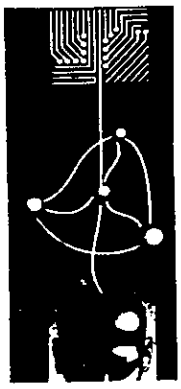

1° Congresso Brasileiro de Redes Neurais

Escola Federal de Engenharia de Itajubá
Itajubá, 24 a 27 de outubro de 1994

AQUISIÇÃO DE CONHECIMENTO DE TEXTOS UTILIZANDO TÉCNICA CONEXIONISTA

I. R. Guilherme
DEMAC/IGCE/UNESP-Rio Claro
A. F. Rocha
DFB/IB/UNICAMP

RESUMO

O sistema JARGÃO é uma ferramenta para aquisição de conhecimento de textos em liguagem natural. O conhecimento obtido é estruturado em redes neurais simbolicas hierárquicas: redes de conceitos, redes de classes e redes de teorias. O conhecimento obtido pode ser utilizado para diversas finalidades.

1. INTRODUÇÃO

A linguagem falada e escrita tem um importante papel no desenvolvimento do aprendizado e na comunicação das pessoas. Com o surgimento dos computadores e fazendo uso do desenvolvimento dos conhecimentos relacionados com a linguagem, tem se desenvolvido a área do processamento da linguagem natural (PLN), que trabalha com problemas como: interfaces em linguagem natural para sistemas computacionais, interpretação e geração de textos, tradução automática, recuperação e filtragem de informação, etc. O desenvolvimento desta área tem sido feito por pesquisadores de duas linhas distintas em suas concepções básicas:

- os cognitivistas: assumem que a unidade básica de processamento é o símbolo e que todas as regras para o processamento são definidas previamente. São fortemente influenciados pelos trabalhos do Chomsky ([CHOM57],[CHOM65]);

- os conexionistas: assumem que a unidade básica de processamento é o neurônio e que estes são agregados (conexões) para a construção de estruturas mais complexas (relações, regras). As relações são obtidas a partir de um conjunto de exemplos ([McCL86][RUME86]).

A aplicação de técnicas conexionista no PLN tem alcançado resultados expressivos em problemas denominados de baixo nível, como reconhecimento de caracteres, reconhecimento e produção da fala, modelagem do efeito do contexto na compreensão da linguagem natural, etc. Para tratar problemas de linguagem de mais alto nível a técnica não tem produzido resultados tão expressivos. Isto ocorre porque o processamento de alto nível é predominantemente simbólico. Porém sabe-se que este processamento simbólico é feito por redes de neurônios no cérebro.

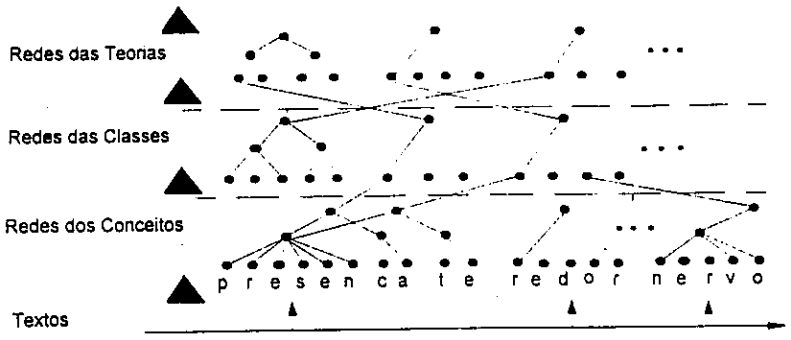

Figura 1: Estrutura Hierárquica das Redes no Jargão

A solução portanto consiste em definir um modelo de neurônio que faça processamento numérico e simbólico. Este trabalho utiliza a definição de neurônio ([ROCH92]) que possibilita o processamento numérico e simbólico na análise de textos.

