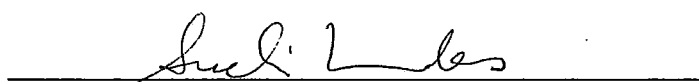
UMA ABORDAGEM BASEADA EM REDES BAYESIANAS PARA A SOLUÇÃO DA AMBIGÜIDADE LÉXICA

Leila Maria Ripoll Eizirik

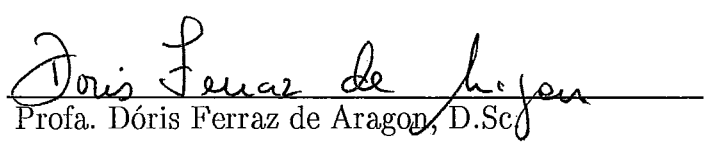
TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS–GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

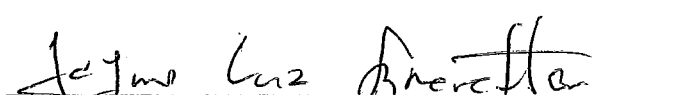


Profa. Sueli Bandeira Teixeira Mendes, Ph.D.
(Presidente)

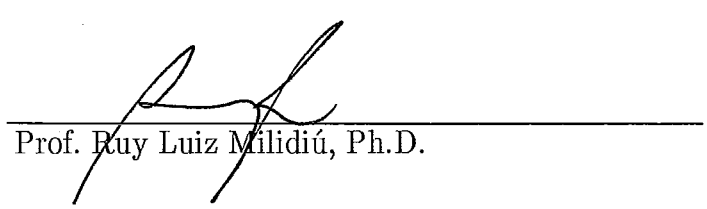
Prof. Valmir Carneiro Barbosa, Ph.D.



Profa. Dóris Ferraz de Aragon, D.Sc.



Prof. Jayme Luiz Szwarcfiter, Ph.D.



Prof. Ruy Luiz Milidiú, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
maio de 1990

EIZIRIK, LEILA MARIA RIPOLL

Uma Abordagem Baseada em Redes de Bayes para a Solução da Ambigüidade Léxica [Rio de Janeiro] 1990

xii , 105 p. 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia de Sistemas e Computação, 1990)

Tese – Universidade Federal do Rio de Janeiro, COPPE.

1. Ambigüidade Léxica, Linguagem Natural, Redes Bayesianas.

I. COPPE/UFRJ II. Título (Série)

À memória de meu pai.

À minha mãe pelo apoio constante.

Ao Nelson que soube me amar e acompanhar meu crescimento profissional.

Às minhas filhas Julia, Alice e Cecilia.

AGRADECIMENTOS

À minha orientadora Sueli B. T. Mendes pelo incentivo constante ao longo do meu doutoramento.

Ao meu orientador Valmir C. Barbosa pelo apoio, paciência, e dedicação durante todas as fases desta tese.

Aos colegas Inês de C. Dutra e Luís Adauto Pessoa pela eficiente implementação dos simuladores.

Ao Projeto de Processamento Paralelo pelo apoio e incentivo.

A todos os meus colegas do Programa de Sistemas, em particular, às minhas colegas da linha de Inteligência Artificial, por me liberarem das atividades acadêmicas.

Ao Nelson pela paciência na correção do português.

Às minhas filhas pelo carinho e compreensão.

Ao Aloysio pelo apoio nos momentos difíceis.

À Daisy pelo cuidadoso trabalho de datilografia.

RESUMO DA TESE APRESENTADA À COPPE/UFRJ COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS (D.Sc.)

UMA ABORDAGEM BASEADA EM REDES BAYESIANAS PARA A SOLUÇÃO DA AMBIGÜIDADE LÉXICA

Leila Maria Ripoll Eizirik
maio de 1990

ORIENTADORES: Valmir Carneiro Barbosa
Sueli Bandeira Teixeira Mendes
PROGRAMA: Engenharia de Sistemas e Computação

Neste trabalho apresentamos uma proposta de solução da ambigüidade léxica das linguagens naturais utilizando um modelo Bayesiano. A rede proposta está dividida em duas sub–redes, a sintática e a semântica. Desenvolvemos algoritmos para a construção automática da sub–rede sintática a partir de uma gramática livre de contexto e apresentamos a organização da sub–rede semântica baseada em uma gramática de casos.

