NLP Assignment 3 Write Up

Carolyn Ryan

**CFG**

The CFG that I ended up generating and using to parse the sentences in the corpus was the following:

S -> NP VP | VP | RB NP VP | NP RB VP | RB NP | VP  
NP -> PRP | DT | DT JJS | DT NN | NP S | NN | CD NNS | NNS  
VP -> VBG VP | VBD VP | VB VP | VBZ VP | TO VP | MD VP | RB VP  
VP -> VBZ NP | VB NP | RB VBD NP  
VP -> VBN PP | VBD PP | VBG PP | VBD PP  
VP -> VBG | VBN | VB IN S | VBZ JJ  
PP -> IN NP | IN RB | IN PP | PP PP  
  
VB -> 'know' | 'go' | 'take' | 'tell' | 'hate' | 'feel' | 'need' | 'am'  
VBD -> 'was' | 'started' | 'spilled' | 'did' | 'fell' | 'spent'  
VBG -> 'Running' | 'crying' | 'playing'  
VBN -> 'born' | 'gone'  
VBZ -> 'is'  
RB -> 'Sometimes' | 'just' | 'slowly' | 'Now' | 'all' | 'not' | 'Then'  
PRP -> 'I' | 'me' | 'it' | 'You' | 'you'  
DT -> 'this' | 'the' | 'a' | 'Any'  
IN -> 'with' | 'like' | 'of' | 'in' | 'if' | 'out'  
NN -> 'leak' | 'goodness' | 'love' | 'shower' | 'Everything'  
NNS -> 'hours' | 'fonts'  
JJS -> 'worst'  
JJ -> 'terrible'  
MD -> 'can'  
TO -> 'to'  
CD -> '7'

This CFG constructed from the following lexicon:

['I', 'hate', 'this', 'running', 'is', 'terrible', 'everything', 'the', 'worst', 'sometimes', 'feel', 'like', 'was', 'born', 'with', 'a', 'leak', 'any', 'goodness', 'started', 'just', 'slowly', 'spilled', 'out', 'of', 'me', 'now', 'it', 'all', 'gone', 'you', 'did', 'not', 'know', 'then', 'fell', 'in', 'love', 'need', 'to', 'go', 'take', 'shower', 'can', 'tell', 'if', 'am', 'crying', 'spent', '7', 'hours', 'playing', 'fonts']

I made the CFG by using the nltk POS tagger to get the parts of speech for individual words and then constructing rules around cases to generalize the sentence structure. I used Stanford’s parsing tool to help with some of the rule decision-making and intermediate steps/transitions at the NP/VP rule level. This CFG is infinite, which is not necessarily ideal but recursion is a feature of a natural language and a sentence can be infinite through nesting, so the infinite CFG is needed to simulate a natural language. This CFG would not have to be infinite, as there could be a direct rule going S 🡪 specific sentence structure, with no recursion. That however would take away some of the fluidity that I was looking for my language subset defined with my CFG to have. The infinite recursion of the CFG means that the typical nltk generator will crash as it will reach the maximum recursion depth and throw an error. This is not ideal, but was a trade off to allow the CFG to be more generalizable to English as a natural language.

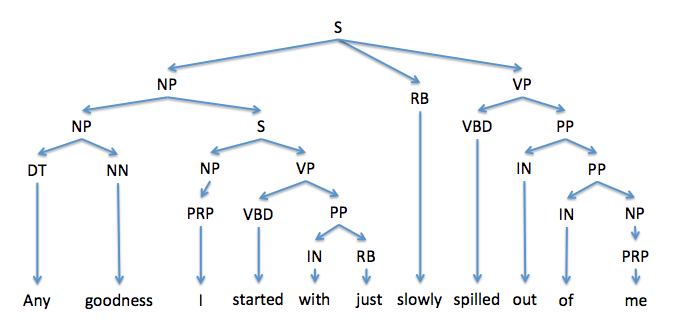


Figure 1. An example parse from my CFG.

This CFG may be used to generate similarly structured sentences (once more terminal words are added to the lexicon), but will struggle with more complex grammar parsing. With some success, you can construct new sentences with the words already in the lexicon, such as ‘I am crying like I am a shower’. But it doesn’t recognize “I love you” due to love being only classified as a noun.

