1 INTRODUÇÃO

Neste primeiro capítulo apresentam-se as principais motivações e objetivos do trabalho realizado e relata-se a seqüência em que foi desenvolvido. Ao final, indica-se a organização do texto desta tese.

1.1 Apresentação

O processamento de linguagens naturais requer o desenvolvimento de programas que sejam capazes de determinar e interpretar a estrutura das sentenças em muitos níveis de detalhe.

As linguagens naturais exibem um intricado comportamento estrutural visto que são profusos os casos particulares a serem considerados. Uma vez que as linguagens naturais nunca são formalmente projetadas, suas regras sintáticas não são nem simples nem óbvias e tornam, portanto, complexo o seu processamento computacional.

A formalização completa de linguagens naturais exige o emprego de modelos computacionais mais elaborados que expressões regulares e autômatos finitos, ou mesmo que gramáticas livres de contexto e autômatos de pilha.

De fato, um dos problemas mais importantes, dentre os usualmente encontrados no processamento de linguagens naturais, corresponde à dificuldade de se expressar, através de um formalismo legível e expressivo, as complexas nuanças estruturais, sempre presentes nas linguagens naturais.

O maior desafio consiste no estabelecimento de uma forma simples e clara para definir as construções lingüísticas de maneira fácil e inteligível, sem que para isso seja necessário lançar mão de mecanismos meta-lingüísticos de difícil interpretação.

A teoria das linguagens regulares e dos autômatos finitos permite lidar confortavelmente com atividades relacionadas à morfologia, dicionários, categorização, etiquetação morfológica, inflexão de palavras, etc. (KARTTUNEN, 2001).

O formalismo que melhor permite representar os aspectos livres de contexto da sintaxe de uma linguagem natural, tais como a estrutura superficial das sentenças, múltiplas árvores de derivação e ambigüidades sintáticas e não-determinismos, é o autômato de pilha.

Esta classe de problemas é muito freqüente no processamento computacional de linguagens de programação, dispondo-se de um grande repertório de algoritmos e modelos que permitem resolvê-los (AHO; ULLMAN, 1972).

O tratamento adequado de não-determinismos e ambigüidades, o tratamento de categorias vazias, de coordenações e de anáforas são exemplos de problemas mais complexos, que requerem formalismos dependentes de contexto, visto que são fenômenos lingüísticos que denotam interdependências, não apenas entre sentenças como também entre elementos de uma mesma sentença.

Linguagens dependentes de contexto geralmente se definem através de regras de substituição, eventualmente condicionais, as quais restringem as situações de contexto nas quais podem ser permitidas as substituições que definem.

Muitos métodos são empregados em sistemas de processamento de linguagem natural, adotando diferentes paradigmas. Assim, para isso dispõe-se de métodos exatos, aproximados, pré-definidos ou interativos, inteligentes ou algoritmicos, etc. (JURAFSKY; MARTIN 2000).

Os métodos estatísticos exploram probabilidades, taxas de minimização de erros, aprendizagem dinâmica. Freqüentemente não requerem gramáticas formais ou qualquer outra

representação exata da linguagem. Uma vez que empregam técnicas aproximadas proporcionam implementações mais simples, as quais infelizmente, tendem a exibir taxas de erro residuais devido à inevitável presença de casos pouco prováveis.

Métodos não-estatísticos empregam conceitos advindos da Teoria dos Autômatos e das Linguagens Formais a fim de representarem linguagens naturais e outras linguagens sensíveis ao contexto. Apesar de apresentarem resultados mais precisos, o desempenho que com eles pode ser obtido é geralmente baixo, devido às explosões combinatórias que freqüentemente devem enfrentar.

Métodos híbridos permitem a fusão e a interação complementar entre as técnicas estatísticas e não-estatísticas, proporcionando ao usuário os benefícios das soluções mais econômicas de cada técnica (MANNING, 2005). Naturalmente o desempenho pode ser diferenciado, de acordo com a abrangência de cada método e a particular aplicação a que se destina. (CHIANG, 2004).

