O Reconhecimento e *Parsing Top-down* de Gramáticas Livres de Contexto

Rayssa Küllian Martins

Instituto de Matemática e Estatística, Universidade de São Paulo

19 de outubro de 2014

**Resumo**

Neste trabalho foi desenvolvida uma implementação do algoritmo de *Earley* para o reconhecimento e *parsing* de Gramáticas Livres de Contexto para sentenças em Português dado um córpus.

Foram realizadas análises e comparações de resultados em diferentes casos para encontrar possíveis necessidades de melhoria em termos de *performance*, manipulação dos dados e precisão.

1. Introdução

Asfafas

1. Fundamentação Teórica

~ "Aqui apresentaremos rapidamente alguns conceitos básicos necessários para a compreensão deste trabalho..."

1. *Top-down* Parser: Earley

Asfasfas

1. Manipulação dos Dados

Asfasfas

1. Arquitetura e Estratégias de Implementação

Asfasfas

1. O problema
2. Metodologia

O córpus *“aires-treino.parsed”,* utilizado nas duas distribuições, está em Português (Brasil) e é constituído de regras de Gramáticas Livres de Contexto, ou seja, cada sentença é sintaticamente anotada com sua respectiva árvore de *parse*, formando um córpus *treebank* com seu POS *tagset* específico (JURAFSKY; MARTIN, 2006). A imagem 3.1 mostra a árvore para a sentença “alguma vez se havia de ver a vaidade sem lugar .”.

(IP

(NP (Q Alguma)

(N vez))

(NP (SE se))

(HV havia)

(PP (P de)

(IP (VB ver)

(NP (D a)

(N vaidade))

(IP

(PP (P sem)

(NP (N lugar))))))

(. .))

Figura 3.1: Exemplo de *parser tree* no córpus “*aires-treino.parsed”*

Para análise e avaliação de desempenho do algoritmo de *Earley* foram realizados cinco experimentos em que se extraiu aleatoriamente as partes de treinamento. A distribuição dada consiste em:

* Gramática extraída de 80% do córpus de treinamento
* Gramática verificada em 20% do córpus de teste

As regras gramaticais foram extraídas do córpus de treinamento e a sentença foi extraída do córpus de teste.

As medidas utilizadas para avaliação foram *cobertura* e *precisão*, tal que cobertura é definida pela proporção de sentenças do córpus de treinamento que foram reconhecidas pela gramática e precisão definida como pela percentagem de sentenças que admitem a árvore desta sentença no córpus.

Caso a sentença não fosse reconhecida, a precisão foi utilizada para descrever a quantidade de *brackets* encontrados pelo *parser* que estão presentes na árvore desta sentença no córpus (e.g., árvore do córpus definida por “IP-> NP NP VB”, o algoritmo *reconheceu* a sentença, mas com a árvore de parse “IP-> NP VB ADV“).

sentença extraída - reconhecida - precisão - tempo

1. Resultados

Asfasfas

1. Experimento 1 (gráficos + breves explc.)

Asfasfasfas

1. Experimento 2

Asfasfasfsa

1. Trabalhos futuros (OPT)

Asfasfas

1. Conclusão

~ BEM breve!

"Apesar do problema SAT ser um problema NP-Completo, é possível resolver eficientemente várias instâncias e a razão M/N é um bom indicador se uma certa distância aleatória pode ser resolvida eficientemente."

Referências

JURAFSKY, Daniel; MARTIN, James H.. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2. ed., Pearson Prentice Hall, 2009.

Citação com autor incluído no texto: Jurafsky e Martin (2006)

Citação com autor não incluído no texto: (JURAFSKY; MARTIN, 2006)

GOETZ, Brian et al. **Java Concurrency In Practice**. Addison Wesley, 2006.

Citação com autor incluído no texto: Goetz et al. (2006)

Citação com autor não incluído no texto: (GOETZ et al., 2006)