

multi-palavra e um protótipo de uma ferramenta de tradução automática baseada em exemplos (*example-based machine translation*).

TrAva e CorTA: avaliação de tradução automática para português

O TrAva (Traduz e Avalia) foi desenvolvido no contexto de uma proposta exploratória de avaliação conjunta para a tradução automática (TA) feita pela Linguateca. Esta ferramenta, desenvolvida no pólo do Porto da Linguateca, permite traduzir frases do inglês para o português em quatro motores de TA disponíveis livremente na Internet, e pede aos utilizadores para classificarem as traduções obtidas, utilizando um quadro de classificação que recorre a dois sistemas gramaticais: para as frases em inglês, o sistema de anotação gramatical utilizado pelo British National Corpus⁴; para as frases em português, uma taxonomia baseada na sintaxe do português [30]. O METRA, o sistema meta-tradutor que envia a frase em inglês para quatro sistemas distintos na rede, e que também pode ser utilizado independentemente, recebe de momento 100 pedidos de tradução por dia.

As avaliações das traduções, efectuadas pelos utilizadores do TrAva, foram armazenadas num corpus especializado, o CorTA (Corpus de Traduções automáticas Avaliadas) [9], que permite pesquisar e consultar essas traduções.

CHAVE

A colecção CHAVE [15] é um dos resultados visíveis da participação da Linguateca na organização em 2004, do CLEF (*Cross-Language Evaluation Forum*, Forum de avaliação conjunta cruzada) [32].

Esta colecção, um recurso útil para investigadores trabalhando na área da recolha de informação (RI), contém os textos completos do jornal diário português PÚBLICO, de 1994 e 1995, bem como uma lista de cinquenta tópicos em português, compilados em cooperação com os restantes organizadores do CLEF; as avaliações (binárias) de cada tópico, ou seja, que documentos são acerca desse tópico; uma lista de 700 perguntas e respostas em português, compiladas em cooperação com os restantes organizadores do QA@CLEF; um conjunto não-exaustivo de documentos que suporta a(s) resposta(s) para 199 dessas perguntas. Está acessível gratuitamente dos nossos servidores, mediante um registo prévio [36].

WPT 03

A colecção WPT 03 [34] é a maior recolha de documentos da web portuguesa existente. Pode ser utilizada como um recurso importante para trabalhos de investigação em várias áreas do processamento da língua portuguesa, linguística e sociologia.

É o resultado de uma parceria entre a Linguateca e o XLDB⁵ que desenvolveu o motor de pesquisa para a Web portuguesa tumba! [31]. Recolhida entre Março e Junho de 2003, contém aproximadamente 3,5 milhões de documentos [5]. Em conjunto com a colecção é também disponibilizado o diário (log) com os registos das pesquisas efectuadas no tumba! ao longo de seis meses, diário esse com mais de um milhão de registos.

Esfinge

O Esfinge [27] é um sistema de resposta automática a perguntas de domínio geral em português que explora a redundância existente na rede, bem como o facto do português ser uma das linguagens mais utilizadas na mesma [38]. O Esfinge é baseado na arquitectura proposta por Brill para o inglês [24].

Este sistema, ainda incipiente, está disponível na rede, onde é possível colocar perguntas em português e obter as dez respostas mais prováveis encontradas pelo sistema.

Avaliação

Processo de dinamização

O primeiro passo tomado pela Linguateca na área da avaliação [13], foi estudar o que já tinha sido feito para as outras línguas nesta área. Dessa forma colheram-se ensinamentos preciosos de forma a evitar alguns dos erros cometidos no passado. Optou-se pela adopção do paradigma de "avaliação conjunta" (*evaluation contest*), segundo o qual os participantes na avaliação participam activamente na organização.

De seguida, partiu-se para a identificação de aplicações e recursos passíveis de avaliação. Criou-se um formulário na rede onde se pediu aos interessados, que indicassem as áreas que tinham mais interesse em avaliar. Este foi um passo essencial para definir quais as áreas em que havia mais potencial e interesse para organizar uma avaliação conjunta.

O passo seguinte foi a criação de uma lista de discussão, a "avalia", destinada à discussão de todos os assuntos relacionados com avaliação de sistemas e recursos relacionados com o processamento do português. O objectivo desta lista é propiciar a discussão sobre a organização de avaliações no futuro, bem como discutir assuntos científicos e técnicos à volta da avaliação de sub-áreas do processamento do português.

