Contributo para o processamento computacional do português: o CRdLP

Pedro Veiga e Diana Santos

O carácter cada vez mais importante que as tecnologias da informação e da comunicação desempenham no nosso dia-a-dia vai obrigar ao uso crescente de sistemas de tratamento automático ou semi-automático da língua portuguesa quer falada quer escrita.

Esta constatação tem vindo a ser feita desde há alguns anos e, no que nos diz respeito, foi em particular nos trabalhos preparatórios do *Livro Verde para a Sociedade da Informação em Portugal* que aumentou a percepção da crescente importância em investir recursos específicos para o português, em particular o português europeu. Também só na parte final da década de 90 a tecnologia informática começou a estar num nível de desenvolvimento (capacidade de processamento, miniaturização, custo, ...) que permite uma maior penetração destas tecnologias que conduza à sua massificação.

Foi assim que na preparação das áreas de intervenção na I&D para a sociedade da informação e do conhecimento se julgou imprescindível investir no desenvolvimento do processamento computacional da língua portuguesa.

O processamento computacional de uma língua é a forma principal de manipular, com o computador, o conhecimento humano, que é expresso, quase exclusivamente, em linguagem natural. De facto, é através da língua que lemos e absorvemos informação, e que a transmitimos, e é através da língua que ensinamos e aprendemos. Maioritariamente, é através dela que comunicamos com o mundo que nos rodeia.

As aplicações desta área são, portanto, múltiplas (veja-se Santos (2001) para uma panorâmica mais abrangente da engenharia da linguagem). Destacamos aqui, a título de exemplo,

- procura de informação ("information retrieval") em grandes quantidades de texto (por exemplo na Web)
- tradução automática e sistemas de apoio à tradução
- ajuda à redacção (correctores ortográficos, sintácticos, estilísticos)
- legendagem automática
- obtenção (semi)automática de informação sobre domínios especializados
- criação semi-automática de recursos lexicais, desde dicionários bilingues a tesauros
- assistentes computacionais que respondam a perguntas e forneçam ajuda não trivial a um utilizador

- sistemas de ensino (de língua e de outras disciplinas)
- criação automática de resumos ou de documentos adaptados a leitores de perfis diferentes
- obtenção de sistemas personalizados com base na história e nos interesses de cada utilizador
- indexação inteligente (e correspondente melhoria significativa da procura)
- tradução entre diferentes modalidades (fala - texto) para cidadãos com necessidades especiais ou para trabalhos ou situações que não permitam a utilização da vista nem o recurso a um teclado
- jogos didácticos – e todo o tipo de entretenimento electrónico em geral – que interagem na língua do utilizador

Não se pense, contudo, que estas aplicações são fáceis de implementar, ou que existem de momento sistemas computacionais que as desempenhem a um nível satisfatório seja para que língua for. Existe, sim, trabalho para atingir estes objectivos, primordialmente para o inglês.

É nossa convicção que um trabalho paralelo terá de existir tendo como alvo o português, visto que não é válida a hipótese ingénua de que "resolvido para uma língua, resolvido para todas": Cada língua é um sistema excepcionalmente complexo com um funcionamento próprio, e é preferível conhecer e saber medir os problemas que se pretende resolver do que adaptar métodos e sistemas de uma língua a outra (Santos, 1999c).

CRdLP – Centro de recursos distribuído para a língua portuguesa

Como uma das primeiras iniciativas, foi decidido lançar um centro de recursos distribuído para o processamento computacional da língua portuguesa. Este centro é uma estrutura que tenta contribuir para a solução de dois problemas identificados nesta área (veja-se o documento preparatório (Santos, 1999a) para a discussão da área do Processamento Computacional do português por ocasião dos debates sobre política científca que informaram o *Livro Branco para o Desenvolvimento Científico e Tecnológico (1999-2006)*):

- a falta de recursos que possam servir de base ao desenvolvimento de aplicações e ao próprio estudo da língua portuguesa
- a falta de métodos e de métricas de avaliação para comparar sistemas e para avaliar o progresso numa dada área

Este centro de recursos é **um** dos vectores de uma estratégia mais vasta a nível do Ministério da Ciência e da Tecnologia na área do processamento computacional da nossa língua e na dinamização da sociedade da informação e do conhecimento.

Os principais **objectivos** do centro podem ser resumidos da seguinte forma:

- facilitar o acesso aos recursos já existentes,
- desenvolver de forma harmoniosa e em colaboração com os interessados os recursos considerados mais prementes,

- organizar avaliações e encontros científicos que envolvam a comunidade como um todo
- velar por que os recursos encaminhados para esta área possam aproveitar ao máximo o progresso desta,
- além de manter e melhorar o portal sobre o processamento computacional do português, http://www.portugues.mct.pt, manter o catálogo de recursos e actores actualizado.

