the probability of w2=man given that t1=N, i.e.,p(w2=man|t1=N)?***

   P(t1 = N|start) * P(t2 = N|t1 = N) * P(man|N) + P(t1 = N|start) * P(t2 = V|t1 = N) * P(man|V)
   = 0.8 * 0.3 * 0.4 + 0.8 * 0.5 * 0.4
   = 0.256

3. N1: P(N|start)P(young|N)=0.8 * 0.2 = 0.16
A1: P(A|start)P(young|A)=0.2 * 1 = 0.2
V1: P(V|start)P(young|V)=0
N2: N1 * P(N|N) * P(man|N) + A1 * P(N|A) * P(man|N) + V1 * P(N|V) * P(man|N) = 0.0192 + 0.064 + 0 = 0.0832
A2: N1 * P(A|N) * P(man|A) + A1 * P(A|A) * P(man|A) + V1 * P(A|V) * P(man|A) = 0 + 0 + 0 = 0
V2: N1 * P(V|N) * P(man|V) + A1 * P(V|A) * P(man|V) + V1 * P(V|V) * P(man|V) = 0.04 + 0 + 0 = 0.04

N3: N2 * P(N|N) * P(wall|N) + A2 * P(N|A) * P(wall|N) + V2 * P(N|V) * P(wall|N) = 0.009984 + 0 + 0.0096 = 0.019584

A3: N2 * P(A|N) * P(wall|A) + A2 * P(A|A) * P(wall|A) + V2 * P(A|V) * P(wall|A) = 0 +0 + 0 = 0

V3: N2 * P(V|N) * P(wall|V) + A2 * P(V|A) * P(wall|V) + V2 * P(V|V) * P(wall|V) = 0.0208 + 0 + 0 = 0.0208

END: 0.019584 * 0.2 + 0.0208 * 0.2 = 0.0080768

4. N1: P(N|start)P(young|N)=0.8 * 0.2 = 0.16 {start}

A1: P(A|start)P(young|A)=0.2 * 1 = 0.2 {start}

V1: P(V|start)P(young|V)=0 {start}

N2: N1 * P(N|N) * P(man|N) OR A1 * P(N|A) * P(man|N) OR V1 * P(N|V) * P(man|N) = max(0.0192, 0.064, 0) = 0.064 {A1}

A2: N1 * P(A|N) * P(man|A) OR A1 * P(A|A) * P(man|A) OR V1 * P(A|V) * P(man|A) = max(0, 0, 0) = 0 {N1 or A1 or V1}

V2: N1 * P(V|N) * P(man|V) OR A1 * P(V|A) * P(man|V) OR V1 * P(V|V) * P(man|V) = max(0.04, 0, 0) = 0.04 {N1}

N3: N2 * P(N|N) * P(wall|N) OR A2 * P(N|A) * P(wall|N) OR V2 * P(N|V) * P(wall|N) = max(0.00768, 0, 0.0096) = 0.0096 {V2}

A3: N2 * P(A|N) * P(wall|A) OR A2 * P(A|A) * P(wall|A) OR V2 * P(A|V) * P(wall|A) = max(0, 0, 0) = 0 {N2 or A2 or V2}

V3: N2 * P(V|N) * P(wall|V) OR A2 * P(V|A) * P(wall|V) OR V2 * P(V|V) * P(wall|V) = max(0.016, 0, 0) = 0.016 {N2}

END: max (0.0096 * 0.2, 0, 0.016 * 0.2) = 0.0032 {V3}

Tag sequence: start -> A1 -> N2 -> V3 -> end
P(tags|words) = P(words|tags) / P(words) = 0.396 (roughly)
Another tag sequences “start-> N1-> V2-> N3-> end” also have substantial probability.

6.Which one is assigned higher probability?

The first one is assigned higher probability than the second one. From linguistics point of view, “preposition” occur a lot of more frequent than “infinitive marker” and thus “preposition” gets assigned higher emission probability than “infinitive marker” does, and thus I am not surprised the first one turn out higher probability than the second one.