As sub–redes sintática e semântica interagem de forma que na solução da ambigüidade o diagnóstico sintático é alterado pelo semântico e vice–versa. A análise das sentenças é feita de forma inteiramente paralela e a interpretação semântica emerge da interação entre a análise sintática e a análise semântica.

Implementamos um simulador para redes Bayesianas e fizemos testes com sentenças contendo ambigüidades de categoria gramatical, de estrutura sintática e de significados das palavras. Concluímos que as redes Bayesianas são adequadas para expressar e compatibilizar as relações semânticas com as informações sintáticas viabilizando um método de análise das sentenças completamente paralelas.

ABSTRACT OF THESIS PRESENTED TO COPPE/UFRJ AS PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF SCIENCES (D.Sc.)

A BAYESIAN–NETWORK APPROACH TO LEXICAL DISAMBIGUATION

Leila Maria Ripoll Eizirik
May, 1990

THESIS SUPERVISORS: Valmir Carneiro Barbosa
Sueli Bandeira Teixeira Mendes

DEPARTMENT: Systems Engineering and Computer Science

In this thesis we present a Bayesian–network approach to the lexical disambiguation of natural language sentences. The network is subdivided into a semantic subnetwork and a syntactic subnetwork. The syntactic subnetwork is obtained algorithmically for a given context–free grammar, and the semantic subnetwork is based on a case grammar.

The sentence analysis is carried out completely in parallel and during the process the two subnetworks interact continually. The semantic interpretation of the sentences emerges from the global interaction of the two subnetworks.

We have implemented a Bayesian–network simulator and performed tests with sentences where both syntactic and semantic ambiguities were present. We found that semantic relations and syntactic information are expressed in an efficient and precise way by using a Bayesian model, which in addition is very amenable to a parallel implementation.

ÍNDICE

Pág.

ÍNDICE DE ALGORITMOS

ÍNDICE DE FIGURAS

Pág.

Pág.

CAPÍTULO I

INTRODUÇÃO

A ambigüidade é uma das características centrais das linguagens naturais. Uma sentença é ambígua quando possibilita a construção de mais de uma interpretação semântica.

Quando afirmamos a existência de um ou mais significados de uma sentença e buscamos solucionar a ambigüidade, estamos presumindo que existe um contexto no qual a sentença está inserida e no qual a ambigüidade é eventualmente resolvida.

Entretanto, a compreensão de qualquer sentença pressupõe um interlocutor/leitor. Então, não há como assegurar uma compreensão única e completa do que está sendo dito/escrito. Acreditamos que isto acontece porque cada pessoa tem seus conceitos (significados de palavras/sentenças) suficientemente gerais para permitir a comunicação com os outros e suficientemente particulares para gerar a ambigüidade de interpretações.

Estas observações levam à conclusão sobre a existência de uma continuidade entre a lingüística e outras formas de conhecimento do mundo. A constatação de que a interpretação semântica de uma sentença não está completa em representações do tipo estrutura profunda*, proposta por CHOMSKY [1965] está presente em seu trabalho posterior [CHOMSKY, 1970].

CHOMSKY [1970] afirma ser natural (embora parcialmente correto) supor que a interpretação semântica de uma sentença é determinada pelo conteúdo semântico intrínseco dos itens léxicos e a maneira como eles estão relacionados ao nível da estrutura profunda. A razão pela qual esta afirmação é apenas parcialmente correta reside, segundo Chomsky, no fato de que fatores não–lingüísticos, tais como crenças e intenções, interferem na interpretação semântica.

Além disso, se analisarmos as afirmações de Chomsky, veremos que ele se

---

* Para uma definição do conceito de estrutura profunda veja–se CRYSTAL [1988].

refere a "uma semântica intrínseca dos itens léxicos", ou seja, os itens léxicos possuem significados claros, pré–definidos e independentes. Acreditamos que significados assim definidos de forma estanque constituem uma descrição falha dos significados que são consensuais e que viabilizam a comunicação entre as pessoas.

