ABORDAGENS AO PROCESSAMENTO SIMBÓLICO DA LINGUAGEM NATURAL

APPROACHES TO SYMBOLIC NATURAL LANGUAGE PROCESSING

João Luís Garcia ROSA*

RESUMO

Segundo Pereira e Grosz (1993), o Processamento de Linguagem Natural (PLN) é dividido basicamente em três abordagens simbólicas: baseada em casos, baseada em princípios e baseada em regras. Na Inteligência Artificial, existe há muito tempo um debate entre as abordagens baseadas em regras e as abordagens baseadas em casos. Já na Lingüística e na comunidade de PLN, há um debate entre abordagens baseadas em regras e baseadas em princípios. Este debate contrasta, por exemplo, descrições de regras de estrutura de frase da sintaxe da linguagem natural em que as regras são específicas da linguagem, com abordagens baseadas em princípios nas quais um conjunto de princípios independentes da linguagem são modulados por certos estabelecimentos de *parâmetros*, por exemplo, com respeito a ordem da palavra, para caracterizar linguagens naturais particulares. O objetivo deste trabalho é discutir estas três abordagens de forma comparativa.

Palavras Chave: Processamento de Linguagem Natural, Lingüística Computacional, Abordagem Simbólica, Inteligência Artificial.

ABSTRACT

According to Pereira and Grosz (1993), Natural Language Processing (NLP) is divided basically into three symbolic approaches: case-based, principle-based, and rule-based approach. In Artificial Intelligence, there is for a long time a debate between rule-based approaches and case-based approaches. In Linguistics and in the NLP community, there is a debate between rule-based and principle-based approaches. This debate contrasts, for instance, natural language syntax descriptions of phrase structure rules in which rules are language dependent, with principle-based approaches in which a set of language-independent principles are modulated by some *parameters*, for instance, in relation to the order of words, to characterize particular natural languages. The aim of this paper is to discuss these three approaches in a comparative way.

Keywords: Natural Language Processing, Computational Linguistics, Artificial Intelligence

* Professor do Instituto de Informática da PUC-Campinas e doutor em Lingüística Computacional pelo IEL-Unicamp. E-mail: joaol@ii.puc-campinas.br

1. INTRODUÇÃO

Pereira e Grosz (1993) dividem o Processamento de Linguagem Natural (PLN) em três abordagens simbólicas: baseada em casos, baseada em princípios e baseada em regras. Mas, o que são casos, princípios e regras? Um *caso* é uma associação entre uma situação prototípica e a informação relevante à tarefa que a segue. Por exemplo, um caso pode representar uma sentença da linguagem natural envolvendo o verbo principal *dar* e alguma outra informação desta sentença, por exemplo, que depois da ação descrita pelo verbo, o agente da ação não tem mais a posse do "paciente" da ação. O raciocínio baseado em casos envolve a descrição de analogias entre situações observadas recentemente e casos relevantes e o uso da informação de tarefa associada para determinar as inferências apropriadas às novas situações.

Um *princípio* é uma restrição aos tipos de situações possíveis: permite que um sistema infira características de situações adicionais a partir de outras características observadas. Por exemplo, um princípio na sintaxe da linguagem natural requer que cada sintagma nominal em uma sentença preencha exatamente uma posição argumental de um item lexical com posição argumental tal como um verbo ou uma preposição. Tal princípio restringe as associações possíveis entre itens lexicais com posição argumental e sintagmas nominais e portanto restringe a faixa de significados que podem ser expressos por uma determinada sentença.

Uma *regra* especifica como certas características de, ou relacionamentos entre, situações seguem de outras. Por exemplo, de novo na sintaxe da linguagem natural, uma regra de algumas línguas estabelece que um sintagma nominal (SN) seguido por um sintagma verbal (SV), havendo concordância em gênero e número, pode formar uma sentença (S), com o SN como sujeito e o SV como predicado.

2. O RELACIONAMENTO ENTRE REGRAS E CASOS

O contraste entre as abordagens baseada em regras e baseada em casos está essencialmente na fonte de generalidade de um sistema. Em sistemas baseados em regras, a generalidade vem da escolha de primitivas descritivas que permitem grandes coleções de situações com resultados similares a serem identificados e trabalhados por regras; em contraste, a generalidade em um sistema baseado em casos vem dos procedimentos de recuperação de caso e unificação (*matching*) que determinam o resultado para uma situação nova a partir de resultados para casos similares armazenados. Permitindo noções de unificação parcial ou aproximada, os sistemas baseados em casos são freqüentemente capazes de agir mesmo quando seu conhecimento de caso não unifica totalmente com a situação sob análise. Por outro lado, as regras previamente projetadas podem resumir e identificar eficientemente os itens comuns em grandes conjuntos de casos, tornando então o conhecimento do sistema mais largamente aplicável.