A **actividade** do centro repartir-se-á entre

1. o assegurar dos serviços básicos de repositório, distribuição e catálogo, lançando as bases para tal vir a ser feito de forma distribuída
2. o desenvolvimento de alguns recursos pelo próprio centro, sobretudo recursos para avaliação ou para calibragem
3. a organização de "avaliações conjuntas" (conferências de comparação de sistemas onde o método de avaliação é discutido e criado pelos participantes)
4. o fornecimento de uma infraestrutura para projectos colaborativos de interesse geral
5. a formação de pessoal especializado em gestão e disponibilização de recursos
6. a gestão de um programa de desenvolvimento de recursos (incluindo recursos de formação) por concurso público, em presumível articulação com a Fundação para a Ciência e a Tecnologia

O centro foi pensado como uma forma de endogeneizar certos padrões de comportamento, e normas de conduta, entre os principais actores do processamento da nossa língua. Pretende-se que seja uma estrutura temporária, mais tarde podendo vir a ser absorvida pelas próprias instituições que o albergam.

Os recursos criados serão sempre do **domínio público**, ainda que a sua disponibilização obedeça à preocupação de acautelar os interesses dos eventuais donos ou detentores dos direitos intelectuais, seja através de compensação monetária, seja pelo desenvolvimento de projectos conjuntos de utilidade mútua ou, ainda, recorrendo a outros mecanismos pertinentes.

Organização

O centro de recursos proposto é uma estrutura distribuída geograficamente, estando de momento os seguintes pólos em fase de arranque:

- Oslo (SINTEF, pólo fundador)
- Braga (Universidade do Minho, Departamento de Informática)
- Lisboa (INESC/ID, Laboratório da fala)

Prevemos a criação de novos pólos durante o ano de 2001.

Além de ser competência do centro facultar informação clara sobre como obter os recursos que o centro mantém ou a que dá acesso, todas as decisões e projectos em que o Centro se envolve serão públicas e disponibilizadas na rede para consulta de todos os interessados.

Trabalho em curso

Lançados pelo projecto Processamento computacional do português (veja-se Santos (2000) para um balanço dos primeiros anos de actividade deste), já se encontram em funcionamento algumas actividades que se enquadram no espírito do centro de recursos, e que se pretende que sejam continuadas de forma mais estruturada e distribuída no âmbito deste:

- o projecto AC/DC http://cgi.portugues.mct.pt/acesso/

Este projecto (cujo nome por extenso é **Acesso a Corpora / Disponibilização de Corpora**) pretende tornar fácil a consulta, para investigação sobre a língua, a grandes quantidades de texto em português, com um mínimo de requisitos técnicos por parte do utilizador. Esta linha de acção foi sugerida em Santos (1999b) como forma de contrariar a grande dificuldade, então identificada, de acesso a este tipo de material. O projecto AC/DC faculta o acesso através da Web a diversos corpora, codificados e processados de forma a serem enriquecidos com informação pertinente para estudos de vária índole: Assim, esses textos contêm associada indicação sobre a segmentação (separação em frases, parágrafos e outras unidades estruturais) e o tipo de texto (jornalístico, ensaio, texto literário, didáctico, etc.), assim como, para cada palavra, a categoria morfossintáctica e informação morfológica, e a função sintáctica (num formalismo de gramática dependencial), etc. (Veja-se Santos & Bick (2000) para uma descrição técnica aprofundada.)

Tal permite a um linguista interrogar um corpus formulando questões que excedem largamente a simples procura de palavras ou formas, bastando para isso ter acesso à Internet.

De momento já se encontram interrogáveis textos perfazendo mais de 230 milhões de palavras (de português de Portugal e do Brasil), incluindo texto jornalístico (a grande maioria), literário, técnico, didáctico, assim como ensaio e correspondência. O trabalho prossegue com a integração de mais material textual, em curso.

- o projecto COMPARA/DISPARA http://www.portugues.mct.pt/COMPARA/

O projecto **COMPARA/DISPARA** é um congénere do AC/DC para textos paralelos em português e inglês. O corpus COMPARA foi contudo criado de raiz (por Ana Frankenberg-Garcia), e a problemática de interrogar um texto e a sua tradução levou ao dsenvolvimento de uma interface especial para corpora paralelos, o DISPARA. Este projecto encontra-se em progresso, facultando contudo já dez pares de textos para consulta e tendo já mais de uma centena com as autorizações tratadas. (Veja-se Frankenberg-Garcia & Santos (2001) para mais informação.)

