

## Anexo B

# Abordagens ao Processamento Simbólico da Linguagem Natural

*“A verdade científica deve ser apresentada de formas diferentes, e deve ser considerada como igualmente científica se aparece na forma robusta e nas cores vivas de uma ilustração física, ou na tenuidade e palidez de uma expressão simbólica”.*  
atribuído a James Clerk Maxwell (1831-79), em *Physics Teacher*, 1969.

## B.1 Introdução

Pereira e Grosz (1993) dividem o Processamento de Linguagem Natural (PLN) em três abordagens: 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.

## B.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 frequentemente 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, nossa falta de seleção de caso e métodos de generalização efetivos força os praticantes atuais a criarem a maior parte da informação de caso manualmente. 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 se inspiram nas idéias das Ciências Cognitivas, que tratam da organização da memória e inferência do senso comum, as abordagens baseadas em regras derivam fundamentalmente das tradições fortes da Lingüística e da Teoria da Linguagem Formal. 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 produ-

zido 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.

### *B.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 de aplicação de regras sofisticadas, 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 língua 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 teste” é composta de um sintagma nominal sujeito (“um estudante”) seguido por um sintagma verbal predicado (“fez todo 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 teste”. A interpretação resultante para a sentença poderia então ser explicada como “existe um estudante que tem a propriedade de ter feito todo 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 específico.

Numa visão baseada em princípios, em processos com componente perceptual como o processamento da linguagem, as restrições impostas pelo sistema são uma fonte 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, provêm 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 tenha 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 manualmente 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 qualquer construção. Os princípios considerados são (Berwick, 1992): teoria X-Barra, filtro de caso, critério temático, *move-a*, 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 re combinado 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, por exemplo, de restrições de papel temático e variação de línguas particulares.

#### *B.4 Parser Baseado em Princípios*

Segundo Crocker (1991), as abordagens tradicionais ao PLN podem ser 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 língua é difícil de assegurar. Em contraste, esforços na pesquisa lingüística têm observado certas regularidades nas línguas 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, frequentemente 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-a*. A aplicação de *move-a* 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 línguas 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<sup>23</sup>. Ou seja, existem várias teorias de performance<sup>24</sup> 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 descrita como ‘fortemente baseada em princípios’.

---

<sup>23</sup> Um exemplo disto é o parser de Marcus (1980), que computa uma estrutura de superfície, incluindo relações antecedente-vestigio. 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’.

<sup>24</sup> *Competência* refere-se ao conhecimento da língua e *performance* ao modo como se usa este conhecimento.

### B.5 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 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 do 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.

## B.6 Conclusão

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.