CS5012: Practical 2 – Grammar Engineering

Student ID: 220029176

Date of Submission: 14/04/2023

Word Count: 825

1. Introduction

The objective of this assignment was to design a grammar to parse a limited subset of English consisting of eight valid sentences and three invalid sentences. Initially, context-free grammar was developed to define rules for parsing sentences. Then, the grammar was refined to become unification grammar which incorporates more rigorous rules to prevent parsing of invalid sentences.

1.1 Compilation and Execution Instructions

1. Download the zip file “CS5012-Assignment2-220029176” and unzip it to a chosen folder location.
2. Open a terminal prompt from within the src folder location.
3. Enter the following command into the terminal “/usr/local/python/bin/python3 parse.py”

2. Interesting Grammar Choices

During the development of this project, three interesting choices were chosen when engineering grammar to create a successful model.

2.1 VP[NUM=?n, TENSE=?t, SUBCAT=?args] -> V[NUM=?n, TENSE=?t, SUBCAT=?args] S

Verb phrases can be very complex expressions, some can embed entire sentences within them called sentential complements. Hence, this verb phrase rule was constructed to fulfil this multifaceted requirement. An example of this structure includes the sentence “Wallace thinks Gromit barks and eats cheese” which can be split into “Wallace thinks” and “Gromit barks and eats cheese”.

2.2 NP[NUM=pl] -> NP CONJ NP *and* VP[SUBCAT=nil] -> VP CONJ VP

These rules were interesting to develop as they are based on the meta rule “X -> X and X” established by Gazdar et al. (1985) which states any non-terminal can be conjoined with the same non-terminal to yield a constituent of the same type.

2.3 QP[SUBCAT=?s, NUM=?n] -> WH | WH[SUBCAT=?s] AUX[NUM=?n]

The most complex of the sentence-level structures within the provided sentences are the various wh- structures containing whom, what, where, why, when, and how. To capture this complexity, a new phase was created called the question phase. This phase allowed all the wh-structures to be accepted.

3. Critical Reflection

The initial development of this project was aided by the module recommended Jurafsky & Martin book (2009) and it was therefore quite straightforward to implement and create a list of parts of speech for each known word, for the context-free grammar model. This was also supported as an English native speaker with an understanding of syntax and grammar.

However, as the complexity of the project increased, there were a variety of challenges that needed to be overcome. Firstly, when incorporating number agreements and subcategorisation into the unification grammar model, there were various significant challenges, particularly in the case of verb tags. The complex nature of some verbs necessitated the creation of several different categories of verb tags, utilizing forms such as the base form, the third person singular form, and the preterit.

Although an attempt was made to add the past participle/gerund verb form to enable the parsing of words like "inventing", time constraints prevented its inclusion. Incorporating this verb form would be necessary in future work, as it would enable the implementation of many new sentence structures.

When moving forward onto the unification grammar model, understanding the unification process became vital to the work. A basic understanding could be derived from the lecture material. However, an advanced understanding was required to overcome the challenges faced within the phases. Specifically, the verb phase contained many examples of complex patterns that needed to be unified using head features, subcategorization, perspective, and pluralisation. Whilst implementing these attributes was challenging, the final grammar model demonstrates an advanced yet concise solution to the problem.

The final grammar model was fully tested against each of the negative and positive sentences, including the extension sentences. Additionally, further sentences were added to test the model using the pre-established lexicon. During this testing phase, some negative sentences seemed to be parsed when testing the wh-structures such as “how does Gromit ate cheese”. The issue is rooted in the tense of the structure and requires further investigation to determine the cause of the issue.

Overall, a final model has been successfully implemented that is able to accurately parse grammatically correct sentences of simple or complex nature, including questions and statements. This model has been rigorously tested for accuracy and flexibility within the English Language. Furthermore, additional complex sentences have been added to the sentence pools to further evaluate the model.

4 Remarks on this Practical

In general, this practical was created to allow the student to gain an in-depth understanding and appreciation for formal grammar within the English Language through a practical means. However, this assessment goes further and allows the students to observe patterns and techniques that are key within natural language processing that can be taken forward into the development of a more complex model, that may be able to embody meaning or semantics from a given language.

5 Extension

The extension for this given assessment was to implement new types of sentences. Therefore, this grammar model was extended to allow sentences from all wh-words; whom, what, where, when, why, and how. Various sentences were added into the positive and negative sentence pools to allow the model to be tested fully.

6 References

Gazdar, G., Klien, E., Pullum, G., & Sag, I. (1985). *Generalised Phrase Structure Grammar.* Blackwell.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing.* New Jersey: Pearson.