- o projecto CETEMPúblico http://cgi.portugues.mct.pt/cetempublico/

O **CETEMPúblico** (Corpus de Extractos de Textos Electrónicos MCT/Público), também acessível através do projecto AC/DC, é um recurso criado com o objectivo de ir mais além na facilitação de recursos: além de dar acesso através da Web, quisemos poder distribuir fisicamente o texto todo, de forma a poder ser manipulado e utilizado pelos investigadores e engenheiros que desenvolvem ferramentas que lidam com o português. O texto, do jornal PÚBLICO, foi dividido em extractos ordenados aleatoriamente para não permitir a reconstrução das notícias integrais, e é distribuído em CD gratuitamente a todos quantos registarem o seu endereço postal. Em Março de 2001,

data de redacção do presente texto, já mais de 150 pessoas ou grupos espalhados pelo mundo receberam este recurso. (Veja-se Rocha & Santos (2000) para documentação técnica, e a página de informações na rede, que é constantemente actualizada.)

Prevê-se para breve o início de um projecto paralelo contendo linguagem jornalística brasileira, em colaboração com o Núcleo Interinstitucional de Lingüística Computacional (NILC) do Brasil.

- o projecto da floresta sintá(c)tica http://cgi.portugues.mct.pt/treebank/

Este projecto tem como objectivo criar um conjunto de unidades correctamente analisadas sintacticamente (revistas por linguistas) que possam servir como instrumento de calibragem de analisadores sintácticos automáticos, assim como melhorar a cobertura descritiva da análise sintáctica computacional, obrigando a um consenso ou pelo menos à formalização das escolhas e alternativas tomadas em áreas "cinzentas" da sintaxe. Se é, por um lado, possível considerá-lo como a continuação natural do projecto AC/DC (por fornecer uma informação mais fina – árvores sintácticas – e com um nível de correcção consideravelmente superior – dada a revisão humana), o projecto **Floresta Sintá(c)tica** é, por outro lado, mais ambicioso, tanto em termos de concepção científica (a forma de criação do banco de árvores é original) como em termos políticos, visto que pretende congregar, após uma primeira fase de experimentação ligada a um analisador sintáctico determinado (o PALAVRAS de Eckhard Bick, Bick, 2000), todos os grupos que se dedicam ao processamento da nossa língua e que aceitem colocar o formato do seu analisador para escrutínio do resto da comunidade.

- o serviço Busca http://cgi.portugues.mct.pt/Busca/

De momento, o Busca resume-se a um sistema de procura sobre as nossas páginas, com algumas capacidades de fornecer informação sobre investigadores relacionados com o processamento da língua portuguesa, dadas as listas de doutorados e de projectos com financiamento público em Portugal a que temos acesso. Contudo, pretendemos vir a tornar este sistema progressivamente mais inteligente, por um lado detectando automaticamente possíveis candidatos para enriquecer o catálogo e, por outro, evoluindo para um sistema de procura de texto na Web dentro de uma determinada área temática (usando categorização semi-automática).

Referências

Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006). 1999. Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia, http://www.mct.pt/Livro-BrancoCT/.

Livro Verde para a Sociedade da Informação em Portugal. 1997. Missão para a Sociedade de Informação, http://www.missao-si.mct.pt/livroverde/livrofin.htm.

Bick, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

Frankenberg-Garcia, Ana & Diana Santos. 2001. "Introducing COMPARA: the Portuguese-English Parallel Corpus". In *Proceedings of The Second International Conference on Corpus Use and Learning to Translate (CULT2000)* (Scuola Superiore di Lingue Moderne per Interpreti e Traduttori

da Università di Bologna, Bertinoro, Italy, November 3-5, 2000), St.Jerome, 2001.

Rocha, Paulo Alexandre & Diana Santos. 2000. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa". In Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, Atibaia, São Paulo, Brasil (19 a 22 de Novembro de 2000), pp.131-140.

Santos, Diana. 1999a. "Processamento computacional da língua portuguesa: Documento de trabalho", versão base de 9 de Fevereiro de 1999; revista a 13 de Abril de 1999, http://www.portugues.mct.pt/branco/.

Santos, Diana. 1999b. "Disponibilização de corpora através da WWW", in Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações: Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998). APL, Lisboa: Colibri, pp.323-346.

Santos, Diana. 1999c. "Toward Language-specific Applications", *Machine Translation* **14** (2), June 1999, pp.83-112. Kluwer Academic Publishers.

Santos, Diana. 2000. "O projecto Processamento Computacional do Português: Balanço e perspectivas", *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, Atibaia, São Paulo, Brasil (19 a 22 de Novembro de 2000), pp.105-113.

Santos, Diana & Eckhard Bick. 2000. "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp.205-210.

Santos, Diana. 2001. "Introdução ao processamento de linguagem natural através das aplicações", in Elisabete Ranchhod (ed.), *Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações*, Lisboa: Caminho, pp.229-259.