A utilidade de uma abordagem baseada em casos depende crucialmente da eficiência dos mecanismos de aquisição e uso da informação específica sobre a distribuição das situações de interesse. No PLN, tais situações envolvem objetos lingüísticos tais como palavras ou unidades fonéticas em determinados contextos. Enquanto as abordagens baseadas em casos devem ser avaliadas por sua habilidade de aprender casos relevantes, generalizá-los apropriadamente e aplicá-los, a falta de seleção de caso e de métodos de generalização efetivos força os praticantes atuais a criarem a maior parte da informação de caso a mão. Dado isto, os problemas mais importantes enfrentados por estes sistemas são a escolha dos traços (*features*) de caso relevantes à seleção de caso, reconhecimento dos casos que se aplicam a uma situação dada e a construção de interpretações para enunciados (*utterances*) complexos a partir de combinações de casos apropriados unificando partes do enunciado.

Enquanto as abordagens ao PLN baseadas em casos têm muito da sua inspiração a partir das idéias da ciência cognitiva que trata da organização da memória e inferência do senso comum, as abordagens baseadas em regras derivam na maior parte das tradições fortes da lingüística e da teoria de linguagens formais. Estas origens têm levado a

arquiteturas de sistema centradas sobre as noções da descrição estrutural e da transdução estrutura a estrutura. Por exemplo, as regras de estrutura de frase são usadas para descrever a sintaxe da linguagem natural e regras adicionais em cascata são então usadas para transformar tais descrições estruturais, através de uma sucessão de representações intermediárias, em uma representação do conteúdo das sentenças originais. Várias representações têm sido usadas, incluindo fórmulas lógicas, redes semânticas e quadros (*frames*). Enquanto as arquiteturas baseadas em regras têm produzido sistemas de processamento de linguagem muito expressivos, elas têm encontrado sérias dificuldades na área da robustez, isto é, a habilidade de produzir saída útil mesmo diante de regras muito específicas ou ausentes e de tratar com fenômenos não composicionais, ou seja, situações nas quais a saída apropriada numa situação complexa não pode ser derivada por uma regra simples a partir das saídas para suas partes.

3. O RELACIONAMENTO ENTRE REGRAS E PRINCÍPIOS

Um outro conjunto de dificuldades com sistemas baseados em regras no PLN surge da rigidez e especificidade das regras. Por exemplo, com a exceção de alguns sistemas recentes que usam formalismos de regra baseados em restrições declarativas e estratégias sofisticadas de aplicação de regras, as considerações baseadas em regras do mapeamento sintaxe-significado são tipicamente unidirecionais; portanto, evita-se o uso das mesmas regras para interpretação e geração da linguagem. Mais fundamentalmente, os sistemas a regras são específicos da linguagem e da construção, portanto requerem esforço maior para serem transportados para outras línguas ou mesmo para outras partes da mesma língua ou outros domínios.

Estas dificuldades podem ser vistas como sintomas da restrição da noção usual de regra, que força uma definição gerativa do relacionamento entre análises e interpretações. Por exemplo, um sistema que mapeia análise sintática para fórmulas lógicas que representam significados da sentença, teria tipicamente uma regra da gramática estabelecendo que uma sentença como “um estudante fez todo o teste” é composta de um sintagma nominal sujeito (“um estudante”) seguido por um sintagma verbal predicado (“fez todo o teste”). Associado com esta regra da gramática haveria uma regra de interpretação estabelecendo que o significado da sentença é igual ao significado do sujeito aplicado ao significado do predicado. No nosso exemplo, o significado do sujeito poderia ser uma fórmula que pode ser explicada como “verdadeira para qualquer propriedade que tem algum estudante”, e o significado do predicado como uma fórmula que podemos explicar como “propriedade de fazer todo o teste”. A interpretação resultante para a sentença poderia então ser explicada como “existe um estudante que tem a propriedade de ter feito todo o teste”. Esta interpretação força o quantificador do sujeito ter escopo mais largo do que o quantificador do objeto. Mas, para adequadamente manipular linguagem natural, o processo de interpretação precisa considerar escopos alternativos antes de escolher aquele que é contextualmente mais apropriado.

A causa fundamental deste problema é que regras privilegiam conexões gerativas particulares entre evidência e interpretação. Em contraste, a evidência específica que pode ser extraída de uma situação natural tal como um enunciado (isto é, a estrutura sujeito-predicado no exemplo anterior) é muito indeterminada para ser confiavelmente modelada como uma transdução entre um domínio de descrições estruturais e um domínio de interpretações.

