CONCEITOS E APLICAÇÕES DA LINGÜÍSTICA COMPUTACIONAL

Gabriel de Ávila Othero

Gabriela Betania Hinrichs Conteratto

(Doutorandos em Lingüística Aplicada pela PUC-RS)

gabnh@terra.com.br

ghinrichs@uol.com.br

Resumo: *A Lingüística Computacional é uma área de natureza transdisciplinar que possui objetivos bastante diversificados, indo da criação de sistemas computacionais até a construção de representações lingüísticas com vistas à implementação ou à verificação de modelos lingüísticos. Neste trabalho, apresentaremos alguns conceitos básicos pressupostos para o desenvolvimento de aplicações computacionais.*

Introdução

Neste trabalho, iremos apresentar algumas aplicações de trabalhos que envolvem o tratamento computacional da linguagem natural. Queremos mostrar que a ciência que se ocupa com esse tipo de trabalho – a Lingüística Computacional – envolve um intercâmbio entre as áreas da Lingüística e da Informática. Pretendemos mostrar alguns conceitos-chave que devem pressupor o trabalho com o tratamento computacional da linguagem.

1. As áreas de aplicação da Lingüística Computacional

Grosso modo, podemos definir a Lingüística Computacional como a área da ciência Lingüística preocupada com o tratamento computacional da linguagem e das línguas naturais. Alguns campos da Lingüística são, por isso mesmo, fundamentais para o desenvolvimento de aplicações em Lingüística Computacional.

A Fonética e a Fonologia, por exemplo, estão presentes em diversos aplicativos de Processamento de Linguagem Natural (PLN). Elas são as áreas da Lingüística preocupadas em estudar o sons das línguas humanas. A Fonética ocupa-se do estudos dos fones, dos sons concretizados na fala. Ela está interessada na parte acústica, articulatória e fisiológica da produção dos sons da fala. A Fonologia, por outro lado, concentra-se em estudar os fonemas e o sistema fonológico subjacente de uma língua. Ela investiga o sistema abstrato que envolve o conhecimento fonológico dos falantes. Entre as aplicações que surgiram e que podem ser ainda desenvolvidas a partir de estudos fonéticos e fonológicos, podemos destacar aplicativos de reconhecimento e de geração de fala. Programas desse tipo apresentam diversas finalidades: podem servir desde meros reconhecedores de fala em um aparelho celular (o que tornará possível a discagem a partir do reconhecimento da voz do proprietário do aparelho), até programas que reconheçam a fala a ponto de digitar um texto ditado por um usuário, ou mesmo o contrário, programas que sejam capazes literalmente de "ler em voz alta" um documento escrito e armazenado no computador. Aplicações como essas com certeza facilitam o acesso de computadores pessoais e aparelhos eletrônicos para pessoas com deficiência visual e até mesmo para leigos, que poderão interagir com a máquina através da linguagem falada.

Outras áreas importantes da Lingüística para o desenvolvimento de programas de PLN são a sintaxe e a semântica. Em Lingüística, a sintaxe pode ser definida como "a parte da gramática que descreve as regras pelas quais se combinam as unidades significativas em frases" (Dubois et al, 1988: 559). Estudos em sintaxe gerativa e funcional têm alcançado bons resultados em programas que lidam com a geração automática de sentenças, como os *chatterbots*, por exemplo. Os *chatterbots* são programas desenvolvidos para "conversar" com humanos através da linguagem natural[1]. Tais programas vêm sendo desenvolvidos especialmente para trabalhar com atendimento virtual e com tutoriais educativos. Um programa desse tipo deve ser capaz de poder manter uma conversa com um humano da forma mais natural possível, daí a Semântica (que estuda o significado das palavras e proposições de uma língua) ser fundamental na programação de tais aplicativos.

Obviamente, outras áreas da Lingüística são necessárias para o desenvolvimento de *chatterbots*, como a Lingüística Textual (que se ocupa das relações intratextuais

[1] Cf. http://bots.internet.com/search/s-chat.htm e também http://cybelle.cjb.net/.

indispensáveis na produção de significado) e a Análise da Conversação (que investiga a maneira como os diálogos se organizam nas diferentes línguas do mundo), entre outras.