O emprego de técnicas de inferência na construção de uma especificação de linguagem (gramática ou autômato) é atrativo sempre que não houver disponibilidade de uma descrição formal da linguagem, e sempre que houver um conjunto significativo de amostras ou de padrões corretos da linguagem desejada.

Técnicas de inferência são freqüentemente empregadas de forma relativamente adequada nas tarefas de formalização de uma linguagem a partir de uma amostra da mesma.

Nesse caso, geralmente é necessária uma interação com o ser humano, tanto para a inserção de dados complementares, para a posterior extração automática de informação, como para o fornecimento de casos irregulares.

O uso de corpora, textos alinhados, textos anotados e outros, é uma técnica fundamentada na captura informal (manual ou automática) e validação do conhecimento sobre a linguagem, através da análise de um conjunto representativo de dados experimentais.

Geralmente, uma grande quantidade de amostras rigorosamente revistas de textos pertencentes à linguagem e admitidos como significativos a partir de critérios oriundos da Lingüística é empregada como base para o treinamento de algoritmos que efetuarão o processamento da linguagem natural.

Tal prática pode ser realizada através de algoritmos de inferência e de dispositivos de reconhecimento capazes de se automodificarem. Ao final, o dispositivo, devidamente treinado, é utilizado para analisar textos desconhecidos, norteado pelo conhecimento acumulado durante a fase de treinamento. Este método tem sido largamente aplicado em atividades de processamento de linguagem natural tais como: tradução automática, extração semântica e mineração de dados.

Exemplos de Corpora para a Língua Inglesa são passíveis de serem acessados em (PENN PARSED Corpora of Historical English.), (PENN-HELSINKI Parsed Corpus of Early Modern English). Corpora para a Língua Portuguesa podem ser acessados em (LINGUATECA), um centro distribuído de recursos para o processamento computacional da Língua Portuguesa .

N-gramas (geralmente N = 2 ou 3) são uma outra aplicação de heurística ao processamento de linguagem natural que também pode eliminar o requisito de descrições formais da linguagem.

Com esta técnica, textos são localmente inspecionados através de uma janela de N palavras. As decisões são, em sua maioria, efetuadas a partir do conhecimento previamente adquirido e da informação presente no conjunto de palavras em inspeção.

Trata-se de um método aproximativo simples, com bom desempenho no tempo e que pode ser eficientemente empregado para a tarefa de desambigüização morfológica de palavras. (KINOSHITA, 1998)

O processamento de linguagens naturais com o auxílio de autômatos e gramáticas exige que se disponha de alguma definição formal da linguagem. Em sua forma original, proporciona uma técnica exata para a análise da linguagem, que usa como base a matemática discreta, autômatos e teoria das linguagens formais.

Muitas ferramentas de processamento de linguagem natural existem, muitas das quais implementadas com linguagens funcionais ou lógicas. Em particular, encontram-se implementações de DCG (*definite clause grammars*) e ATN (*augmented transition networks*) desenvolvidas em Prolog (MATTHEWS, 1998), (DOUGHERTY, 1994) e Lisp (GAZDAR; MELLISH, 1989), respectivamente.

A Gramática de Adjunção de Árvore (TAG - "*Tree Adjoining Grammar*") introduzida em (JOSHI, 1985) é um formalismo cuja estrutura de dados fundamental é a árvore. As árvores podem ser do tipo inicial ou do tipo auxiliar. As árvores iniciais são combinadas por duas operações denominadas substituição e adjunção.