In order to generalize successfully, one would need to add any word they want to parse to the grammar, otherwise it will just throw an error and the program will crash. Adding in words corresponding to their parts of speech may work to help expand the current grammar structures to further sentences.

**Translation and BLEU Scores**

I implemented a direct word-to-word translation for each sentence from English to Spanish. By this, I mean that I looked only at the word level and didn’t include context or utilize my CFG in the translations. This brute force approach led my BLUE scores to be lower than they would’ve been had my translation been able to account for context (both the English context/structure and then the grammar structures of the Spanish language). The BLEU scores of my system were rather low, with all but one sentence getting 0 precision for trigram and quadgram matches, and about half getting 0 precision for bigram matches. All sentences did have at least one matching unigram with the Google translated form of that sentence. We would expect for lower ngram models, the score would be higher, which is represented in my scores. This is because it has less to do with context, structure, and conjugation and thus is more likely to be able to match without having a grammar issue that Google would catch.

Table . Precision and BLEU Scores for Sentences.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sentence | Unigram | Bigram | Trigram | Quadgram | Total BLEU SCORE |
| 1 | 0.66666 | 0.5 | 0 | X | 0.58333 |
| 2 | 0.66666 | 0.5 | 0 | X | 0.58333 |
| 3 | 0.75 | 0.33333 | 0 | 0 | 0.54167 |
| 4 | 0.363636 | 0.1 | 0 | 0 | 0.23182 |
| 5 | 0.454545 | 0.2 | 0 | 0 | 0.32727 |
| 6 | 0.6 | 0 | 0 | 0 | 0.6 |
| 7 | 0.2 | 0 | 0 | 0 | 0.2 |
| 8 | 0.14286 | 0 | 0 | 0 | 0.14286 |
| 9 | 0.25 | 0 | 0 | 0 | 0.25 |
| 10 | 0.28571 | 0 | 0 | 0 | 0.28571 |
| 11 | 0.5 | 0.14286 | 0 | 0 | 0.32143 |
| 12 | 0.625 | 0.42857 | 0.33333 | 0.2 | 0.39673 |

Total for System: 0.37201

Table . My translation and Google's translations of the different sentences.

|  |  |  |
| --- | --- | --- |
| No. | My Machine Translation | Google Translate |
| 1 | Yo odio esto | Odio esto. |
| 2 | Corriendo es terrible | Correr es terrible. |
| 3 | Todo es el peor | Todo es lo peor. |
| 4 | A veces yo sentir como yo era nacido con un fuga | A veces siento que nací con una fuga. |
| 5 | Cualquier bondad yo comenzo con solo lentamente derramado fuera de yo | Cualquier bondad que empecé con sólo se derramó fuera de mí. |
| 6 | Ahora ello es todo ido | Ahora todo se ha ido. |
| 7 | Usted hizo no saber yo | No me conoces. |
| 8 | Entonces usted cayo en amor con yo | Entonces te enamoraste de mi. |
| 9 | Ahora usted saber yo | Ahora me conoces. |
| 10 | Yo necesidad a ir tomar un ducha | Necesito ir a bańarme. |
| 11 | Yo puede no decir si yo soy llorando | No puedo decir si estoy llorando. |
| 12 | Yo solo gastado 7 horas jugando con fuentes | Acabo de pasar 7 horas jugando con las fuentes. |

To improve these scores/translations, we would want to have a CFG of the Spanish language and of the English language and be able to better map between them. The direct translations can reach a basic meaning (sometimes), but struggle when trying to phrase that meaning in a grammatical way. Using framing, conjugation, and context clues could help improve the translations. For example, in Spanish it is common to drop the ‘yo’, so Google’s translations do not have ‘yo’. Or running doesn’t in the context of sentence 2 should translate (Spanish back to English) as “to run is terrible,” just from Spanish’s grammar structures. Breaking verbs into their base form and having who did the action and when they did it could help improve the translations from the conjugation aspect. Taking into account things like when you say ‘the worst’ the translation is ‘lo peor’ and not ‘el peor’ would improve the score. Having more knowledge and using rules from both languages would greatly help the translations similarities to Google’s.