Esta questão da existência de significados "discretos" das palavras reflete–se nos sistemas computacionais mediante uma representação mais ou menos localizada dos significados. Ou seja, um determinado significado de uma palavra pode ser representado pelo preenchimento de determinadas características básicas (micro–características) ou mediante uma única entidade que representa o significado como um todo.

Segundo WILKS [1988], a existência de um ou mais significados distintos para cada palavra é fundamental para a discussão da ambigüidade: "... the issue at the heart of lexical ambiguity resolution – namely, are there, in any real sense, discrete word senses, or is it all just a matter of fuzzy, numerical–boundary classifications of individual examples of use ?".

Do ponto de vista computacional, essa questão está diretamente ligada à forma de representação dos conceitos. Nesse sentido, as abordagens fundamentadas na lógica são demasiado rígidas para modelar os significados das palavras.

Além disso, os sistemas computacionais em que a análise sintática e semântica estão implementadas como processos autonômos não possuem mecanismos que viabilizem a utilização de informações da análise semântica e contextual para alterar a análise sintática e vice–versa. Esse modo de funcionamento dificulta a solução da ambigüidade, que exige em geral informações sintáticas, semânticas, contextuais e até mesmo de conhecimento do mundo.

Esta tese apresenta uma proposta para o tratamento da ambigüidade léxica, onde os processos de análise sintática e semântica são implementados de forma independente do ponto de vista de estrutura interna. Porém, a comunicação entre os processos sintático e semântico é bilateral e a interpretação da sentença emerge da interação dos processos.

Foi feita uma implementação completa da rede sintática considerando sentenças de comprimento máximo 5. A rede sintática foi testada para diversas entradas, incluindo entradas ambíguas. Além disso, foram feitos alguns testes com

parte da sub–rede semântica. A rede funcionou corretamente no caso de ambigüidades de categoria gramatical e de significado.

Foram feitos, ainda, alguns testes com sentenças cujas ambigüidades de categorias gramaticais resultam em mais de uma estrutura sintática correta. Nesses casos, a solução da ambigüidade deu–se no nível semântico, a sub–rede sintática adaptou–se à solução semântica e optou pelo diagnóstico sintático correto.

Concluindo os testes realizados, apresentamos à rede uma entrada completamente agramatical do ponto de vista sintático, porém com itens semanticamente relacionados. Obtivemos então diagnósticos sintáticos que localizavam corretamente o verbo principal. Além disso, a inexistência de uma estrutura sintática correta não impediu uma interpretação parcial da semântica associando corretamente os itens relacionados.

No Capítulo II apresentamos uma classificação das ambigüidades de acordo com as suas origens: ambigüidade estrutural, caso em que a ambigüidade provém das diversas formas possíveis de agrupar os constituintes de uma sentença, ou ambigüidade léxica, caso em que provém dos significados distintos de uma mesma palavra. Ilustramos com alguns exemplos os tipos de ambigüidade existentes e os seus contextos de solução. Apresentamos também uma breve descrição dos principais sistemas computacionais que tratam do problema e estabelecemos algumas comparações com o nosso trabalho.

No Capítulo III descrevemos o modelo de redes Bayesianas, enunciando os teoremas principais que estabelecem o poder de expressão do modelo e que provêm um método seguro de especificação de redes consistentes com um modelo probabilístico.

Inicialmente apresentamos um critério de separação em DAGs (Directed Acyclic Graphs) que expressa mais acuradamente as dependências do modelo probabilístico. A seguir enunciamos os teoremas que estabelecem a vizinhança mínima que deve ser consultada para atualização do valor de um nó da rede e descrevemos as restrições que devem ser atendidas para a especificação das probabilidades condicionais associadas a um DAG. Finalmente, discutimos os problemas de atualização de valores em redes que contêm ciclos, apresentando o

método de simulação estocástica que foi por nós utilizado.

Os Capítulos IV, V e VI constituem a parte central desta tese, apresentando respectivamente a descrição geral da abordagem adotada, a descrição detalhada da parte sintática, e a descrição detalhada da parte semântica.

