UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA APLICADA

Disciplina Processamento de Linguagem Natural
Docente Carlos Augusto Prolo
Discente Felipe Cortez de Sá

Extração de gramática

1 Introdução

Este relatório descreve a implementação de um programa para extrair a gramática de um corpus e convertê-la para a *Chomsky Normal Form*.

2 Programa

O programa foi desenvolvido utilizando a linguagem de programação Python 3 apenas com as bibliotecas padrão, sendo elas `argparse` para tratar argumentos de linha de comando e `collections` para contar a frequência de cada regra.

2.1 Uso

Para rodar o programa, digite na linha de comandos

```
$ python3 grammar.py input_file [-m "pre" | "post"]
```

em que `input_file` é o arquivo do corpus contendo a gramática que se deseja extrair.

A flag `-m` é utilizada para imprimir as regras pré ou pós conversão para CNF. Caso seja omitida, o padrão é a impressão das regras convertidas.

As regras são escritas na saída padrão `stdout`.

3 Conversão para forma normal de Chomsky

Foi utilizado o corpus `traindata-TOP`, garantindo que há uma regra base que serve como símbolo `START`.

3.1 Terminais

O método `cnf_term` cria, para cada terminal, uma regra não-terminal no formato `TERM_NT -> TERM`. As ocorrências de terminais nas produções são então substituídas pelos não-terminais criados.

3.2 Binarização

O método `cnf_bin` transforma regras com mais de dois não-terminais no lado direito em múltiplas regras, cada uma com dois não-terminais no lado direito.

Para a regra `NP -> DT_NT NN_NT CC_NT NN_NT NN_NT`, por exemplo, temos

```
NP   -> DT_NT NP_1
NP_1 -> NN_NT NP_2
NP_2 -> CC_NT NP_3
NP_3 -> NN_NT NN_NT
```

4 Resultados

As 1000 regras mais comuns aparecem 690502 vezes dentre um total de 783701 produções, cobrindo 88.10% do corpus, e podem ser vistas no arquivo `grammar.txt`. As regras na forma normal de Chomsky encontram-se no arquivo `grammar-cnf.txt`