Robert Frank explora em (FRANK, 2002) o papel deste formalismo na teoria da sintaxe, motivado principalmente pelo trabalho Programa Minimalista de Chomsky (CHOMSKY, 1993). Trata-se de um formalismo com o auxílio do qual é possível encontrar um conjunto de árvores elementares para cada uma das sentenças em $L_1 = \{a^n b^n c^n \mid n \in N\}$. Por outro lado, nas linguagens $L_2 = \{a^n b^n c^n d^n e^n \mid n \in N\}$ e $L_3 = \{www \mid w \in \Sigma^*\}$ não são geradas por nenhuma TAG

A alternativa para o *processamento* e *representação* de linguagem natural empregada nesta pesquisa é a utilização do formalismo adaptativo (NETO, 1993), que traz como vantagem, a possibilidade de inclusão automática de mecanismos de reconhecimento dependentes de contexto, a possibilidade de exploração das características de aprendizado e o potencial desempenho diferenciado para problemas de alta complexidade.

Nesse trabalho, o autor descreve um método poderoso para a construção automática de reconhecedores sintáticos com operação *hierarquizável*, fundamentado no uso do autômatos de pilha estruturados. Tal método possibilita a obtenção de reconhecedores diretamente a partir de descrições convencionais, livres de contexto, de uma linguagem, e a imposição de limitantes para certos parâmetros do reconhecedor que efetua o tratamento da linguagem. O método explora o fato de que os textos a serem analisados são sempre finitos na prática, e que, dada uma aplicação existe sempre um limite, além do qual os recursos dos reconhecedores gerais raramente seriam utilizados.

Este modelo permite explicitamente representar os elementos que podem tornar não-regular uma linguagem livre de contexto, armazenando-os em uma memória organizada como pilha, e permitindo que os demais símbolos da cadeia, que não participam da parte não-regular da sentença, sejam tratados como se constituíssem sentenças de uma sub-linguagem regular, utilizando para isso mecanismos similares aos autômatos finitos. O formalismo resultante é capaz de aceitar qualquer linguagem livre de contexto determinística em tempo linear. (NETO, 1987), (NETO, 1993), (NETO; PARIENTE, 1999).

O autor, em seguida, introduz a conceituação do autômato e do transdutor adaptativo, modalidades de dispositivos de reconhecimento e transdução sintática, os quais incorporam a possibilidade de auto-modificação ao longo de sua operação, estendendo-se e adaptando-se às necessidades de cada particular texto de entrada que lhes seja submetido.

Os dois grandes resultados relatados projetam-se na área de linguagens extensíveis, bem como na área do tratamento de linguagens que exibem *dependências de contexto.*

O dispositivo aí proposto é o autômato adaptativo, dispositivo reconhecedor com poder de máquina de Turing, portanto um modelo abrangente de computação, o qual pode ser empregado como mecanismo de definição e de representação para linguagens naturais.

Os autômatos adaptativos incorporam ainda, na forma de transições adaptativas, um *mecanismo básico de aprendizagem*, com o qual tais máquinas se tornam capazes de memorizar informações acerca do histórico da sua operação, bem como de alterar seu próprio comportamento em função desse aprendizado.

Em (NETO, 1998) está apresentada uma coleção de soluções para alguns problemas de alta complexidade, cujas alternativas por métodos clássicos podem se mostrar inadequadas ou excessivamente ineficientes.

Essa publicação mostra que, empregando-se o formalismo adaptativo, a solução dos problemas apresentados pode alcançar o desempenho O(n).

Os problemas em questão são: reconhecimento eficiente de palíndromes, *a elaboração de aceitadores para linguagens dependentes de contexto*, aceitadores de anagramas constituídos por um conjunto finito de símbolos, aceitadores para seqüências incorporadas assincronamente. Também discute o problema dos aceitadores para linguagens ambíguas e do aceitador simultâneo para um conjunto de linguagens.

Em (IWAI, 2000), é proposto um formalismo dual ao dos autômatos adaptativos, visando facilitar o desenvolvimento de linguagens complexas ou outras aplicações que necessitem especificar linguagens dependentes de contexto na forma de gramáticas.