O Capítulo IV apresenta uma visão global da rede. Descrevemos em linhas gerais como estão organizados os módulos que compõem a rede e como interagem para obter a solução da ambigüidade. A rede está dividida em duas sub–redes: a sub–rede sintática e a sub–rede semântica. Apresentamos mediante exemplos, o funcionamento geral da rede e a forma como interagem as sub–redes sintática e semântica. Além disso, descrevemos duas formas de implementar testes de concordância gramatical para verificar a correção da entrada e também para utilizar estas informações na solução da ambigüidade.

O Capítulo V trata especificamente da análise sintática. O método de análise sintática está baseado em uma gramática livre de contexto. Esta gramática define a estrutura sintática das sentenças analisáveis pelo sistema.

Apresentamos um algoritmo que tem como entrada uma gramática livre de contexto G e fornece como saída uma rede Bayesiana que reconhece exatamente L(G). O resultado da análise sintática é sumarizado por alguns nós especiais da rede que simbolizam estruturas sintáticas corretas. Se existir ambigüidade sintática todos os possíveis diagnósticos sintáticos resultarão igualmente prováveis.

A rede construída pelo algoritmo contém alguns nós inúteis, já que não integram nenhum diagnóstico sintático. Apresentamos então um algoritmo que elimina todos os nós inúteis.

Além disso, se a gramática de entrada estiver fatorada de uma forma especial a rede obtida é mais eficiente. A eficiência a que nos referimos diz respeito ao grau de paralelismo existente na rede. Apresentamos finalmente um algoritmo para obter a forma fatorada de uma gramática livre de contexto.

No Capítulo VI descrevemos a sub–rede responsável pela análise semântica e detalhamos a forma como se dá a interligação das sub–redes sintática e semântica. Mostramos como foram construídos os diversos níveis da sub–rede semântica e como devem ser especificadas as probabilidades condicionais

associadas aos nós. Apresentamos também um algoritmo para construção da interligação das sub–redes sintática e semântica. Finalmente, descrevemos detalhadamente a análise de algumas sentenças ambíguas.

No Capítulo VII apresentamos os resultados da implementação, extraímos algumas conclusões, mencionamos algumas possíveis extensões do sistema e alguns tópicos que permanecem em aberto.

No Apêndice I apresentamos os algoritmos de simulação utilizados, bem como algumas informações sobre o número de simulações necessárias para a convergência da rede.

No Apêndice II apresentamos uma análise da complexidade dos algoritmos apresentados.

CAPÍTULO II

O PROBLEMA E SOLUÇÕES ANTERIORES

II.1 – INTRODUÇÃO

No Capítulo I, discutimos a importância da solução da ambigüidade na questão da compreensão das linguagens naturais e destacamos as dificuldades de abordar esse problema que está presente em todos os níveis de análise da sentença.

Neste capítulo, apresentaremos uma classificacão das ambigüidades tendo em vista as suas origens e os seus foros de solução. Além disso, descreveremos os principais sistemas computacionais de compreensão das linguagens naturais que abordam a questão da ambigüidade, bem como estabeleceremos algumas comparações com o nosso trabalho.

II.2 – CLASSIFICAÇÃO DAS AMBIGÜIDADES

As diversas interpretações semânticas de uma sentença podem ter origem na existência de vários significados para uma palavra ou nas formas de agrupar os constituintes que compõem a estrutura da sentença.

Quando a ambigüidade refere–se a significados distintos dos itens léxicos, é chamada ambigüidade léxica. A ambigüidade léxica pode ser classificada de acordo com o tipo de distinção entre os diversos significados de um item léxico. Uma palavra tem significados polissêmicos quando estes significados, embora distintos, estão relacionados. Por exemplo, a frase "O menino descobriu a menina" é ambígua pois o verbo "descobrir" tem, segundo HOLANDA [1986], vinte e quatro significados. Entre estes, descobrir (retirar a coberta) e descobrir (encontrar) são significados polissêmicos do verbo "descobrir", que resultam em interpretações semânticas distintas.

Um item léxico tem significados homônimos quando os significados são distintos e não estão relacionados. Por exemplo, na frase "O menino correu para o banco", o substantivo "banco" tem significados homônimos: banco (objeto para sentar) ou banco (instituição financeira). Ambos são plausíveis, resultando na ambigüidade da sentença.