A análise de entrevista, relatório ou legislação é uma atividade complexa e trata-se de um requisito comum em vários setores da atividade humana. Análise automatizada de textos é também uma tarefa complexa e as implementações existentes consistem em produzir representação detalhada de um número pequeno de textos. Com a evolução da capacidade de armazenamento e processamento dos sistemas computacionais aumentou-se a necessidade de sistemas para a análise de grande número de textos ([JACO93]). Este tipo de análise, devido a sua complexidade computacional não pode ser feita de forma rigorosa. A forma de análise implementada no sistema JARGÃO e que será descrita neste trabalho, consiste em fazê-la da mesma forma que nós humanos trabalhamos, ou seja explorando a imprecisão e a incerteza, e fazendo uso de outra importante heurística que é o contexto.

2. ESTRUTURA DO SISTEMA JARGÃO

O sistema Jargão gera três níveis de redes hierarquicamente organizadas: redes de conceitos, redes das classes e as redes das teorias. Estas redes são criadas contendo as informações comuns de um grupo de textos, contidos na base de dados em linguagem natural analisada. A primeira rede consiste da rede de conceitos que é construida com as palavras ou associações de palavras que representam os conceitos mais significativos e freqüentes presentes nos textos. A redes de classes é construída com as classes mais significativas e freqüentes nos textos. e que consiste da conexão da saída das redes de conceitos nas entradas das redes de classes. Essas redes de classes são usadas como entradas das redes de teorias. que consistem dos textos padrões ou significativos contidos na base de dados (Figura 1).

3. OBTENÇÃO DAS REDES DE CONCEITOS

Um conceito pode ser um conceito simples ou composto. O conceito simples consiste de ter associado apenas uma palavra. Conceitos compostos podem ter necessidade de várias palavras. Neste contexto, a rede de conceitos simples é construída como redes neurais de três camadas. Aos neurônios da camada

inferior, denominado de entradas serão associado códigos ASCII. Os neurônios da camada intermediária são de agregação, e os neurônios das camadas superior são denominados de saída. Os neurônios da camada inferior podem ser agrupados em partes. A parte denominada Germe é composta pelos neurônios associados aos caracteres iniciais que são idênticos nas palavras incorporadas na mesma rede e que são agregados por um neurônio da camada intermediária. O Germe serve como índice dos conceitos. A parte denominada Halo é composta pelos neurônios associados aos caracteres que complementam o germe na formação da palavra e que são agregados por um neurônio da camada intermediária. A rede gerada para conceitos diferentes das existentes não necessitam de camadas intermediárias e os neurônios de entradas são agregados por um neurônio de saída. Nas outras redes, a agregação do germe com os halos é feita por neurônios de saída.

As redes de conceitos compostos são construídas utilizando redes de conceitos simples já criadas. A obtenção dos conceitos compostos pode ser feita utilizando os recursos do sistema ou pelo próprio usuário, que neste caso consiste em agrupar as redes de conceitos simples na formalização do conceito composto.

Todas as palavras contidas nos textos são lidas e os módulos das redes de conceitos são geradas de acordo com a topologia descrita acima.

O usuário faz a análise dos conceitos encontrados e aqueles que não têm relação com o contexto trabalhado ou que sejam muito pouco freqüentes são eliminados. Deve-se também agrupar os conceitos que são sinônimo em um único modulo de rede. Gera-se então o dicionário de conceitos, sobre o qual passa-se a operar.

4. OBTENÇÃO DAS REDES DE CLASSES

4.1 Definição da sintaxe

O primeiro aspecto a ser levando em consideração para a geração das classes é verficar a característica da base analisada no que refere-se o tipo de expressões nela existente. As expressões podem ser: descritivas (contém informações ou a descrição de um símbolo) ou procedural (descreve uma ação, e a ação é representada por verbos).

Portanto, de acordo com a característica das expressões contidas nos textos analisados, deve-se definir inicialmente quais os conceitos que são termos chaves (símbolos ou verbos), e as outros conceitos devem ser classificadas como complemento. Podem existir conceitos classificados como termo chave e complemento.

O passo seguinte consiste em definir as classes, e que pode ser feita utilizando as classes gramaticais encontradas na gramática transformacional, gramática de casos, etc. Pode-se, por exemplo, adotar, utilizando as definições de gramática de casos as seguintes classes: transitividade do verbo (transitivo direto, transitivo indireto, bitransitivo), agente, coagente, lugar, objeto, origem e destino, transporte, etc.