Em contraste às abordagens baseadas em regras, nas abordagens baseadas em princípios, os princípios fornecem restrições fundamentais gerais entre tipos de evidências em diferentes níveis de descrição. A análise da linguagem, interpretação ou geração é vista não como um processo de rescrita, mas como uma busca pela hipótese que melhor explica a evidência observada; o espaço de busca é implicitamente definido pelos princípios e por restrições de domínio.

Numa visão baseada em princípios, em processos com componente perceptual como o processamento da linguagem, as restrições impostas pelo sistema natural fundamental são uma fonte crucial de princípios. Na linguagem natural, uma visão correspondente (devido à teoria lingüística de princípios e parâmetros de Chomsky e seus seguidores) assegura que os princípios, ancorados em critérios determinadores evolucionários tal como

aprendizagem, eficiência comunicativa e carga cognitiva, provêem um sistema de regularidades "legais" que definem os espaços de representações possíveis nos vários níveis relevantes de descrição e as restrições entre estes níveis. Na linguagem, estes níveis incluem sintaxe, semântica, discurso e prosódia, ainda que o trabalho de PLN baseado em princípios tem se concentrado somente nos níveis sintático e lexical. Abstratamente, os princípios não são apenas independentes de modelos de processamento específicos mas também de línguas particulares ou domínios de discurso. Entretanto, para ser usado na prática, um sistema baseado em princípios deve ser "preenchido" com conhecimento sobre objetos particulares - línguas, palavras, conceitos - nos termos dos princípios. Noções de aprendizagem têm um papel importante na concepção das teorias baseadas em princípios; entretanto, algoritmos efetivos para aprendizagem do conhecimento específico necessário ainda não estão disponíveis. Por ora, o conhecimento específico deve ser descrito à mão como também é o caso dos sistemas baseados em regras e baseados em casos, mesmo que a generalidade dos princípios em algumas instâncias permita especificação mais concisa do conhecimento específico.

Enquanto nas abordagens baseadas em regras tem-se uma regra para cada construção, nas abordagens baseadas em princípios, considera-se alguns princípios apenas, que combinados, atendem a qualquer construção. Os princípios considerados são (Berwick, 1992): teoria X-Barra, filtro de caso, critério temático, *move-α*, teoria de vestígios e teoria de ligação. Apesar de, aparentemente, este tipo de abordagem parecer ineficiente computacionalmente, é bem interessante pelas seguintes razões:

1. o mesmo conjunto (pequeno) de princípios pode ser recombinado várias vezes, de diferentes formas, resultando em muitas sentenças de superfície, e variando os parâmetros, diferentes dialetos e línguas;
2. princípios abstratos e heterogêneos, estabelecidos como um conjunto de restrições declarativas, ao contrário de uma representação mais uniforme como um conjunto de regras livres de contexto;
3. ênfase na importância do léxico, fonte de restrições de papel temático e variação de línguas particulares.

4. *PARSER* BASEADO EM PRINCÍPIOS

Segundo Crocker (1991), as abordagens tradicionais ao PLN podem ser consideradas baseadas em construção. Isto é, elas empregam regras específicas da linguagem orientadas à superfície, ou na forma de Redes de Transição Aumentadas (ATN, de Woods, 1970), gramáticas lógicas ou algum outro formalismo de gramática ou *parsing* (Pereira e Warren, 1980). Os problemas de tais abordagens são claros, pois envolvem grandes conjuntos de regras, freqüentemente *ad hoc*, e sua adequação com respeito à gramática da linguagem é difícil de assegurar. Em contraste, esforços na pesquisa lingüística têm observado certas regularidades nas linguagens naturais. De fato, uma tentativa inicial foi caracterizar esses universais lingüísticos, que definiriam a classe das linguagens naturais, resultando na teoria da Gramática Universal (GU). A melhor teoria da GU desenvolvida por Chomsky e outros foi um paradigma de Princípios e Parâmetros, freqüentemente referido como a teoria da Regência e Ligação (GB para *Government and Binding*). GB é uma teoria dedutiva e modular da gramática que trabalha com vários níveis de representação relacionados por uma regra transformacional, a *move-α*. A aplicação de *move-α* é restringida pela interação de vários princípios que agem como condições às possíveis representações ou derivações. Associados com os princípios estão os parâmetros que dão conta das variações entre as línguas. Então a gramática para uma determinada língua é especificada pelo estabelecimento de parâmetros apropriados e por um léxico.

As abordagens baseadas em princípios não são apenas independentes de modelos de processamento específico mas também de linguagens ou domínios de discurso particulares.