Trata-se das Gramáticas Adaptativas. É apresentada uma notação para esse formalismo gramatical, que pode ser empregada como metalinguagem de entrada para uma ferramenta que permite transformar uma gramática no autômato equivalente e vice-versa. Iwai também formula teoremas que provam a equivalência entre os autômatos adaptativos e as gramáticas adaptativas.

Os aspectos de implementação prática de uma ferramenta dessa natureza foram posteriormente explorados em (RICCHETTI, 2005)

Muitas são as vantagens proporcionadas pelos autômatos adaptativos na área da Computação (NETO, 2003), destacando-se:

- Resgate do caráter sintático das dependências de contexto;
- Inclusão de recursos conceituais suficientes para dar ao seu usuário a opção de dispensar estruturas explícitas de dados, geralmente empregadas nos métodos convencionais de análise e de compilação de linguagens de programação usuais;
- Eliminação da distinção e separação entre os tratamentos léxico, sintático e semântico de uma linguagem, proporcionando descrições mais uniformes e completas da mesma;
- Tratamento sintático dos escopos dos nomes – incorporado à análise léxico-sintática, o tratamento de escopos para os nomes pode passar a ser efetuado pela manipulação da estrutura topológica do autômato, criando-se ou bloqueando-se acessos às partes do mesmo que representam elementos da linguagem sujeitos a regras de escopo, de acordo com a visibilidade contextual de tais elementos ao longo da análise;
- Tratamento sintático dos atributos associados aos nomes;
- Tratamento sintático da declaração e da chamada de procedimentos em linguagens de programação;

No que diz respeito ao tratamento de Linguagens Naturais, (TANIWAKI, 2001) demonstra a capacidade que têm os autômatos adaptativos de representarem o conhecimento de forma análoga à encontrada em formalismos consagrados dessa área, tais como ATN, DCG, Frames, XG, RTN. Para verificar a viabilidade da utilização dos formalismos adaptativos, no procedimento de análise sintática de linguagem natural, a autora apresenta propostas de algoritmos de mapeamento do formalismo ATN e da Gramática Baseada em Restrição para o Formalismo Adaptativo equivalente.

Em (MENEZES, 2000), foram discutidos diversos aspectos da aplicação prática da tecnologia adaptativa, tanto no projeto como na implementação de etiquetadores morfológicos para linguagens naturais.

Dos seus primórdios, em 1985, aos dias de hoje, marcado significativamente pela introdução do formalismo dos autômatos adaptativos, inúmeras dissertações, teses, ferramentas (PEREIRA; NETO, 1997), (PEREIRA, 1999), (PISTORI, 2003), (PISTORI; NETO, 2003a), (PISTORI; NETO, 2003b), e aplicações em software se desenvolveram no Laboratório de Tecnologia Adaptativa do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo.

Atualmente os principais projetos do Laboratório ligados ao tema desta tese incluem: a representação e manipulação do conhecimento, a representação e processamento de linguagens naturais, a tradução automática de idiomas, a inferência gramatical, os sistemas inteligentes de tomada de decisão, e a busca semântica de informação.

Em todos esses assuntos, o uso de tecnologia adaptativa, em particular o emprego de autômatos adaptativos, tem se mostrado muito adequado, visto que as técnicas adaptativas se mostram bastante adequadas na representação e manipulação de fenômenos como esses, de natureza inerentemente dinâmica.

1.2 Objetivos e método utilizado

A pesquisa relatada nesta tese tem como principal objetivo, identificar alguns caminhos para a aplicação da potencialidade dos formalismos adaptativos – em particular os transdutores adaptativos e a gramática adaptativa – no tratamento dos aspectos dependentes de contexto da linguagem natural, e propor uma arquitetura adaptativa para processamento de linguagem natural

Esta arquitetura adota para os aspectos livres de contexto, pertinentes à estrutura superficial da sentença, o formalismo fundamentado em autômatos de pilha estruturados, e para os aspectos mais complexos das linguagens, mecanismos adaptativos, que permitem explorar ao máximo as potencialidades de desempenho do mecanismo não adaptativo subjacente (NETO, 1993), (NETO, 1998)

1.3 Histórico desta Pesquisa

Em (NETO; MORAES, 2002), relatou-se uma visão preliminar do tratamento de não-determinismos e ambigüidades

Em um exemplo ilustrativo indicou-se a construção de um desses autômatos a partir de uma gramática que define uma aproximação livre de contexto de um pequeno, porém significativo, subconjunto da língua portuguesa.