Utiliza-se conceitos conhecidos em neurofisiologia que é a noção de transmissor, receptor e controlador para codificar as informações das classes lingüísticas nos neurônios. A afinidade transmissor/receptor é utilizado para codificar as regras sintáticas que especificam as classes de concatenação nesta sintaxe. Neste contexto deve-se definir:

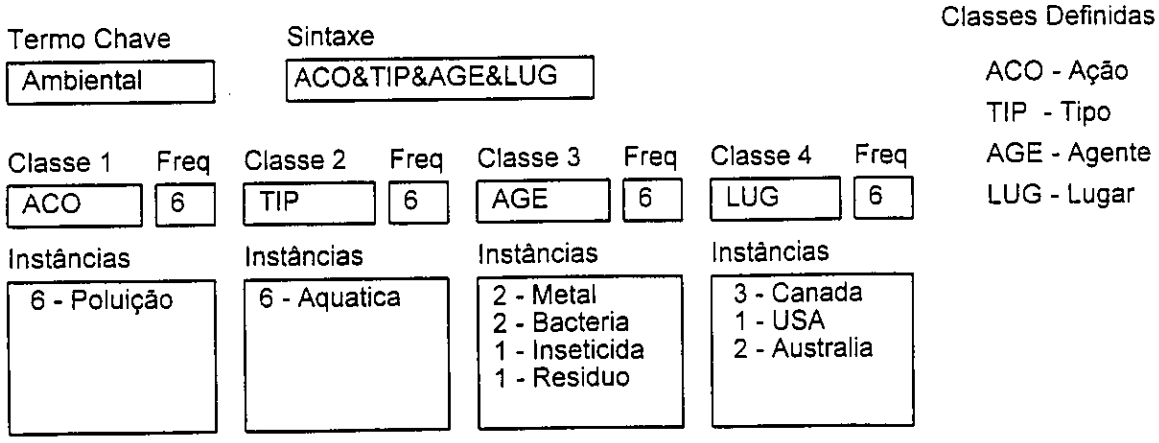

Figura 2: Apresentação das Instâncias de uma rede de classe.

a) os trasmissores que são representado com símbolo como vtd, age, coa, lug, obj, ori, etc, codificando respectivamente agente, coagente, lugar, objeto, origem e destino, etc.

b) os receptores que são representados como , VTD, VTI, AGE, COA, LUG, OBJ, ORI, etc.

c) os controladores do tipo vtdAGE, ageCOA, vtdOBJ com objetivo de implementar regras sintáticas condicionais, como por exemplo:

AGE ^ ageOBJ >> OBJ ou
VTD ^ vtdAGE >> AGE

Utilizando-se as regras acima, define-se os símbolos associados aos transmissores, receptores e moduladores que representam a sintaxe que será adotada. Tendo sido feito a definição ou a escolha da sintaxe a ser utilizada o usuário deverá:

a) atribuir diferentes classes de receptores para os termos do dicionário, classificados como termo chave, usando as cadeias correspondentes (AGE, VTD, VTI, etc).

b) atribuir diferentes classes de transmissores ou moduladores para os termos do dicionário, classificados como complementos, usando as cadeias correspondentes (age, obj, lug, etc ou vtdAGE, ageCOA, etc).

Tendo sido feito as definições para os conceitos do dicionário será então gerado as redes de classes, usando como conjunto de treinamento as frases dos textos.

4.2 Obtenção das classes

As redes de classes são estruturas de mais alto nível que são instanciadas pelos conceitos. A topologia consite em: serem criadas tantas redes de classes quanto forem o número de termos chaves definidos; o primeiro neurônio na rede de classes é o termo chave, esse neurônio caracteriza o módulo; os outros neurônios de entrada são classes complementos e eles produzem receptores para diferentes categorias sintáticas aceita pelos verbos e seus próprios complementos. Nesse caso, são criadas tantos neurônios de entradas quantos forem as classes sintáticas requeridas pelos termos chaves e seus complementos.

A definição sintatica funciona como heurística na busca das combinações. Esta codificação especifica as informações contextuais desejadas pelo usuário. Quanto mais restrita for a sintaxe definida maior será seu efeito sobre a explosão combinatorial.

Utilizando as definições sintáticas podemos criar classes como estruturas de representação do conhecimento conhecidas (frames, dependência conceitual, redes semânticas).