Além dos *chatterbots*, conhecimentos em sintaxe e semântica são também fundamentais para aplicativos como tradutores automáticos, *parsers*, geradores automáticos de resumos, corretores ortográficos e gramaticais e classificadores automáticos de documentos digitais[2].

2. A arquitetura de sistemas de PLN

Para compreender como ocorre o processamento automático da linguagem, precisa-se conhecer a arquitetura de sistemas que interpretam e geram a linguagem natural. É importante mencionar que a interpretação da linguagem natural baseia-se em mecanismos que tentam compreender frases, buscando traduzi-las para uma representação que possa ser compreendida e utilizada pelo computador.

Já na geração de linguagem, ocorre o oposto, pois o computador traduz uma representação para sua expressão em alguma língua. Ou ainda, na geração, o computador produz textos o mais próximo possível de textos produzidos por pessoas. Uma aplicação que necessita tanto da interpretação quanto da geração da linguagem é a tradução automática.

Os sistemas para PLN são geralmente modulares. Os diferentes níveis de processamento (morfológico, sintático, semântico, discursivo e pragmático) são executados em módulos distintos. Esses módulos se comunicam pela passagem de representações intermediárias do texto sob análise. Apenas o fluxo de informação muda, de acordo com a tarefa do sistema: interpretação ou geração. Nos sistemas para interpretação da LN, tem-se o texto como entrada, e uma representação formal como saída.

Todos os sistemas para o PLN utilizam as chamadas "Bases de Conhecimento", que são arquivos externos onde informações necessárias ao processamento são codificadas

[2] Cf. Cole et al. (197), Garside, Leech and McEnery (1997) e Vieira (2002).

declarativamente. Barros (1997) propõe uma figura representando cinco Bases de Conhecimento.

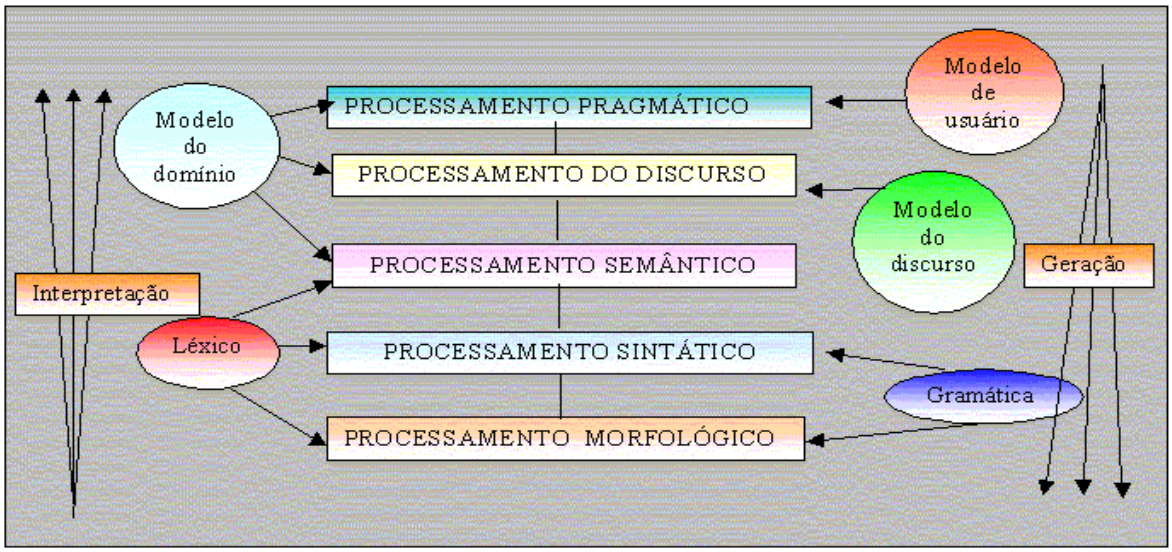

Figura (1) – Bases de Conhecimento

Na figura (1), mostram-se as cinco "*Bases de conhecimento*": o léxico, a gramática, o modelo do domínio, o modelo do usuário e o modelo do discurso. No léxico, cada palavra pode estar associada às suas características morfológicas, sintáticas e semânticas. É importante mencionar que existem vários formalismos para representar estas informações armazenadas no léxico.

