Context-Free Grammar Engineering

Tatiana Matejovicova
University of St Andrews
20 April 2018

Introduction

The goal of this work is to construct a context-free grammar to parse a given set of sentences. However, a frequent result of the CFG formalism is that the grammar overgenerates and ungrammatical sentences are allowed. This is mitigated by adding grammar unification. We account for verb frames i.e. what structures follow a verb and the correct use of person, number and the verb form.

Sentences

The desired grammar has to parse the following set of sentences.

Bart laughs
Homer laughed
Bart and Lisa drink milk
Bart wears blue shoes
Lisa serves Bart a healthy green salad
Homer serves Lisa
Bart always drinks milk
Lisa thinks Homer thinks Bart drinks milk
Homer never drinks milk in the kitchen before midnight
when Homer drinks milk Bart laughs
when does Lisa drink the milk on the table
when do Lisa and Bart wear shoes

Choice of part of speech

The following parts of speech are selected:
ProperNoun -> Bart | Homer | Lisa
Verb -> laughs | laughed | drink | wears | serves | drinks | thinks | wear | laugh | puts
Noun -> milk | shoes | salad | kitchen | midnight | table
Adjective -> blue | healthy | green
Conjunction -> and | or
Wh-Conjunction -> when | while
Preposition -> in | on | before
Determiner -> a | the
Adverb -> always | never

Wh-Adverb -> when | why
Aux -> do | does

As suggested in the specification, the word when has two possible parts of speech. It is and adverb in the sentences 'when do Lisa and Bart wear shoes' and 'when does Lisa drink the milk on the table'. However it is and conjunction in the sentence 'when Homer drink the milk on the table'. To reflect that these can occur at the beginning of a sentence, a special parts of speech Wh-Conjunction and Wh-Adverb are used. The words while and why are added as a Wh-Conjunction and Wh-Adverb respectively to better illustrate these parts of speech.

Sentence forms

The production rules to produce the sentences themselves are the following:
S -> SENT
SENT -> NP VP
S -> Wh-Conjunction SENT SENT
S -> Wh-Adverb Aux SENT
The last two sentences correspond to 'when Homer drinks milk Bart laughs' and 'when do Lisa and Bart wear shoes' respectively. On the right hand side of these productions, the symbol SENT is rather than NP VP, so that the grammar is more compact.

Recursion in CFG

The full CFG was designed and is enclosed in the script. It is designed such that it accepts larger set of correct sentences than that given. Recursion can be observed in several production rules on various levels that allow for this. For example

1. NP -> NP Conjunction NP means that any number of NPs can be joined by conjunctions (e.g. and, or). Therefore the grammar not only accepts 'Bart and Lisa drink milk' but also 'Bart and Lisa and Homer drink milk' etc.
2. VP -> VP PP means that any number of PPs can be attached after a VP. This means that the grammar accepts 'Homer never drinks milk in the kitchen before midnight' but also 'Homer never drinks milk in the kitchen before midnight on the table' etc.
3. Similarly for the rule NP -> NP PP.
4. Nominal -> Adjective Nominal, Nominal -> Nominal Noun mean that NP can be composed of any number of adjectives followed by any number of nouns e.g. 'healthy green tuna salad'.
5. The most interesting is probably the rule VP -> VERB SENT. This rules allows for infinite sentences of the type 'Lisa thinks Homer thinks Bart thinks Lisa thinks Homer drinks milk.'

PP attachment ambiguity

In the CFG we have two rules with PP on the right hand side VP -> VP PP and NP -> NP PP corresponding to VP and NP attachment respectively. However this causes PP attachment ambiguity in parsing. For example the sentence 'Lisa drinks the milk on the table' has two possible parses:

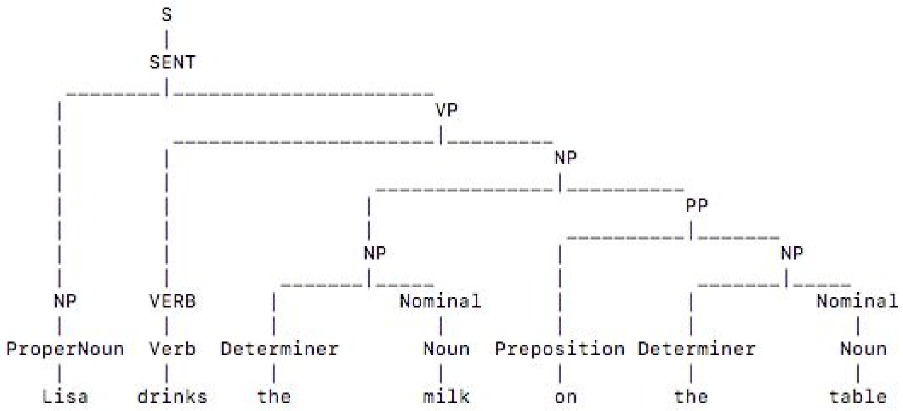

Figure 1 - NP attachment of PP

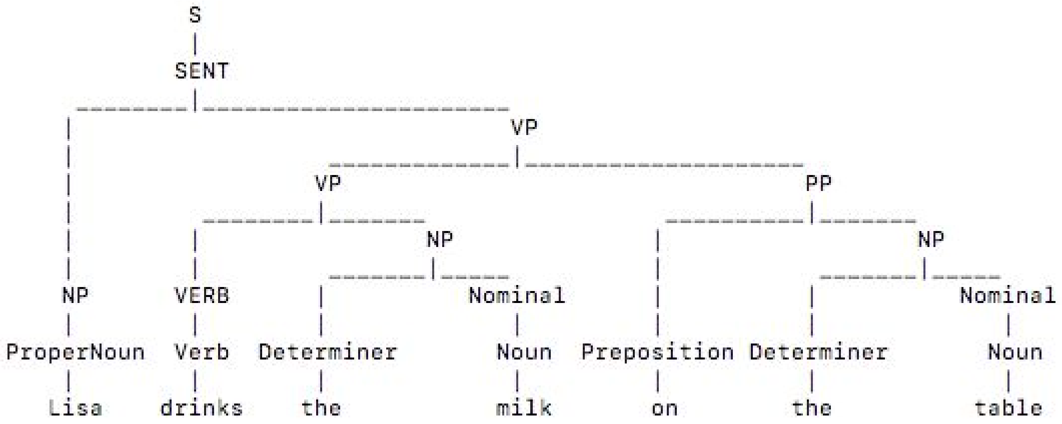

Figure 2 - VP attachment of PP

The meaning of the first parse (Figure 1) is that the milk itself is on the table whereas the second meaning is that Lisa is on the table while drinking the milk. Ambiguities like these are often when dealing with CFGs and to resolve this probabilistic distribution of the rules is added resulting in P-CFGs.

Subcategory

Subcategory of VPs is implemented as outlined in the specification. Verbs in the corpus can be followed by zero one or two NPs or alternatively by SENT. We have the following options:

laugh: (nothing)
drink, wear, serve: NP
serve: NP NP
think: SENT

Note that this setting enables both sentences 'Bart serves Lisa and Bart serves Lisa a healthy green salad', since serve can be followed by either NP or NP NP. With our unification grammar, as before, any VP can be followed by any number of PPs.

Implementing subcategory means that incorrect sentences such as 'Bart laughs the kitchen' or 'Lisa serves' are not accepted.

Number and verb form

Finally number is specified for NPs and form for VPs. Consequently, only the correct combinations are allowed as outlined by the following rules:

S -> SENT
SENT -> NP[NUM=sg] VP[SUBCAT=nil, FORM=vbz]
SENT -> NP[NUM=pl] VP[SUBCAT=nil, FORM=base]
SENT -> NP VP[SUBCAT=nil, FORM=pret]
S -> Wh-Conjunction SENT SENT
S -> Wh-Adverb Aux[FORM=vbz] NP[NUM=sg] VP[SUBCAT=nil, FORM=base]
S -> Wh-Adverb Aux[FORM=base] NP[NUM=pl] VP[SUBCAT=nil, FORM=base]

The NUM argument is passed up from words to NPs. The FORM argument is passed up from words to VPs.

Implementing number and verb form coordination means that incorrect sentences such as 'Bart laugh' or 'when do Homer drinks milk' do not get accepted.

Conclusion

In this project we successfully designed and tested a context-free grammar for a given corpus of sentences. The overgeneration of incorrect sentences was accounted for by extending the CFG to a unification grammar. However with CFG there is still the issue of ambiguity i.e. when one sentences has multiple possible parses for example as we illustrated with the PP attachment. For these cases, probabilistic distribution of the rules would be needed. Moreover, despite the sentences being grammatically correct, their semantics might still be nonsense. For example a sentence 'when does table drink salad before midnight' is nonsensical but syntactically correct. To conclude whether the meaning is plausible, semantic analysis is required.