Crocker (1996), no contexto da teoria da gramática dos princípios e parâmetros assumida, discute a noção "baseado em princípios": um modelo que usa os princípios da gramática diretamente na recuperação de uma análise sintática. Isto é, ele *não* usa uma

gramática compilada, transformada[1]. Ou seja, existem várias teorias de performance[2] que podem ser consideradas baseadas em princípios, no sentido de que elas usam os princípios da gramática *on-line*, mas que tomam decisões na base de critérios 'não lingüísticos', tal como eficiência computacional ou complexidade representacional. Tais modelos claramente contrastam com teorias de processamento que operam de acordo com estratégias baseadas em gramática, sugerindo um relacionamento mais próximo entre *parser* e gramática. Uma teoria de performance que é baseada em princípios e incorpora estratégias que são baseadas em gramática (no sentido descrito), é chamada de 'fortemente baseada em princípios'.

5. APLICAÇÕES DO *PARSING* BASEADO EM PRINCÍPIOS

Cornell (1993) mostra que o interesse de se trabalhar com abordagens baseadas em princípios se justifica também pelas conseqüências teóricas interessantes, como por exemplo, uma visão mais clara do lugar da teoria GB dentro do universo do formalismo lingüístico. Ele trabalha com sistemas de princípios gramaticais através do problema mais simples do *parsing* com sistemas de condições de licenciamento[3]. O autor desenvolve um formalismo para expressar as Gramáticas de Licenciamento, isto é, gramáticas construídas inteiramente de condições de licenciamento, e explora a interpretação de tais gramáticas como sistemas de restrições (sua interpretação tradicional) e como sistemas de produção, tratados como regras de rescrita.

O trabalho de Stabler Jr. (1993) trata da abordagem baseada em princípios onde os mesmos são estabelecidos como axiomas de uma linguagem de primeira ordem. Stabler usa a resolução no seu *parser* baseado em lógica de primeira ordem que usa a teoria lingüística transformacional de Chomsky: gera árvores para DS, SS e LF a partir da entrada PF [4]. Stabler frisa que o Prolog puro não pode resolver todos os problemas da linguagem natural. O Prolog puro usa a resolução SLD, ou seja, cláusulas definidas onde somente um literal pode ser positivo[5]. Em outras palavras, numa implicação, só se pode ter um conseqüente. Ele propõe uma teoria enriquecida chamada de G1, onde pode-se concluir que certas construções *não* são sentenças, coisa que só cláusulas de Horn não conseguem. Para isto, Stabler tenta especificar as propriedades computacionais de seu modelo, ou seja, necessidade de componentes não-Horn para o problema do *parsing*:

1. princípio da categoria vazia: nenhum constituinte pode estar vazio.
2. filtro de caso: não falta caso para nenhum SN não vazio.
3. subjacência: nenhuma cadeia pode cruzar mais de uma "barreira".
4. c-comando.

Uma outra aplicação para o *parsing* baseado em princípios é o trabalho de Mchale (1995). Neste artigo, o autor propõe uma abordagem combinada entre um *parser* baseado em princípios e um dicionário legível por máquina semanticamente. O *parser* é implementado com a GB do Chomsky e sua cobertura sintática é uma função do tamanho e riqueza de seu léxico (extraído do *Longman's Dictionary of Contemporary English*) e enriquecido semanticamente usando o *Roget's International Thesaurus*.

Sua pesquisa investiga:

(1) o impacto de usar um dicionário legível por máquina como o léxico para um *parser* baseado em princípios;

(2) a extração automática de papéis temáticos deste dicionário; e

[1] Um exemplo disto é o *parser* de Marcus (1980), que computa uma estrutura de superfície, incluindo relações antecedente-vestígio. O *parser* faz isto sem tornar explícito o uso dos princípios da gramática, mas no entanto ele os obedece. Este *parser* é 'fracamente baseado em princípios'.

[2] *Competência* se refere ao nosso conhecimento da língua e *performance* ao modo como usamos este conhecimento.

[3] Condição de licenciamento é uma condição da sintaxe das linguagens naturais que permite que uma determinada construção sintática possa ser considerada como gramatical.

[4] DS = Estrutura-D; SS = Estrutura-S; LF = Forma Lógica; e PF = Forma Fonética.

[5] Cláusulas de Horn são disjunções de literais da lógica que contêm no máximo um literal positivo. Cláusulas Definidas são cláusulas de Horn que possuem um literal positivo, ou seja, uma única conclusão definida para a implicação. Um programa Prolog puro consiste de cláusulas definidas.

(3) métodos para enriquecer estes papéis usando o *Roget's*.

