MAC5701
Tópicos em Ciência da Computação
Plano de Estudos: **Métodos Probabilísticos de Análise Sintática para o Português**

Fábio Natanael Kepler
Orientador: Marcelo Finger

25/04/2005

1 Introdução

A Lingüística Computacional estuda a linguagem humana, buscando métodos computacionais para entender e descrever seu comportamento. Uma das categorias de conhecimento de linguagem necessária para o empenho nesta tarefa é a sintaxe. A sintaxe consiste no estudo das relações estruturais entre as palavras. Mais especificamente, entre as palavras de uma sentença. Desta forma, a análise sintática procura identificar as estruturas entre as palavras de uma sentença e se elas estão corretamente relacionadas, podendo, além disso, disponibilizar esta relação entre as estruturas através de uma árvore sintática.

Para se identificar estas estruturas, é preciso algum conhecimento da linguagem analisada. Geralmente este conhecimento é representado sob a forma de gramáticas, que apresentam regras de utilização das estruturas de uma sentença. Entretanto, nenhuma gramática é capaz de representar todas as formas de utilização da linguagem. Assim, apenas gramáticas parciais e aproximadas podem ser obtidas. Dentre estas gramáticas, o conjunto de gramáticas chamado de Gramáticas Livres de Contexto (GLC) é um dos mais utilizados. Não vamos descrevê-las aqui, mas para uma explicação detalhada veja [2, 7].

1.1 Análise Sintática por Tabulação

Dada uma GLC e uma sentença, queremos ser capazes de achar as estruturas para a sentença de acordo com a gramática. Um algoritmo que faz isso é chamado de *analisador sintático livre de contexto*. Em particular, um destes algoritmos bastante conhecido é o de *análise sintática por tabulação*.

Um analisador sintático deste tipo possui três estruturas de dados principais: uma *lista de chaves*, uma *tabulação*, e um conjunto de *arestas*. Uma tabulação é um conjunto de *entradas* onde cada uma consiste no nome de um símbolo terminal ou não-terminal da gramática (por exemplo, um sintagma nominal ou uma preposição), no ponto de início da entrada, e no seu comprimento. A lista de chaves é uma pilha de entradas que estão esperando serem inseridas na tabulação. E por fim as arestas indicam as regras que podem ser aplicadas às entradas da tabulação para construir entradas maiores sobre elas. Uma descrição detalhada deste algoritmo é dada em [2].

1.2 Algoritmos Probabilísticos

Algoritmos como o acima são chamados de determinísticos, pois consistem na aplicação em séria de regras gramaticais viáveis sobre a sentença de entrada, transformando-a na árvore sintática final. Entretanto, uma sentença pode ser sintaticamente ambígua. Neste caso, um algoritmo destes gera mais de uma árvore sintática para a sentença. Para tentar encontrar a árvore que melhor explica a correta estrutura sintática da sentença, adicionamos probabilidades às regras da gramática. Com isto obtemos uma Gramática Livre de Contexto Probabilística (GLCP), e um algoritmo que a utiliza pode retornar a árvore mais provável para uma sentença. Um algoritmo como este é chamado de probabilístico.

2 Métodos de Análise Sintática

Além da utilização de Gramáticas Livres de Contexto e de Gramáticas Livres de Contexto Probabilísticas, outros métodos para análise sintática existem e têm sido propostos tanto para o Inglês quanto para o Português.

2.1 Método Proposto por Charniak

O método proposto por Charniak [3] limita a geração de árvores a apenas utilizar as regras gramaticais vistas nos dados de treinamento do algoritmo. Além disso, Charniak também enfoca o aprendizado automático de gramáticas livres de contexto a partir de dados de treinamento não anotados manualmente, apresentando técnicas de agrupamento por propriedades das palavras e técnicas para decidir o sentido de uma palavra de múltiplos sentidos, através do contexto ou de informações externas. Este método foi utilizado para o Inglês.