Vale lembrar que a representação do léxico deve ser escolhida de acordo com a representação da gramática, pois essas duas bases de conhecimento interagem durante o processamento do texto. A gramática faz a verificação, através de regras, de quais são as cadeias de palavras válidas em uma língua. Allen (1995) chama a atenção para o fato de que essa verificação é feita em termos de categorias sintáticas, e não de uma lista exaustiva de frases, pois isso seria inviável, uma vez que qualquer língua possui um número infinito

de frases gramaticalmente aceitáveis. Para exemplificar estas duas *Bases de Conhecimento*, citam-se dois exemplos[3]:

(1) a. comprou
<categoria> = verbo
<tempo> = pretérito imperfeito
<número> = singular
<arg1> = SN
<arg2> = SN

b. mesa branca
SN→ Subs Adj
<Subst gênero> = <Adj gênero>
<Subst número> = <Adj número>

Na entrada do léxico em (1a), tem-se o verbo *comprar* na terceira pessoa do singular do pretérito perfeito. As características sintáticas indicam que ele tem dois argumentos: um SN[4] sujeito (argumento externo) e um SN objeto direto (argumento interno). Já no exemplo (1b), mostra-se um tipo de regra gramatical que traz restrições associativas, pois o SN indica que se tem um substantivo seguido de um adjetivo (modificador) e as restrições <Subst gênero> = <Adj gênero> e <Subst número> = <Adj número> determinam que o gênero e o número do substantivo e do adjetivo devem concordar.

As outras três Bases de Conhecimento de um sistema para PLN fornecem *o contexto* para o processamento de cada frase. No modelo de domínio, armazena-se o *contexto enciclopédico* (para alguns formalismos são conhecimentos a respeito das entidades, relações, eventos, etc). O modelo do usuário fornece o *contexto interpessoal*, armazenando conhecimento a respeito do sistema (os seus objetos, planos, intenções, funções) através de

[3] O formalismo usado nos exemplos é o PATR-II (Shieber, 1984) - muito usado em sistemas para PLN.
[4] Sintagma Nominal

representações como planejamento hierárquico ou atos da fala (Allen, 1995). O modelo do discurso fornece o *contexto textual.* Logo, depois de apresentar a arquitetura de sistemas de PLN, acredita-se que estes sistemas serão mais ou menos eficientes dependendo da consistência das informações lingüísticas armazenadas e organizadas nas *Bases de Conhecimento*.

3. *Parsers* sintáticos

O estudo da formalização das regras de funcionamento sintático das línguas naturais é vital para o funcionamento de diversos aplicativos desenvolvidos em PLN, especialmente a construção de *parsers*. A palavra *parsing* em si não remete ao processamento sintático mediado por computador (ou processamento sintático computacional). O termo vem da expressão latina *pars orationes* (partes-do-discurso) e tem suas raízes na tradição clássica. De acordo com Mateus & Xavier (1992: 886), *parsing* pode ser entendido como o "processo de atribuição de uma estrutura e de uma interpretação a uma sequência linguística".

No contexto da Lingüística Computacional, entretanto, *parsing* diz respeito à interpretação automática (ou semi-automática) de sentenças de linguagem natural por meio de programas de computador conhecidos como *parsers*. Esses programas são capazes de classificar morfossintaticamente as palavras e expressões de sentenças em uma dada língua e atribuir às sentenças a sua estrutura de constituintes, baseando-se em um modelo formal de gramática[5].

De acordo com Covington (1994: 42), efetuar o *parsing* de uma sentença é "determinar, por um processamento algorítmico, se a sentença é gerada por uma determinada gramática, e se ela for, qual estrutura que a gramática atribui a ela"[6]. Para Bateman, Forrest & Willis (1997: 166), autores do capítulo *The use of syntactic annotation tools: partial and full parsing* (In: Garside, Leech & McEnery, 1997),

[5] Existem também *parsers* semânticos, preocupados em formalizar a estrutura semântica das sentenças em linguagem natural, mas não trataremos deles aqui.