Crocker (1996) também aplica o conceito de *parsing* baseado em princípios no seu processador de sentenças. Baseado no conceito de modularidade da mente e na natureza da competência e performance lingüísticas, o autor defende duas hipóteses fundamentais sobre a arquitetura e princípios básicos do mecanismo de processamento de sentenças humano: a modularidade, que diz que o processador de sentenças constitui um sistema distinto dentro da faculdade da linguagem e a incrementalidade, que diz que a operação do processador de sentenças e seus módulos constituintes é determinada pelo princípio da compreensão incremental, que assegura que o máximo uso da informação lingüística é possível, assim que cada palavra de um enunciado é encontrada.

Ao invés das teorias de sintaxe serem baseadas em sistemas de regras grandes e complexos, estas teorias dependem da interação de um pequeno conjunto de princípios universais que são parametrizados entre as línguas. Fong (1992a) descreve um sistema de *parsing* baseado em princípios que alcança cobertura lingüística substancial e eficiente enquanto mantém um nível de representação de princípios próximo ao usado na literatura da lingüística. Usando um pequeno conjunto de vinte e cinco princípios, o sistema demonstra dar conta corretamente de centenas de construções diferentes a partir de um livro texto introdutório de lingüística. Também ilustrando a natureza universal das teorias baseadas em princípios, o mesmo conjunto de princípios demonstra cobrir exemplos de dados do japonês assim como do inglês. O autor também investiga o problema da configuração de princípios para *parsers* eficientes. O sistema de *parsing* incorpora parâmetros de controle flexíveis independentes das definições de princípios. Estes parâmetros de controle efetivamente definem uma família de *parsers* que incorporam o mesmo conhecimento lingüístico, mas com diferentes características de performance. Fong, através da investigação do efeito das variações nos estabelecimentos de controle, obtém uma caracterização das propriedades computacionais relevantes dos princípios que determinam as configurações de controle mais apropriadas para um *parsing* eficiente.

Merlo (1995) discute o problema relacionado à eficiência dos *parsers* baseados na GB. Ela argumenta que a GB não é uma teoria computacionalmente modular e por isso, os *parsers* baseados nesta teoria lingüística não são eficientes. Ela diz que um *parser* eficiente e confiável pode ser construído tirando vantagem da forma como os princípios são estabelecidos. Para sustentar este ponto de vista, duas características de um *parser* implementado são discutidas. Primeiro, configurações e informação lexical são pré-compiladas separadamente em duas tabelas (uma tabela X-Barra e uma tabela de co-ocorrência lexical). Segundo, a pré-computação de traços sintáticos (papéis theta, caso, etc.) resulta em computação eficiente de cadeias, porque reduz muitos problemas de formação de cadeia para uma computação local, evitando busca extensiva da árvore para um antecedente ou *backtracking* extensivo. A autora mostra também que este método de construção de dependências de longa distância pode ser computado incrementalmente.

Fong (1992b) construiu um *parser* baseado em lógica, o PO-*PARSER* (*parser* de ordenação por princípios) para investigar e demonstrar os efeitos da ordenação de princípios. O PO-*PARSER* foi propositadamente construído de uma forma altamente modular para permitir uma flexibilidade máxima na exploração de ordenações alternativas de princípios. Por exemplo, cada princípio é representado separadamente como uma operação atômica do *parser*. Uma estrutura é considerada bem formada apenas se ela passa por todas as operações do *parser*. O escalonamento das operações do *parser* é controlado por um mecanismo dinâmico de ordenação que tenta evitar trabalho desnecessário através da eliminação de estruturas mal formadas o mais rápido possível. Apesar da preocupação principal ser a exploração das propriedades computacionais dos princípios para construir *parsers* mais eficientes, o PO-*PARSER* é também capaz de tratar uma grande variedade de fenômenos lingüísticos, como os princípios da teoria theta, teoria de caso, teoria de ligação, subjacência, princípio da categoria vazia (ECP), movimento em nível da forma lógica etc.

Tradicionalmente, a informação semântica em léxicos computacionais é limitada a noções tais como restrições selecionais ou restrições específicas de domínio, codificada numa representação

'estática'. Esta informação é tipicamente usada no PLN por um simples mecanismo de manipulação de conhecimento limitado à habilidade de unificar instâncias de palavras relacionadas estruturalmente. O dispositivo mais avançado para estruturar a informação lexical é a herança, para os níveis objeto (itens lexicais) e para os meta níveis (conceitos lexicais) do léxico. Pustejovsky e Boguraev (1993) apresentam uma visão de um léxico computacional através da descrição de uma teoria de semântica lexical. Esta teoria faz uso de uma representação do conhecimento que oferece um vocabulário mais rico e mais expressivo para a informação lexical.