2.2 Método Orientado ao Núcleo-Léxico

O método probabilístico orientado ao núcleo-léxico foi desenvolvido por Michael Collins em sua tese de doutorado em Ciência da Computação na Universidade da Pensilvânia em 1999 [4]. Este método tenta realizar a análise sintática a partir das informações léxicas mais relevantes presentes na sentença. Isto é feito através de uma parametrização do problema, que considera dois pontos críticos:

1. Como as árvores sintáticas podem ser quebradas em fragmentos menores?

2. Como essa escolha pode ser feita para se obter um modelo estatístico?

A argumentação de Collins é que a influência em uma árvore da propriedade da localidade de um item lexical núcleo deve motivar a modelagem de escolhas no problema da análise sintática. Em outras palavras, a propriedade da localidade que as palavras de uma sentença possuem define o domínio no espaço sobre o qual a palavra núcleo influencia as demais palavras da sentença. Cada palavra da sentença afeta um domínio limitado na sentença. O propósito então é escolher uma decomposição que permita uma parametrização que reflita melhor a influência local dos itens lexicais núcleos. Assim, o item lexical núcleo deve ser gerado antes da estrutura que depende dele. Com esta decomposição, aplica-se hipóteses de independência sobre relacionamentos que estão fora do domínio de influência (localidade).

2.2.1 Adaptações para o Português

O trabalho realizado por Collins naturalmente foi aplicado ao Inglês, mas Fabiano Sousa, em sua dissertação de mestrado em Ciência da Computação na Universidade de São Paulo [5], e Andréia Bonfante, em sua tese de doutorado em Ciência da Computação na Universidade de São Paulo [1], apresentaram analisadores sintáticos para o Português baseados no trabalho de Collins.

Os resultados obtidos pelos dois, entretanto, foram inferiores aos obtidos para o Inglês. Mas os dois trabalhos chegam a conclusões semelhantes quando apresentam algumas razões para que isso tenha acontecido, como principalmente o fato de terem contado com poucos dados de treinamento (às vezes 10 vezes menores que para o Inglês) e com uma pequena gama de conhecimento coletado sobre o problema da análise sintática do Português.

Os dois trabalhos explicam a adaptação do método de Collins para o Português, e contribuem para o conhecimento da área quando relacionam as dificuldades e os problemas encontrados no desenvolvimento de analisadores sintáticos para o Português.

2.3 Método Baseado em Máxima Entropia

Ratnaparkhi propõe um modelo de analisador sintático baseado em uma técnica de aprendizado de máquina chamada de Máxima Entropia [8]. Neste modelo, não há uma gramática que dita quais ações são possíveis. Todas as ações que levam a árvores bem formadas são permitidas, e o modelo é usado para pontuar essas ações. A pontuação de cada ação na derivação é usada para computar a pontuação da derivação inteira. Para analisar uma sentença, um analisador sintático baseado neste modelo usa uma rotina de busca que explora eficientemente o espaço de todas as árvores possíveis e tenta encontrar a de maior probabilidade.

3 Plano de Estudos

A idéia para a escrita da monografia é estudar os algoritmos probabilísticos de análise sintática na aplicação para o Português. Embora vários métodos usados no Inglês tenham sido bastante descritos por Sousa [5] e Bonfante [1], nosso objetivo final é mais específico: queremos levantar métodos probabilísticos de análise sintática tendo em vista a construção de um analisador sintático para o Português. Não iremos propriamente construir um analisador, mas desejamos, sim, obter informação e conhecimento suficientes para mostrar como a construção de um analisador sintático específico para o Português poderia ser feita. Isto poderá ser através de um novo método, ou então através de um método híbrido baseado em outros existentes. Como um exemplo, ainda não encontramos na literatura trabalhos sobre a utilização de Cadeias de Markov de Alcance Variável [6] na análise sintática. Com o estudo dos diversos métodos existentes, esperamos descobrir se a aplicação a este problema de cadeias de Markov deste tipo seria viável ou não.

Referências

[1] Andréia Gentil Bonfante. *Parsing Probabilístico para o Português do Brasil*. Tese de doutorado, Programa de Pós-Graduação em Ciência da Computação, Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, Junho 2003.

[2] Eugene Charniak. *Statistical Language Learning*. The MIT Press, 1993.

[3] Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. In *AAAI/IAAI*, pages 598–603, 1997.

[4] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. Tese de doutorado, Department of Computer and Information Science, University of Pennsylvania, 1999.

[5] Fabiano de Carvalho e Sousa. Analisador sintático estatístico orientado ao núcleo-léxico para a língua portuguesa. Dissertação de mestrado, Programa de Pós-Graduação em Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, Outubro 2003.

[6] Fábio Natanael Kepler. Um etiquetador morfo-sintático baseado em cadeias de markov de tamanho variável. Dissertação de mestrado, Programa de Pós-Graduação em Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, Abril 2005.

[7] Christopher D. Manning and Hinrich Schütze. *Foundations Of Statistical Natural Language Processing*. The MIT Press, 1999.

[8] Adwait Ratnaparkhi. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34:151 — 175, February 1999.