[6] Trecho original: "(...) to determine, by an algorithmic process, whether the sentence is generated by a particular grammar, and if so, what structure the grammar assigns it".

> um dos principais objetivos da área de PLN nos últimos dez anos tem sido produzir um "analisador gramatical", ou ***parser***, de **abrangência ampla**. Para muitos aplicativos de PLN, o desafio é produzir um *parser* que poderá ser capaz de analisar automática e estruturalmente de maneira correta, de acordo com um esquema de *parsing* definido, qualquer sentença do inglês que possa ocorrer naturalmente, **sem restrições**, de uma gama de gêneros textuais tão vasta quanto possível.[7] (grifos dos autores)

Vários *parsers* já foram desenvolvidos ao longo dos anos, porém nenhum deles foi ainda capaz de alcançar o objetivo proposto por Bateman, Forrest & Willis. Por isso, esse ainda continua sendo um dos principais objetivos de PLN, já que um *parser* com tal capacidade de análise ainda não foi criado[8].

Considerações finais

A Lingüística Computacional é uma área que vem se desenvolvendo consideravelmente nos últimos anos, pois ela desenvolve trabalhos que visam a facilitar cada vez mais a interação *homem x máquina*. Essa área "híbrida" e transdisciplinar envolve o comprometimento de pesquisadores com formação em Lingüística e em Informática. A cooperação entre esses dois campos de estudo é de crucial importância para desenvolvimentos em Lingüística Computacional, porque ela é responsável por desenvolver *aplicativos computacionais* que trabalhem com a *linguagem natural*. Como vimos, o objetivo de muitos desses aplicativos é permitir que a interação entre computador e usuário seja cada vez mais simples e amigável para o usuário. Afinal, se temos de trabalhar e lidar com computadores, acreditamos que o ideal seja que eles aprendam a se comunicar através da nossa linguagem e não o contrário.

---

[7] Trecho original: "one of the major aims of NLP over the past ten years has been to produce a **wide-range** 'grammatical analyser' or **parser**. For many NLP applications, the challenge is to produce a parser which will automatically be able to structurally analyse correctly, according to a defined parsing scheme, any sentence of naturally occurring **unrestricted** English, from as wide a range of genres as possible".

[8] Para saber mais sobre *parsers*, cf.Bick (1996), Cole et al. (1997), Othero (2004) e Menuzzi & Othero (2005).

Referências

ALLEN, James. Natural ***Language Understanding***. Benjamin/Cummings, 2nd edition, 1995.

BARROS, Flávia de Almeida; ROBIN, Jaques. ***Processamento de Linguagem Natural***. In: Congresso da Sociedade Brasileira de Computação, Recife**,** 1997.

BICK, Eckhard. Automatic parsing of Portuguese. In: GARCÍA, Laura Sánchez (Ed.). *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado*. Curitiba: CEFET-PR. 1996.

COLE, R. A. et al. (eds.). ***Survey of the state of the art in human language technology***. Cambridge: Cambridge University Press / Giardini, 1997. [www.dfki.de/~hansu/HLT-Survey.pdf].

COVINGTON Michael. A. ***Natural language processing for Prolog programmers***. New Jersey: Prentice Hall, 1994.

DUBOIS, Jean et al. ***Dicionário de lingüística*****.** São Paulo: Cultrix, 1988.

GARSIDE, Roger; LEECH, Geoffrey; McENERY, Anthony. ***Corpus annotation: linguistic information from computer text corpora***. London / New York: Longman, 1997.

MATEUS, Maria Helena Mira; XAVIER, Maria Francisca (Orgs). ***Dicionário de termos lingüísticos***. Lisboa: Edições Cosmos, 1992.

MENUZZI, Sérgio de Moura; OTHERO, Gabriel de Ávila. ***Lingüística Computacional: teoria & prática***. São Paulo: Parábola, 2005.

OTHERO, Gabriel de Ávila. ***Grammar Play: um* parser *sintático em Prolog para a língua portuguesa***. Porto Alegre: PUCRS, 2004. Dissertação de Mestrado.

VIEIRA, R. Lingüística computacional: fazendo uso do conhecimento da língua. *Entrelinhas*, ano 2, n. 4, São Leopoldo: UNISINOS, 2002.