6. *PARSER* BASEADO EM CASOS

Para conectar o texto de entrada ao conhecimento e metas prévios do entendedor, deve-se ter um modelo no qual o acesso ao conhecimento e metas prévios seja uma parte integral do processo de *parsing*. O *parsing* baseado em casos atende a esse requisito. O objetivo de um *parser* baseado em casos é reconhecer quais estruturas de memória já existentes são mais relevantes à entrada, onde esta "relevância" é determinada pelos planos e metas do entendedor. Esta abordagem difere dos modelos tradicionais de *parsing* que tenta construir uma análise sintática ou uma estrutura de significado conceptual para um texto. O *parsing* baseado em casos é primariamente um processo de *reconhecimento* (Martin, 1989).

Como o *parsing* baseado em casos objetiva uma meta diferente dos outros *parsers*, o algoritmo para um *parser* baseado em casos também é diferente. Certos aspectos do algoritmo codificam conhecimento sintático ou conceptual, mas o algoritmo tem a principal preocupação de prover acesso às estruturas de memória preexistentes o mais cedo possível no curso do PLN. Esse acesso às estruturas de memória é essencial ao entendimento. Portanto, a organização da memória é fundamental a um *parser* baseado em casos. Este deve ser a ponte entre os itens lexicais primitivos da entrada e as estruturas de memória direcionadas identificadas como a saída do processo de entendimento. O *parsing* baseado em casos depende dos planos e metas idiossincráticos do entendedor.

Um texto pode - e deve - se referir a muitas estruturas de memória, sendo que cada estrutura é uma caracterização diferente da entrada nos termos relacionados à meta. A noção de um único significado para um texto é abandonada no *parsing* baseado em casos.

Os *parsers* convencionais, que constróem uma representação do significado de um texto de entrada, geralmente retornam uma representação como a saída do processo de *parsing*. Um *parser* baseado em casos, entretanto, pode ser chamado a reconhecer múltiplas estruturas de memória no curso do processamento de um texto de entrada. O que é significativo sobre essas estruturas de memória não são suas representações individuais, mas suas conexões com outras estruturas de memória que podem também ser relevantes ao texto. O conjunto de expectativas e referências do *parser* determina quais estruturas de memória serão reconhecidas. É este conjunto de expectativas e referências que constitui um resultado do processo do *parsing* bem sucedido.

A saída de um *parser* baseado em casos pode ser caracterizada como *um novo estado de memória*. Algumas estruturas terão sido referenciadas por um texto de entrada ou processo de inferência e algumas serão esperadas. Ainda que novas estruturas de memória sejam adicionadas, quando a informação específica já não estiver na memória, é o estado de referência e expectativas que constitui a saída real do sistema.

Um *parser* baseado em casos usa itens lingüísticos tais como palavras individuais para direcionar o processo de busca aos conceitos na memória. A tarefa de busca na memória consiste em conectar essas referências espalhadas, achando as unidades organizacionais para a memória que melhor organizem a entrada.

O conhecimento do processamento de um *parser* baseado em casos está na forma de *expectativas*. As expectativas são baseadas na idéia de senso comum de que as pessoas são capazes de fazer previsões sobre o que deve acontecer no futuro baseado no que aconteceu no passado e na sua experiência anterior.

As expectativas são derivadas de exemplos estereotípicos do uso da linguagem,

que apontam para as unidades organizacionais da memória. Estes são os índices para um *parser* baseado em casos. Portanto, estes *parsers* usam exemplos específicos de uso da linguagem para indexar estruturas de memória.

A hipótese fundamental do *parsing* baseado em casos é que o acesso ao conhecimento prévio na forma de estruturas de memória dinâmicas, específicas do domínio, é crucial nos estágios mais iniciais do entendimento da linguagem natural. É a idéia do *parsing* através da lembrança. Realizar o *parsing* a partir de casos significa lembrar instâncias passadas do uso da linguagem tal que possam ser reconhecidas de novo e lembrar conceptualizações passadas tal que possam ser chamadas novamente. O *parsing* baseado em casos é muito diferente da análise conceptual. Questões de organização de memória, indexação e busca na memória são de importância central para um *parser* baseado em casos porque este deve operar dentro de um modelo de memória existente.

A tarefa de *parsing* é um problema de busca na memória. Conceitos não são construídos a partir de pedaços derivados da entrada. Ao invés disso, já existem conceitos que preenchem muitas necessidades do entendedor. A tarefa é usar os indícios supridos pela tarefa para localizar os conceitos mais relevantes e modificá-los quando necessário para refletir as diferenças entre o que é visto e o que já é conhecido. Já que um *parser* baseado em casos faz uso da memória, ele pode fazer uso das expectativas derivadas desta memória. Estas expectativas dirigem o processo de *parsing*.