Morfolimpíadas

As Primeiras Morfolimpíadas para o português [8] [11], organizadas pela Linguateca, foram a primeira actividade

de avaliação conjunta realizada para o português, dedicada à análise morfológica. Decorreram de Março a Junho de 2003, culminando com a sessão final, no âmbito do encontro AVALON' 2003 - Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português (um encontro satélite do PROPOR' 2003 - VI Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada), organizado pela Linguatca no dia 28 de Junho de 2003 na Universidade do Algarve.

A nossa opção de começar por um exercício de avaliação na área dos analisadores morfológicos tem diversas justificações: a constatação, derivada das sondagens descritas na secção anterior, de que havia um considerável número de interessados nessa avaliação; o facto de os analisadores morfológicos serem um componente presente na grande maioria das aplicações que processam a língua portuguesa; finalmente, e atendendo ao facto de ser um acontecimento pioneiro e dada a relativa escassez de recursos materiais e humanos, a intuição de que seria a área mais realista para empreender a primeira avaliação.

O sistema PALMORF, desenvolvido por Eckhard Bick, obteve os melhores resultados, considerando os critérios previamente estabelecidos. No entanto, todos os participantes tiveram um desempenho comparável, com diferenças apenas na faixa dos dez pontos percentuais.

Os dados e resultados desta primeira avaliação conjunta para o português podem ser obtidos a partir do nosso portal, constituindo um recurso único para quem quiser estudar os desafios que a morfologia portuguesa ainda apresenta, assim como investigar diferentes métricas de avaliação.

Nessa data foi também lançada a ideia de um livro que apresentasse ao público as actividades de avaliação conjunta do processamento computacional da língua portuguesa que têm sido levadas a cabo ou inspiradas pela Linguatca. Embora o primeiro acontecimento deste tipo, as Morfolimpíadas, ocupe uma parte significativa do livro, muitas outras áreas são focadas, visto que o objectivo principal do livro não é o de relatar simplesmente uma experiência, mas sim servir de referência a este paradigma da engenharia da linguagem em português [16].

HAREM

Além do CLEF, já mencionado acima a propósito da colecção CHAVE, a Linguatca promove neste momento o HAREM - Avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas ("named entity recognition"). A chamada à participação ocorreu em Setembro de 2004, e contamos com que a avaliação dos sistemas participantes tenha lugar no final de Janeiro de 2005.

Escolhemos esta área para continuar o nosso esforço de avaliação em processamento da língua portuguesa por várias razões: por um lado já existe uma vasta história a nível internacional, e foi grande o interesse suscitado por uma experiência preliminar no princípio de 2003. Além disso, o HAREM permite avaliar outro tipo de capacidades das que foram objecto de estudo nas Morfolimpíadas e no CLEF: Por um lado, a tarefa é mais semântica, por outro, é menos complexa do que a resposta automática a perguntas ou a própria identificação do tópico de um documento.

O futuro

Pensamos que a Linguatca, nestes quase cinco anos de existência, mudou efectivamente as condições de trabalho de quem se dedica ao estudo da língua portuguesa e ao desenvolvimento de sistemas que a processam. Chegou a altura de, tendo erguido a infra-estrutura, nos podermos dedicar a questões mais do foro da investigação (aplicada). Além do modelo IRA, que continuaremos naturalmente a seguir, planeamos debruçar-nos sobre os seguintes problemas: categorização automática de texto em português; extracção inteligente de terminologia bilingue sobre corpora comparáveis; investigação de padrões de uso de serviços na rede; construção semiautomática de ontologias baseadas em linguagem natural; e desenvolvimento sistemático de estruturas de indexação e procura baseadas num trabalho terminológico de base.

Agradecimentos

O agradecimento é devido a todos os que colaboram e colaboraram com a Linguatca, visto que o trabalho aqui descrito é o resultado da conjugação dos seus esforços. Adicionalmente agradecemos a tradução do resumo do artigo para espanhol e francês a Paulo Rocha e Isabel Marcelino respectivamente. A Linguatca é financiada pela Fundação para a Ciência e Tecnologia (FCT) através do projecto POSI/PLP/43931/2001, co-financiada pelo POSI.

Diana Santos

Linguatca, pólo de Oslo, SINTEF ICT

Luís Costa

Linguatca, pólo de Oslo, SINTEF ICT