4.3 Definição semântica das classes

Cada uma das redes de classe associadas aos termos chaves é mostrada (Figura 2) para que o usuário elimine as classes incoerentes com o contexto e defina a semântica das classes corretamente construídas. Para auxiliar nesta definição, são mostradas as frases que foram utilizadas no treinamento e as estruturas sintáticas de todas as classes obtidas. Tendo estas informações o usuário deverá:

a) guardar as classes que inequivocadamente definem um significado específico no contexto trabalhado. Poderá associar uma frase ou um símbolo que represente aquela classe, caso contrário, o sistema cria um símbolo para representar a classe.

b) descartar as classes que sejam semanticamente e sintaticamente incoerentes.

Nas situações onde os resultados obtidos não são satisfatórios deve-se refazer a definição sintática afim de corrigir as eventuais distorções.

5. OBTENÇÃO DAS REDES DAS TEORIAS

Este nível é semelhante ao nível das redes de classes. Inicialmente constroi-se um dicionário contendo as classes mais freqüentes. Define-se uma sintáxe que represente as teorias desejadas, e então associa-se as classes. Em função das definições das classes e das suas ocorrências nos textos são gerados as redes de teorias. Em [ROCH92b] descreve-se a geração de teorias utilizado como classe os conceitos linguísticos de tema e rema. As redes de teorias encontradas são mostradas aos usuários.

6. APLICAÇÕES

O sistema JARGÃO tem sido utilizado para aquisição de conhecimento de um conjunto de textos em linguagem natural, numa dada área de especialização. O conhecimento obtido tem sido utilizado no desenvolvimento de sistemas especialistas conexionistas([ROCH92b et all]), para a estruturação dos casos em sistemas de raciocínio baseados em casos, etc.

Com o crescimento do volume de informação disponiveis tem tornado-se necessário a criação de eficientes sistemas de indexação de textos, filtragem e recuperação de informações, obtenção de padrões nos textos, etc. Alguns dos sistemas desenvolvidos ([JACO93]) tem criado mecanismos para o usuário codificar o seu conhecimento que é utilizado nas consultas. O conhecimento (crenças, metas, etc) pode ser constantemente ajustado. O sistema JARGÃO pode operar desta forma, ou ser utilizado para obter o conhecimento a ser utilizado em consultas e ao também para recuperar os textos desejados. A eficiência da recuperação da informação depende do nivel de conhecimento especificado. Caso o usuário especifique o conhecimento a nível de conceito a quantidade de informações recuperadas será enorme e conterá informações indesejadas, se especificar a nível de classes a eficiência deverá aumentar considerávelmente e a nível de teoria obtem-se somente as informações associadas com a teoria que se desejada.

7. BIBLIOGRAFIA

[CHOM57] Noam Chomsky. Sintactic Structures, the Hague: Mouton(1957).

[CHOM65] Noam Chomsky. Aspects of the theory of sintax, MIT Press, Cambridge, USA(1965)

[JACO93] P. S. Jacobs, L. F. Rau. Inovations in text interpretation, Artificial Intelligence, Vol 63, 1993..

[McCL86] J. L. McClelland, D. E. Rumelhart, Parallel Distributed processing: Explorations in Microstructure of Cognition (Vol. 2), Cambridge, MA: Bradford Books, 1986.

[ROCH92a et all] A. F. Rocha, I.. R. Guilherme, M. Theoto, A. M. K. Miyadahira, and M. S. Koizumi, A neural Network for extracting Knowledge from Natural Language Data Bases, IEEE Transactions on Neural Network, Vol. 3, Num. 5, September 1992.

[ROCH92b et all] A. F. Rocha, I. R. Guilherme, R. J. Machado, Knowledge Aquisition: An Connectionist Approach, Proceedings of 3th Annual Simposiun of the International Association of Knowledge Engennering- IAKE, November, 1992, Washington, USA

[ROCH92c] A. F. Rocha, *The theory of Brains and Machines*, in Lectures and Notes in Artificial Intelligence, New York: Springer Verlag, 1992, 400pp.

[RUME86] D. E. Rumelhart, J. L. McClelland, Parallel Distributed processing: Explorations in Microstructure of Cognition (Vol. 1), Cambridge, MA: Bradford Books,1986.