O *parser* baseado em casos difere de outras abordagens, principalmente em relação aos seguintes pontos:

1. Ao invés de acessar uma gramática geral da sintaxe da linguagem para determinar os elementos relacionados de um enunciado, um *parser* baseado em casos captura as formas idiossincráticas nas quais a linguagem é usada para referenciar determinados conceitos na memória.
2. Ao invés de construir conceptualizações para representar o significado de um texto de entrada, um *parser* baseado em casos faz uso dos elementos de análise conceptual para direcionar o processo de busca à memória aos conceitos que organizam estes elementos.
3. Ao invés de depender da memória para resolver ambigüidades de possíveis interpretações depois do *parsing* realizado, um *parser* baseado em casos usa as metas e expectativas na memória para resolver ambigüidades da entrada durante o processo de *parsing*.

A saída de um *parsing* baseado em casos inclui mudanças nas estruturas de memória e no contexto de expectativas na memória. A moderna teoria lingüística busca "capturar as generalizações significativas" no uso da linguagem. Isto tem levado quase que exclusivamente à busca dos padrões sintáticos. O *parsing* baseado em casos, entretanto, se preocupa mais com a caracterização de como o texto se refere a conceitos.

7. APLICAÇÕES DO *PARSING* BASEADO EM CASOS

Jones (1993) propõe um novo método para o processamento de linguagem natural, chamado de processamento baseado em analogias ou em exemplos (na verdade, outro nome que se dá para a abordagem baseada em casos). O paradigma baseado em exemplos é comparado e contrastado com o processamento convencional baseado em regras e o autor discute as vantagens e desvantagens de ambos. Um dos principais temas de seu trabalho é como melhor representar exemplos em um ambiente baseado em exemplos tal que alguma capacidade gerativa possa ser dada a um sistema sem recurso às técnicas baseadas em regras, portanto reduzindo a inflexibilidade associada com a computação lingüística dirigida analogamente. É sugerido que, no campo do processamento de linguagem baseado em analogias, os pesquisadores assumam que as técnicas baseadas em regras convencionais devam ser usadas em algum ponto da arquitetura de tais sistemas. Onde o processamento baseado em regras não é usado, tenta-se realizar uma arquitetura baseada em analogias.

Kolodner (1993) cita em sua biblioteca de casos a aplicação Parse-O-Matic de Goodman (1991), que trata do *parsing* de linguagem natural. O sistema usa uma

biblioteca de casos de 50.000 micro-casos para criar uma representação semântica de questões comuns da língua inglesa. Resolução de ambigüidade do sentido da palavra e morfológica, referência pronominal e elisão são tratados de uma maneira integrada. Um micro-caso é uma ação individual realizada no curso do raciocínio. O sistema tem como ênfase o uso de casos codificados como seqüências temporais de micro-casos, comparação de eficiência, velocidade e tempo de engenharia de conhecimento para criação e manutenção de sistemas baseados em casos e baseados em regras. Este sistema foi construído na metade do tempo necessário para construir um sistema baseado em regras com eficiência comparável.

Jones e Boguraev (1987) descrevem uma aplicação de representações baseadas em casos para o processamento da linguagem. Eles usam um analisador de linguagem que constrói representações de significado expressando papéis de caso semântico; especificamente, o analisador de Boguraev (1979) constrói árvores de dependência com sentidos de palavras definidos por fórmulas 'primitivas de categoria semântica', e com rótulos de caso, isto é, 'primitivas de relação semântica', nas estruturas dos constituintes do verbo (e outras categorias). No seu estudo de casos, os autores trazem uma lista de casos, usando exemplos de sentenças, para ser usada por aqueles que desejam a aplicação no PLN.

Martin (1989) implementou um *parser* baseado em casos, o micro DMAP, que usa uma arquitetura de passagem de marcadores para identificar estruturas de memória relevantes a partir do texto de entrada e das expectativas na memória. Dois tipos de marcadores são usados no sistema: marcadores de *ativação*, que capturam informação sobre o texto de entrada e as estruturas de memória referenciadas correntemente, e os marcadores de *previsão*, que indicam quais estruturas de memória podem se tornar referenciadas. Se um marcador de previsão é *passado* a um conceito e um marcador de ativação é subseqüentemente também passado a este conceito, diz-se que existe uma *interseção* a este conceito. Os marcadores podem carregar mais informação do que a representada pelo conceito, neste caso o conceito pode ser *refinado* a um conceito mais específico. Por último, o conceito refinado é *reconhecido*.