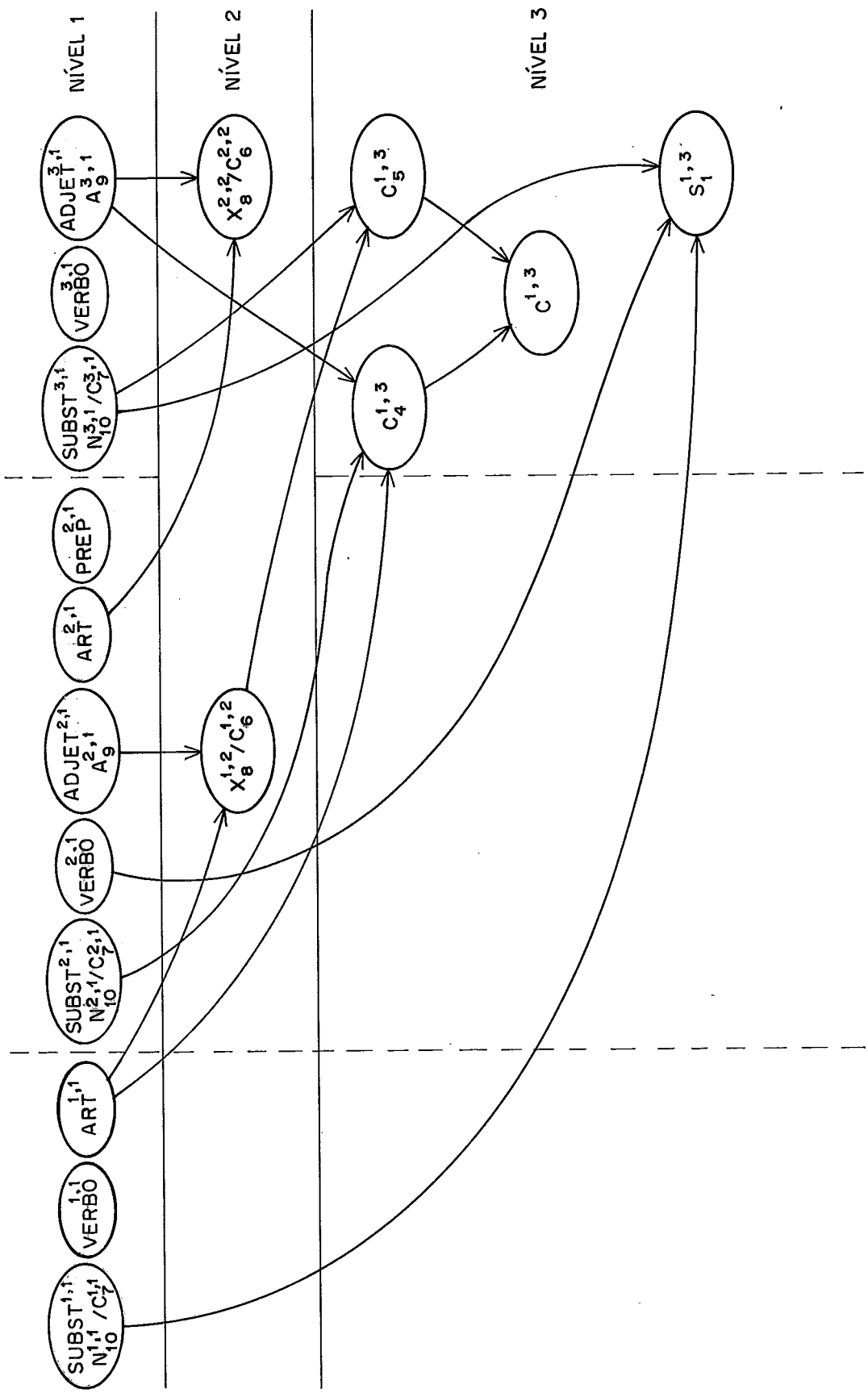


Fig. V.3 – Fragmento da rede sintática com três grupos de entrada incompletos