

Capítulo 1

Introdução

“Os limites da minha língua são os limites de meu mundo”

Ludwig Wittgenstein (1889-1951), *Tractatus Logico-Philosophicus* (1922)

1.1 Considerações Iniciais

O interesse da Inteligência Artificial no **Processamento de Linguagem Natural** (PLN) vem desde a década de 1950, quando Alan Turing projetou o *Teste de Turing*, com a intenção de fornecer uma definição operacional satisfatória para a inteligência (Turing, 1950). Ele definiu comportamento inteligente como a habilidade de alcançar performance “humana” em todas as tarefas cognitivas. No teste, o computador deveria ser interrogado por um ser humano através de um terminal e seria considerado inteligente se o interrogador não pudesse dizer se havia um homem ou um computador do outro lado da linha. Entre as capacidades que este computador deveria possuir, estava o PLN (Russell e Norvig, 1995).

A *língua* (chamada equivocadamente de linguagem natural no PLN) é um meio de comunicação. É organizada em um sistema com níveis de regras complexos, dos níveis dos sons e ritmos (*prosódia*), aos níveis do significado e do relacionamento com o mundo (*pragmática*). Cada nível trata de um aspecto do processo de comunicação e forma um subsistema completo com seus próprios elementos e regras de combinação (Sowa, 1984).

Um sistema que tem a pretensão de processar a língua natural deve considerar todos esses níveis. Mas, devido à sua grande complexidade, a maioria considera apenas alguns deles, preferencialmente a sintaxe e algumas vezes, a semântica. Teorias lingüísticas e psicolingüísticas procuram explicar o funcionamento do processador de linguagem humano.

A maioria dos sistemas encontrados na literatura são sistemas eficientes computacionalmente, mas pecam por tratar de forma ingênua as questões lingüísticas. O sistema proposto nesta Tese é baseado na teoria dos **papéis temáticos**, à qual a teoria lingüística da Regência e Ligação deu visibilidade e importância (Chomsky, 1995).

1.2 A Abordagem Simbólica

A abordagem mais usada nos sistemas de PLN é a abordagem baseada em **regras gramaticais (abordagem simbólica)**. O termo gramática, nesse contexto, se refere a um conjunto de **regras de produção** (ou **regras de rescrita**) que descrevem quais sentenças são parte de uma determinada linguagem. As **gramáticas livres de contexto** (tipo 2, segundo a **hierarquia de Chomsky**) são as mais empregadas nos sistemas simbólicos de PLN, por se tratar de gramática de fácil implementação computacional e de dar conta da maior parte das construções sintáticas da língua natural.

Mas além do fato de nem todas as construções serem livres de contexto, uma abordagem simbólica ao PLN não permite que o sistema reconheça como válida uma construção sintática que empregue palavra não presente em seu **léxico** ou que apresente uma categoria não prevista em seu conjunto de regras. O sistema simbólico só é capaz de aumentar o seu “conhecimento”, se mais regras forem incorporadas à sua base.

1.3 A Abordagem Conexionista

Mais recentemente, diversos pesquisadores têm trabalho em sistemas que empregam as **redes neurais artificiais** para realizar o PLN. As grandes vantagens desta abordagem são a sua capacidade de aprendizado e **generalização**. Sua característica distribuída permite que o sistema seja **tolerante a falhas**, isto é, mesmo diante de uma entrada incompleta, o sistema muitas vezes é capaz de analisar sentenças corretamente. Esta análise é mais próxima dos modelos de processamento de informação linguística humanos.

Apesar da **abordagem conexionista** ser tão antiga quanto a Inteligência Artificial, somente na década de 1980 é que houve um grande crescimento desta área, principalmente depois do lançamento dos dois volumes sobre **PDP – Processamento Distribuído Paralelo** (McClelland e Rumelhart, 1986; Rumelhart e McClelland, 1986). No volume 2 do PDP, McClelland e Kawamoto (1986) apre-

sentam um sistema conexionista para fazer a atribuição do caso semântico de Fillmore (1968). Este trabalho se tornou um clássico do uso da abordagem conexionista para o PLN, principalmente pelo uso das chamadas **microcaracterísticas semânticas** para a representação das palavras.

Muitas críticas ao **conexionismo** apareceram a seguir (Fodor e Pylyshyn, 1988). Mas, mesmo assim, vários sistemas utilizando esta abordagem foram construídos (St. John e McClelland, 1989 e 1990; Jain e Waibel, 1990; Jain, 1991; Miikkulainen e Dyer, 1991; Miikkulainen, 1993 e 1996; Rosa, 1993 e 1997; Rosa e Netto, 1994; Chan e Franklin, 1998 e outros).

1.4 A Abordagem Híbrida Simbólico-conexionista

Combinar as vantagens da abordagem simbólica (poder expressivo das representações lógicas gerais e facilidade de representação) com as vantagens do conexionismo (aprendizado, generalização e tolerância a falhas) é muito interessante para qualquer sistema de PLN. Esta é a intenção dos sistemas **híbridos** simbólico-conexionistas. O sistema proposto nesta Tese – o HTRP (*Hybrid Thematic Role Processor*) é um sistema híbrido deste tipo. Neste sistema, introduz-se conhecimento simbólico, baseado num conjunto de regras de produção para os papéis temáticos, como **pesos de conexão** entre as unidades de uma **rede conexionista**. A rede é em seguida treinada e fornece um “conhecimento” simbólico revisado. Além disso, o sistema revela a **grade temática** para sentenças do português semanticamente bem formadas. Duas versões do HTRP foram implementadas: uma com regras iniciais e outra com **pesos iniciais** aleatórios. A opção pela realização das duas versões do HTRP, se deveu ao fato de que desta forma se torna mais interessante a verificação das teorias lingüísticas no sistema. Em outras palavras, a versão *com* conhecimento inicial corresponde a teorias lingüísticas que postulam conhecimentos inatos como base para a competência sintática nos seres humanos. Assim, o sistema tal como foi implementado permite modelar diferentes hipóteses lingüísticas.