8. CONCLUSÃO: CASOS, REGRAS E PRINCÍPIOS NOS SISTEMAS DE PLN

A faixa das organizações de sistemas constituída por princípios, casos e regras forma um *continuum* multifacetado no qual muitas opções diferentes podem ser consideradas (Pereira e Grosz, 1993). Em uma faceta, os princípios podem ser vistos como fornecedores do conhecimento inicial crucial na especificação do espaço de casos possíveis e representações apropriadas, adquiridas ou recuperadas. Os mecanismos baseados em casos podem ser usados como uma reserva, que entram em ação quando os princípios conhecidos são insuficientes para derivar a interpretação de uma situação particular ou para decidir entre interpretações alternativas compatíveis com os princípios. Em uma outra faceta, a informação derivada de caso pode ser altamente abstraída pelos projetistas de sistemas para as restrições específicas da linguagem tal como a ordem da palavra e sistemas flexionais ou representações e restrições específicas do domínio tais como aquelas que especificam as propriedades sintáticas, semânticas e de domínio de determinadas entradas lexicais. Em ainda uma outra faceta, as regras podem ser vistas como codificações orientadas computacionalmente de determinadas instâncias de princípios apropriados às tarefas ou situações particulares; como a computação direta a partir dos princípios é em geral muito difícil, as regras podem ser preferidas por razões computacionais.

9. REFERÊNCIAS

- Berwick, R. C. (1992), Principles of Principle-Based *Parsing*. In Berwick, R. C. & Abney, S. P. & Tenny, C. (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics*. 1-37. Kluwer Academic Publishers.
- Boguraev, B. K. (1979), Automatic Resolution of Linguistic Ambiguities. *Technical Report 11*, Computer Laboratory, University of Cambridge. *apud* Jones & Boguraev (1987), pág. 65.
- Cornell, T. L. (1993), Description Theory, Licensing Theory, and Principle-Based Grammars and *Parsers*. *Dissertation*

Abstracts International vol. 54, no. 2, August, 500-A.
- Crocker, M. (1996), *Computational Psycholinguistics - An Interdisciplinary Approach to the Study of Language*. Kluwer Academic Publishers.
- Crocker, M. W. (1991), A Principle-Based System for Syntactic Analysis. *Canadian Journal of Linguistics* 36 (1): 1-26. March.
- Fong, S. (1992a), Computational Properties of Principle-Based Grammatical Theories. *Dissertation Abstracts International* vol. 52, no. 10, April, 5364-B.
- Fong, S. (1992b), The Computational Inplementation of Principle-Based *Parsers*. In Berwick, R. C. & Abney, S. P. & Tenny, C. (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics*. 65-82. Kluwer Academic Publishers.
- Goodman, M. (1991), A case-based, inductive architecture for natural language processing. Unpublished paper presented at AAAI Spring Symposium on Machine Learning of Natural Language and Ontology. *apud* Kolodner (1993), pág. 608.
- Jones, K. S. & Boguraev, B. (1987), A Note on a Study of Cases. *Computational Linguistics*, Volume 13, Numbers 1-2, January-June, 65-68.
- Jones, D. B. (1993), The Processing of Natural Language by Analogy with Specific Reference to Machine Translation. *Dissertation Abstracts International* vol. 53 no. 8, February, 4218-B.
- Kolodner, J. (1993), *Case-Based Reasoning*. Morgan Kaufmann Publishers, Inc.
- Marcus, M. P. (1980), *A Theory of Syntactic Recognition for Natural Language*. The MIT Press.
- Martin, C. E. (1989), Case-Based *Parsing*. In Riesbeck, C. K. & Schank, R. C. (Eds.), *Inside Case-Based Reasoning*. 319-372. Lawrence Erlbaum Associates, Publishers.
- Mchale, M. L. (1995), Combining Machine-Readable Lexical Resources with a Principle-Based *Parser*. *Dissertation Abstracts International*, vol. 57/02-A, page 491.
- Merlo, P. (1995), Modularity and Information Content Classes in Principle-Based *Parsing*. *Computational Linguistics* vol. 21, Number 4, 515-541.
- Pereira, F. C. N. & Grosz, B. J. (1993), "Introduction" (to the Special Volume on Natural Language Processing). *Artificial Intelligence* 63, 1-15.
- Pereira, F. C. N. & Warren, D. H. D. (1980), Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks. *Artificial Intelligence* 13, 231-278.
- Pustejovsky, J. & Boguraev, B. (1993), Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence* 63, 193-223.
- Stabler Jr., E. P. (1993), *Parsing* as non-Horn Deduction. *Artificial Intelligence* 63, 225-264.
- Woods, W. A. (1970), Transition Network Grammar for Natural Language Analysis. *Communications of the ACM*, vol. 13, no. 10, October, 591-606.