Em (NETO; MORAES 2003), empregou-se o formalismo gramatical adaptativo no tratamento computacional do problema da concordância nominal. Na ocasião, identificou-se também a aplicabilidade do mesmo método ao tratamento do problema de coordenação, o que levou à determinação de uma solução particular para o tratamento destes problemas.

Procurou-se, a partir de então, investigar o processamento das dependências referenciais (anáforas, pronomes e expressões-R), conjugado àquele referente aos não-determinismos e ambigüidades; partindo-se do pressuposto que o inventário das categorias vazias é paralelo ao das categorias fonéticas(RAPOSO, 1998)., vislumbrava-se que estes problemas tambem seriam resolvidos Constatou-se que o processamento adequado do c-comando (FRANK, 2001), (FRANK, 2004) é crucial.

Tal fato suscitou a busca por uma solução adaptativa para o processamento do c-comando e por um mecanismo adaptativo de cópia identificado em (NETO; MORAES, 2002) como fundamental no tratamento de não-determinismos e ambigüidades.

A partir de então, procurou-se em (NETO, 1993), e (NETO, 1998), onde se apresenta a resolução de problemas que ocorrem no processamento de linguagens de programação, analisar as características inerentes ao formalismo, no que tange à representação e processamento de dependências de contexto.

De fato, o exame detalhado dos mecanismos apresentados nessas referências, permitiu a proposição de uma arquitetura adaptativa para processamento em Linguagem Natural. Como desdobramento, vislumbrou-se uma representação da Linguagem Natural para o usuário desta arquitetura.

1.4 Estrutura da Tese

O primeiro capítulo desta tese faz uma apresentação da pesquisa, aduzindo as suas motivações e finalidades e a descrição da organização do texto que a compõe.

O segundo capítulo contém a exposição de aspectos dos formalismos gramática e autômatos de pilha, cujos desdobramentos foram identificados como primordiais no tratamento de linguagens dependentes de contexto.

Também relatam-se aqui as conclusões obtidas do estudo detalhado de alguns exemplos apresentados em (NETO, 1993): sua aplicação em tarefas como a análise de concordâncias, a delimitação de estruturas coordenadas e subordinadas.

No terceiro capítulo propõe-se uma arquitetura para Processamento em Linguagem Natural, que se fundamenta na existência de um transdutor adaptativo cuja configuração inicial é tal que permite a sua expansão na ocasião em que regras da gramática de uma

Linguagem Natural são fornecidas ao sistema. Como resultado desta expansão, o transdutor atinge uma configuração dotada de uma camada capaz de realizar as tarefas de análise léxica e sintática de sentenças desta Língua.

No quarto capítulo, detalha-se o processamento da especificação de linguagem natural empregando os mesmos mecanismos adaptativos descritos em (NETO, 1993). Justifica-se a adoção desta estratégia pelo desempenho de cada uma destas técnicas.:é: proporcional ao número de símbolos presentes na sentença a ser processada, ao número de ocorrências de determinados símbolos na fita de entrada e eventualmente ao número de símbolos do alfabeto de entrada da Linguagem tratada. Ainda, tais técnicas incluem o tratamento dinâmico e incremental de não-determinismos e ambigüidades.

No quinto capítulo, são apresentadas a descrição e o processamento de algumas dependências sintáticas em linguagem natural. As aplicações, perspectivas e conclusões desta pesquisa são também aduzidas.

Ao final, são apresentadas as referências bibliográficas utilizadas ao longo da realização deste trabalho.