Part 1

1. The sentences are getting better but they are getting more “canned”. Meaning, they seem to be coming directly from the corpus.

2. Basically, we are using longer and longer sequences and there are a limited amount in the corpus. Thus it feels like many of the responses are “canned”. The probability of any word following a long sequence goes to 0 except for an exact match as n goes to infinity.

3. We could use a larger corpus or (as question 4 suggests) we could teach it how to handle new words.

4. First we create a fixed vocabulary. We can construct this vocabulary by taking all the tokens in our corpus and randomly remove a certain percentage of the words and replace them with an $< UNK >$ symbol. We the train on this new processed corpus to we can estimate the probability of an unknown word.

   Selected 4-gram and 5-gram sentences:

   ```
   ngram_generator(4,3,treebank.words())

   Sentence 1:  This is n't Buick 's first travel-related promotion .

   Sentence 2:  The House approved an amendment offered by Rep.  Peter DeFazio -LRB-
   D. , Iowa -RRB- said during House floor debate yesterday .

   Sentence 3:  Mr.  Stern was chairman and chief executive officer , said customers
   were n't willing to commit to an expensive NBI hardware systems because of the
   purchase 's possible effects on the U.K. market for distribution of replacement
   tires .

   ngram_generator(5,3,treebank.words())

   Sentence 1:  Harry Millis , an analyst at UBS Phillips & Drew in London , said
   , ‘‘ An implication that we failed to return investor funds is inappropriate and
   inaccurate .

   Sentence 2:  And though the size of the underlying equity market , exchange officials
   said .

   Sentence 3:  But for small American companies , it also provides a growing source
   of capital and even marketing help .
   ```

Part 2

Arguments for the First Employee

At a high level, the first employee's main argument is that we should model the external "observables" of language rather than attempting to model the internal cognitive process that generates language.

This seems very reasonable since language is, by definition, an "external" phenomenon - it exists only to communicate "meaning" to another person. At a philosophical level, it follows that a valid sentence is not something that follows particular grammatical or stylistics rules; rather, it is anything that effectively encodes some sort of "meaning". Proper grammar is just an attribute of the sentence to increase the likelihood that it is understood. Thus, it would appear that modeling the corpus has strong philosophical footing.

Moreover, at a practical level, it is not clear that it is possible to model any cognitive process in such a way that it would would be useful.

Language is constantly changing. Not only are new words being added to the vernacular but entirely new ways of communication are being devised. Technological advancements throughout history (the invention of the printing press, the Pony Express, the telephone, instant messaging, SMS) as well as historical and cultural phenomenons (gangs and slang talk) have changed language. Let assume that we could model the cognitive machinery inside the language centers of the human mind; there is still no way that we could predict the future and foresee, for example, the invention of the Internet and represent the impact it will have on language.

In fact, this view is consistent with Abney's statement: $grammatical(s) \leftrightarrow lim_{n \to \infty} P_n(s) > 0$. What this means is that the "gramaticallness" of a sentence can be modeled by an $n - gram$ provided that the corpus is large enough. Thus, it would make sense that we should spend time and care to effectively model the corpus. Additionally, this gives us hope for the future: as we collect more data our model will continue to improve.

Lastly, from a product perspective, this make the most sense moving forward. Sitting down and brainstorming a new model for representing language may not go anywhere. Practically speaking we cannot figure out if we will even discover the "real" process that models language. However, the fact of the matter is that $n - grams$ are understood (at a theoretically level) and convenient (to use from an implementation perspective). The storage and computational resources are available to implement an $n - gram$ system and thus, at least, progress can be made.

The fact is that language is in the "eye of the beholder" and thus the best way to model language is to model what is "beheld" - the corpus.

Second Employee

At a high level, the second employee's argument is that we should transcend simply looking at data and strive to find the true model of language that resides in human brain.

Just like the first employee, this position seems defendable from a philosophical perspective since language is, after all, generated by humans. If the human linguistic system can be modeled we can easily model language itself. Thus, this is a lower-level (and perhaps more principled) approach to take.

The first employee makes a practical argument that we may never actually understand and discover this language model. However, the first employee's strategy also is impractical. For it work in the long run we need to store perhaps an infinite amount of data and spend an infinite amount of time computing probabilities. This, in the long run, in impractical and it surely makes more sense to have a more principled and less brute-force approach: we have a hammer (data clusters and distributed systems capable of crunching large numbers and aggregating statistics) and thus everything starts to look like a nail.

Furthermore, the efficacy of the the first employee's approach is highly dependent on the corpus that is chosen. While the first employee can argue that "we may never find the right model for language" the second employee may argue "we may never find the right corpus to train on".

Additionally, the first employee is overly pessimistic about not finding a better model for language where there are many rich avenues of exploration. Simple things like parts-of-speech tagging could lead to a better model that, as Chomsky describes, cleanly partitions sentences into G and $\bar{G}$. We can still collect data and build a corpus and use statistical methods, but not ones as simplistic as $n - grams$.

The fact is that there is a more intelligent model that just counting probabilities and the proof is the human mind. Obviously the human mind does not compute and store conditional probabilities when constructing sentences; rather, there is something more principled taking place. As Chomsky says: "we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds."