Desafios do Processamento de Línguas Naturais

Vera Lúcia Strube de Lima[1], Maria das Graças Volpe Nunes[2], Renata Vieira[3]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC) – PUC-RS
Av. Ipiranga 6681 prédio 32 sala 647 – 90619-900 Porto Alegre – RS – Brasil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC) – ICMC-USP
Caixa Postal 668 – 13560-970 São Carlos – SP – Brasil

[3]Programa Interdisciplinar de Pós Graduação em Computação Aplicada – UNISINOS
Av. UNISINOS, 950 - 93022-000 São Leopoldo – RS – Brasil

vera.strube@pucrs.br, gracan@icmc.usp.br, renatav@unisinos.br

Abstract. *This paper presents the challenge of Natural Language Processing, in particular, the case of Portuguese language in the scope of Computer Science and its disciplines. Questions related to natural language processing are associated to the challenges of knowledge access, information management in data intensive repositories, and the complex and interdisciplinary problems of artificial, natural and socio-cultural systems modeling. The processing of Portuguese language constitutes a crucial demand for the participative and universal knowledge access for Brazilian citizens. Computer Science, in this scenario, plays a central role.*

Resumo. *Este artigo apresenta o desafio do Processamento de Línguas Naturais e, em especial, da Língua Portuguesa, no âmbito da Ciência da Computação e suas disciplinas. Questões relacionadas ao processamento da língua se associam aos desafios do acesso ao conhecimento, da gestão da informação em grandes volumes de dados e dos problemas complexos e interdisciplinares da modelagem computacional de sistemas artificiais, naturais e sócio-culturais. O processamento da língua portuguesa constitui demanda crucial para o acesso participativo e universal do cidadão brasileiro ao conhecimento. A Computação, nesse cenário, é chamada ao papel principal.*

1. O Cenário

As tecnologias da informação e da língua são, cada vez mais, elementos chave do ingresso na Sociedade da Informação e na Sociedade do Conhecimento. Por sua natureza de viabilização, especialmente na geração, no acesso a conteúdos e na constituição de redes de comunicação, essa temática perpassa a interatividade e contempla aspectos do multilingüismo e do acesso à informação, que promovem a cidadania, o desenvolvimento tecnológico e a redução do abismo digital. A importância dessas tecnologias se amplia quando passa a conviver com a realidade social, na busca da universalização do acesso à informação. Este é o caso, por exemplo, do atendimento

ao portador de necessidades especiais, seja ele o deficiente auditivo[1] ou portador de outros tipos de deficiência, ou da atenção ao idoso[2], o qual apresenta uma perda da mobilidade e das capacidades visual e auditiva. Estas questões permeiam não apenas as camadas menos favorecidas da população: dizem respeito a todas as situações sociais, e remetem a todas as etnias e culturas.

As demandas pelas chamadas tecnologias sociais e tecnologias assistivas, onde se inserem as tecnologias da informação e da língua, voltam-se a uma série de necessidades sociais entre as quais podemos destacar: impulsionar o acesso à Sociedade da Informação, através da redução de barreiras lingüísticas e culturais, promovendo a cidadania pelo acesso à informação; reduzir o abismo digital; aumentar a integração social e reduzir as barreiras de comunicação e de empregabilidade, e contribuir, entre outros, à integração e ao bem-estar de todos os cidadãos, e à inclusão dos idosos, descapacitados e deficientes.

De um ponto de vista sócio-econômico, a realidade hoje é de amplificação dos sistemas de acesso à informação, tais como ferramentas para prover acesso a conteúdos textuais e multimídia, governo eletrônico, monitoração e prospecção de mercados, etc. Como um país que necessita e busca desenvolver a formação de seus cidadãos em escala, por meio do ensino a distância, também necessitamos de ferramental com as interfaces adequadas para diálogos, através de tutores inteligentes e melhoria dos aspectos psicopedagógicos em sistemas tradicionais de ensino. Em todas essas direções, o processamento da língua é elemento chave.

Já na esfera econômica, necessitamos criar novas alternativas de produtos e serviços tecnológicos, gerando novas oportunidades e aumentando a competitividade, através das assim chamadas indústrias da língua. Necessitamos promover nas empresas a criação de produtos e serviços com valor agregado, para que proporcionem melhor atendimento às demandas de seus clientes.

Nesse cenário de necessidades e oportunidades é importante situar, desde já, o estado de desenvolvimento de nosso tema central. O Processamento Automático da Língua Natural, ou Processamento da Língua Natural (PLN), mostrou evolução significativa nas últimas décadas, entretanto ainda não proporciona a infra-estrutura exigida para oferecer o desejado suporte à Sociedade de Informação. O uso de um enfoque racionalista [Manning and Schütze 1999] entre as décadas de 1960 e 1980, caracterizou-se pela crença de que ‘uma parte significativa do conhecimento na mente humana não deriva das experiências proporcionadas pelos sentidos, mas é previamente estabelecida, presumivelmente por herança genética’. Esse argumento foi o adotado por Noam Chomsky, que difundiu a idéia da ‘faculdade inata da linguagem’. Na Inteligência Artificial, essa crença racionalista levou à construção de sistemas inteligentes com boa dose de conhecimentos codificados, e mecanismos de raciocínio associados, de modo a

[1] Conforme informação divulgada pela FENEIS – Federação Nacional de Educação e Integração dos Surdos, as escolas da rede pública brasileira possuem hoje 53.000 alunos com algum tipo de deficiência auditiva.

[2] Segundo dados da OMS, até 2025 o Brasil será o sexto país do mundo com o maior número de pessoas idosas.

promover o suposto modo de operação do cérebro humano. O enfoque empirista hoje em pauta atenua o valor depositado pelo primeiro nas habilidades cognitivas inatas: 'não é possível que a aprendizagem parta de uma *tabula rasa*, mas sim de operações gerais de associação, reconhecimento de padrões e generalização'. Este enfoque recupera hoje sua importância nas tarefas do processamento da língua, conduzindo à aplicação da estatística, reconhecimento de padrões e aprendizagem de máquina junto a grandes volumes de amostras da linguagem.

Este artigo tem por objetivo apresentar, no contexto dos Grandes Desafios da Pesquisa em Computação no Brasil, o desafio do Processamento da Língua Natural. No cenário internacional das pesquisas em computação, o processamento da língua natural faz parte da agenda de pesquisa das mais importantes universidades e centros de pesquisa. Entretanto, além do muito que ainda existe por ser feito, as pesquisas no processamento da língua remetem a uma forte componente de regionalização, apresentando maior densidade de resultados para algumas línguas do que para outras.

De contornos delineados, ao longo do tempo, por aspectos sensivelmente econômicos, os resultados destas pesquisas se voltam ainda timidamente à língua portuguesa. A maior densidade de produtos e resultados se concentra na língua inglesa, e estes resultados não são, na área do processamento da língua, diretamente transportáveis para outras línguas. São necessários estudos fundamentais que envolvem a construção de recursos de base, como léxicos, ontologias, grandes coleções de textos ou de amostras de fala processadas e anotadas – os corpora, gramáticas e analisadores, que somente se tornam disponíveis após um trabalho árduo de estudo, seja este estatístico-computacional ou lingüístico-cognitivo, do comportamento da linguagem e do seu uso. Somente dispondo dos resultados desta pesquisa básica, os recursos fundamentais, é que podemos prosseguir com a construção de ferramentas a serem disponibilizadas para as diferentes aplicações.

Em suas demandas por pesquisa básica e por pesquisa aplicada, o processamento da língua natural não constitui um foco autocontido de interesse. Os resultados dessas pesquisas são hoje de importância crucial a outras áreas do conhecimento, seja diretamente, seja como motor de aceleração. A Bioinformática, por exemplo, é área que se beneficia diretamente destes trabalhos, ao empregar métodos para união, mapeamento e construção de ontologias de termos, nas pesquisas *in silica*. O acesso à informação, dentro das perspectivas de Tim Berners-Lee e seu conceito de Web Semântica, pressupõe a representação do conhecimento semântico e pragmático em associação com a representação de grandes bases textuais ou multimídia. A Recuperação de Informação, nesse sentido, busca hoje a associação com recursos de processamento da língua, tais como as ontologias, para atingir um novo patamar de resultados.

O leitor encontrará o processamento da língua associado, aqui, aos desafios do acesso participativo e universal do cidadão brasileiro ao conhecimento, da gestão da informação em grandes volumes de dados multimídia distribuídos e dos problemas complexos e interdisciplinares da modelagem computacional de sistemas artificiais, naturais e sócio-culturais. Desafios estes que requerem soluções articuladas entre diferentes disciplinas para serem resolvidos, onde a língua natural e seu conseqüente processamento computacional desempenham um papel de destaque.

Este artigo discute os principais desafios atuais do Processamento da Língua Natural, e da Língua Portuguesa, em três seções, seguidas de considerações finais e de referências bibliográficas. A primeira seção, esta motivação inicial, cerca a problemática e o tema que abordamos. A segunda seção apresenta em maior detalhe os entraves tecnológicos da área de PLN, com destaque para o processamento do Português, no contexto do acesso à informação digital. A terceira seção aponta para algumas áreas correlatas e desafios de domínio conexo a esse tema de pesquisa.

2. Os desafios do processamento da língua natural e do Português como facilitador do acesso à informação digital

O cenário atual aponta para aplicações voltadas ao acesso à informação em diferentes níveis e formas. A disponibilidade cada vez maior de informação em diferentes formatos e mídias, aliada à popularização do computador como ferramenta de apoio às mais variadas tarefas, faz do acesso à informação digital, entre todas as possíveis operações numa interação usuário-computador, o principal alvo de atenção dos pesquisadores. Nesta seção focalizamos os tipos de aplicação de acesso à informação para cujas soluções a área de PLN tem especial relevância, e cujos desafios deverão receber a atenção dos pesquisadores nos próximos anos. Adicionalmente, comentamos o modo como se posiciona o processamento do Português brasileiro nesse cenário.

Sistemas de busca e recuperação de documentos ou informações a partir de padrões textuais são bastante populares atualmente, dada a imensa quantidade de informação veiculada na web. A tecnologia empregada varia de simples máquinas de estados finitos a sistemas mais sofisticados, que levam em conta algum tratamento lingüístico da consulta e também características do perfil do usuário. O enriquecimento desse processo pode-se dar pela consideração de informação lingüística e contextual, em ordem crescente de complexidade, nos níveis morfológico (a formação das palavras da consulta e do texto pode guiar, de alguma forma, a busca), morfossintático (as classes das palavras – nome, verbo, etc. – são indicadores de relevância), sintático (a inter-relação entre os elementos de um texto – nome-adjetivo, sujeito-verbo, verbo-objeto, etc. adiciona informação importante), semântico (em nível lexical – o sentido daquela ocorrência – ou em nível sentencial – o significado de uma sentença, ambos contribuem para o sucesso de uma busca) e pragmático-discursivo (pela consideração de questões culturais, por exemplo). Vale ressaltar que o estado-da-arte possibilita o uso dos primeiros níveis (morfológico e morfossintático), o sintático (razoavelmente acessível para as línguas mais favorecidas, como o inglês) e, mais recentemente, o semântico. Para o Português do Brasil, recursos para os primeiros níveis – lexical, morfossintático e sintático – têm sido produzidos em diversos contextos, com bons resultados [Nunes *et al.* 1996, Strube de Lima *et al.* 1997, Aires *et al.* 2000, Bick 2000, Santos 2005]. Em nível discursivo, é possível dispor de ferramentas de análise retórica monodocumento [Pardo and Nunes 2006], com boa precisão para o gênero científico, ou de análise de expressões referenciais e anaforicidade [Abreu and Vieira 2006, Coelho *et al.* 2006]. Em nível semântico, trabalhos como [Gamallo *et al.* 2005] voltam-se à extração de relações semânticas a partir de relações sintáticas em textos em Português, porém esta solução ainda não é diretamente aplicável para associar informação semântica aos dados textuais. Outros trabalhos voltam-se à aplicação de novas teorias lingüístico-

computacionais, como a do Léxico Gerativo de James Pustejovsky (1996) [Gonzalez and Strube de Lima 2004], a nossa língua, na busca de uma representação mais adequada.

Pesquisas têm mostrado que as medidas de precisão e cobertura para a recuperação de informação não crescem na mesma proporção do esforço empregado para que isso aconteça [Feng-Yang *et al.* 2004]. Técnicas lingüisticamente motivadas que empregam tecnologias acessíveis a línguas menos favorecidas (isto é, com menor disponibilidade de recursos lingüístico-computacionais), como o Português, apresentam melhora pouco significativa se comparada ao esforço necessário para empregá-las [Voorhees 1999, Gonzalez 2005, Arcoverde *et al.* 2006].

Por outro lado, a demanda por buscas mais especializadas cresce à revelia dos gargalos tecnológicos. Em aplicações mais sofisticadas, a busca por padrões mostra-se insuficiente, e a busca por conteúdo faz-se imprescindível. Busca baseada em conteúdo é aquela na qual a consulta apresentada pelo usuário não é usada como padrão para a busca, mas sim, para dar uma pista sobre o conteúdo do documento ou o tipo de informação que o mesmo procura. Assim, um usuário poderia solicitar um artigo de jornal com "opiniões contrastantes sobre a crise da aviação civil brasileira". As cadeias da consulta podem ser usadas para a busca, o que constitui os primeiros passos nessa direção [Gonzalez *et al.* 2006] mas certamente não são suficientes. É necessário algum tipo de entendimento dos textos candidatos. E aí reside um grande gargalo do PLN.

Recursos lingüísticos de base são criados para que seja possível dar um passo a mais no processamento da língua humana mas, comparados à riqueza desta língua, muito ainda há por ser feito. Exemplos desses recursos são as *wordnets*, os bancos sintáticos e proposicionais, por exemplo. *Wordnets* são bases de dados em que unidades lexicais (palavras e expressões), pertencentes às categorias dos substantivos, verbos, adjetivos e advérbios, organizam-se por meio de relações semânticas, como sinonímia, hiperonímia/hiponímia, holonímia/meronímia, etc. A *Wordnet* de Princeton, desenvolvida para o inglês americano [Fellbaum 1998], foi a pioneira e, a partir dela, têm sido geradas outras, para várias línguas, incluindo a do Português brasileiro, em elaboração e ainda indisponível [Dias-da-Silva *et al.* 2006]. Tais recursos possibilitam a inferência com base na semântica lexical, permitindo identificar, por exemplo, termos correlatos, relações entre palavras de um texto, etc. O futuro acesso à *wordnet* do Português brasileiro, aguardado há vários anos, certamente promoverá uma série de avanços e aplicações – hoje incipientes – que dependam de algum tipo de informação semântica.

Bancos sintáticos (*treebanks)* são constituídos por sentenças analisadas sintaticamente, e são úteis tanto para pesquisas lingüísticas quanto para o aprendizado automático de conhecimento sintático. Qualquer aplicação mais sofisticada de PLN requer processamento sintático. Construir uma gramática para essa tarefa tornou-se muito mais fácil com o uso de modelos estatísticos, que fazem uso de grandes bancos sintáticos. O maior exemplo de um banco sintático talvez seja o Penn TreeBank [Marcus *et al.* 1993], composto por notícias jornalísticas do Wall Street Journal. Com base nesse banco, analisadores sintáticos automáticos foram produzidos com precisão surpreendente (em torno de 92%) para a língua inglesa, possibilitando o uso de tais

informações para o processamento de textos. Para o Português brasileiro há a Floresta Sintática[3], com vários milhões de palavras, necessária para a construção de ferramentas básicas como etiquetadores PoS (*Part os Speech*) e *parsers*.

Em relação aos recursos necessários para o desenvolvimento dessa área, temos muito que avançar no que concerne à língua portuguesa. Para o desenvolvimento de sistemas, sua avaliação e aperfeiçoamento, é necessário um conjunto de dados lingüísticos iniciais, principalmente se considerarmos que o desenvolvimento da área tem-se dado por técnicas estatísticas e de aprendizado de máquina [Manning and Schütze, 1999]. Há ainda uma distância muito grande entre os recursos que possuímos para o Português e o que hoje existe disponível para o trabalho, estudo e desenvolvimento de tecnologia para a língua inglesa[4], por exemplo. Esses dados lingüísticos são de cara construção: para o desenvolvimento da área é necessário investimento básico. O *site* da Linguateca (de iniciativa portuguesa) reúne a maior parte do que há disponível hoje, em termos de ferramentas e corpora da língua portuguesa. A maior parte do corpus com frases anotadas em sua estrutura sintática, a Floresta Sintática, entretanto, corresponde à anotação automática não corrigida. A versão denominada Bosque corresponde à parte corrigida (que pode ser considerada como exemplo ou padrão) e possui em torno de 35.000 palavras em relação a um ideal de 1.000.000 de palavras para esse tipo de dado, sendo que esse é um dos primeiros níveis de anotação. Um dos primeiros corpora da língua Portuguesa, reunindo anotações de vários níveis, está em fase final de construção [Abreu *et al.* 2007] porém, devido à complexidade da tarefa, esse é um corpus de dimensão bastante reduzida em relação aos citados anteriormente.

Atualmente a área conta com dados lingüísticos bem mais sofisticados, de nível semântico e de discurso, onde são informados os papéis semânticos dos complementos verbais, e as relações retóricas entre os segmentos do discurso. Os bancos proposicionais são um exemplo, e constituem-se de grandes repositórios de informações sobre o comportamento das palavras de uma língua, ou seja, como elas se relacionam com outras palavras para constituir sentenças semanticamente bem formadas. O PropBank [Kingsbury and Palmer 2002], para o inglês, armazena as estruturas argumentais possíveis para uma grande quantidade de verbos. Tais recursos possibilitam a inferência do significado em nível sentencial. Outro exemplo de recurso semântico auxiliar são as ontologias, atualmente utilizadas em domínios bem definidos, e construídas (semi-) automaticamente a partir de grandes corpora [Staab and Studer 2004]. Somos ainda carentes desses recursos e suas tecnologias derivadas. Um dos primeiros projetos para a construção de um corpus para o processamento da língua Português do Brasil é o PLN-BR[5]. Esforços voltados a ontologias dão seus primeiros passos [Chaves and Strube de Lima 2004].

Nas aplicações em que a recuperação da informação é apenas uma parte da tarefa a ser executada, como nos sistemas de apoio à decisão, operações mais sofisticadas, como inferências lógico-semânticas, são desejáveis. No mesmo exemplo anterior

[3] www.linguateca.pt
[4] LDC (Linguistic Data Consortium) http://www.ldc.upenn.edu/
[5] http://www.nilc.icmc.usp.br:8180/portal/

("opiniões contrastantes sobre a crise da aviação civil brasileira"), concluir que as opiniões são contrastantes requer mais do que compreender sentenças: é necessário reconhecer o relacionamento retórico entre elas e seu papel no discurso como um todo. Sistemas de análise discursiva são apenas um exemplo de recurso para essa tarefa. A inferência completa em língua natural consiste, ainda, em um grande desafio a ser superado, exigindo ontologias e regras de dedução. Questões importantes, nesse mesmo tema, dizem respeito à organização da terminologia específica, os tesauros, e a definição clara de conceitos (as ontologias) que possam guiar a coleta e o armazenamento de dados, bem como a codificação da informação.

Extrair conhecimento de texto (não estruturado) e transformá-lo em conhecimento estruturado é o foco central da área conhecida com extração da informação. Geralmente uma tarefa de extração de informação está relacionada à busca de conhecimento sobre assunto específico, busca esta feita sobre coleções de textos. O conhecimento lingüístico e as técnicas do processamento da língua empregados nesses sistemas são bastante sofisticados: possuir esse conhecimento sobre nossa língua e ter disponíveis ferramentas e recursos para desenvolvê-las e aperfeiçoá-las são questões estratégicas para o país. Cumpre mencionar, aqui, que a área de extração de informação conta com considerável investimento por parte dos países desenvolvidos. Um exemplo de investimento para o avanço do conhecimento e tecnologias de processamento de texto é o programa ACE (*Automatic Content Extraction* - http://www.nist.gov/speech/tests/ace/), coordenado pelo NIST (Instituto Nacional de Padrões e Tecnologias) dos Estados Unidos. Este instituto possui uma divisão de acesso a informação, da qual o programa de tecnologias da linguagem humana é parte (http://www.itl.nist.gov/iad/).

Relacionada à extração de informação e de relevância em vários domínios, a área de sistemas de perguntas e respostas visa desenvolver sistemas com capacidade de interpretar uma pergunta e buscar em documentos a informação necessária, com capacidade de compor informações de diversas fontes, sintetizar e apresentar ao usuário, de maneira textual coerente, a resposta relevante para a pergunta realizada. Claro que nos interessa não apenas a disponibilidade de tais sistemas para o nosso uso, mas nos interessa, sobretudo, o conhecimento necessário para construí-los.

Tarefa relacionada à produção de perguntas e respostas é a produção de síntese do resultado produzido pela busca e recuperação de informação. Buscas simples retornam trechos do texto pesquisado ou índices de documentos completos da base considerada. No caso de grandes bases, buscas mais sofisticadas poderiam trazer resumos dos documentos selecionados, ou ainda um único resumo de todos eles (sumarização multidocumento). Além disso, a língua do resumo poderia ser distinta das línguas dos documentos recuperados (sumarização *cross-language*). Em sistemas como o Google[6], esses "resumos" são constituídos por segmentos de textos. Para se obter mais qualidade, é preciso lançar mão de técnicas de sumarização automática [Mani 2001]. Em vista da grande complexidade de se abordar essa tarefa do ponto de vista de compreensão do texto-fonte, seguida de síntese do resumo, as técnicas extrativas são as

[6] www.google.com.br

mais usadas e estudadas. Trata-se de técnicas baseadas na seleção e justaposição de segmentos relevantes do texto-fonte, com base em pouca ou nenhuma informação lingüística. O estado-da-arte para a sumarização monodocumento parece ter se aproximado de seu limite [Leite *et al.* 2007], enquanto que a sumarização multidocumento e a sumarização *cross language* são atualmente alvos de intensas pesquisas. A sumarização de fontes multimídia (resumo de fontes em diferentes mídias) se apresenta como um grande desafio para o futuro próximo. Para o Português, bons sistemas de sumarização extrativa monodocumento estão disponíveis [Pardo *et al.* 2003, Rino *et al.* 2004, Leite *et al.* 2007], mas a geração de sumários multidocumento ainda é intenção de pesquisa, e não se conhecem trabalhos para o tratamento multilingüe.

A web semântica, onde as informações estarão melhor estruturadas e organizadas com o intensivo uso de metadados lógicos baseados em ontologias, pode ser vista como uma resposta ao desafio da busca por conteúdo. Padrões de formatação de documentos, como XML e XCES[7] visam à etiquetagem semântica para posterior manipulação computacional. Enquanto as pesquisas avançam na direção de padrões que atendam a demanda atual e futura, a criação ou adequação dos documentos codificados para a web semântica ocorre em passos muito mais lentos. Isso se deve à necessidade de trabalho manual de codificação, já que a atribuição automática de etiquetas semânticas esbarra em questões lingüísticas ainda intratáveis. No Brasil, os padrões têm sido estudados e usados conforme direciona o estado-da-arte. No entanto, ainda não se percebe intercâmbio de documentos e tratamento semântico mais sofisticado do Português nesse cenário. A disponibilização de grandes corpora (Lacio-Web[8] e PLN-BR[9], no formato XCES) ou coleções classificadas (como Folha-RICol, disponível na Linguateca) representa um passo importante para futuras pesquisas e aplicações. Em paralelo a essa linha de desenvolvimento, há um grande esforço despendido na relação entre o conhecimento representado em textos e sua exploração para a construção de ontologias de domínios, necessárias à web semântica e também a diversas outras aplicações, veja por exemplo [Cimiano *et al*. 2005].

A aquisição de conhecimento, seja este lingüístico ou não, é um processo altamente oneroso e complicado. Visando automatizar esse processo, projetos de criação de bases semânticas de apelo público, como o Open Mind Common Sense [Push Singh *et al.* 2002] e o Verb Ocean [Chklovski and Pantel 2004], são alternativas a muito longo prazo. Qualquer usuário que tenha acesso à rede digital pode contribuir com tais projetos. Acredita-se que, um dia, munida de uma gigantesca quantidade de informações, a máquina seja capaz de raciocinar sobre essas informações, preencher as lacunas preenchidas até então por humanos e, quem sabe, gerar conhecimento novo. Técnicas de *Text Mining* ou mineração de textos estão dando os primeiros passos nessa direção. A versão em Português da base Open Mind Common Sense está em construção[10].

[7] http://www.cs.vassar.edu/XCES/
[8] http://www.nilc.icmc.usp.br/lacioweb/
[9] http://www.nilc.icmc.usp.br:8180/portal/
[10] www.sensocomum.ufscar.br

A web nos permite o contato com um imenso mundo multilingüe que, no entanto, fica limitado a uma pequena parte quando a língua representa uma barreira. Mesmo sendo o inglês a língua franca da Internet e a segunda língua de vários povos, não é desprezível o conteúdo da web em outras línguas, tampouco é pequeno o número de usuários não falantes do inglês. Nesse sentido, o acesso à informação multilingüe se apresenta como outro desafio. Fazer consultas em uma língua e receber como resposta documentos em várias línguas, ou ainda na mesma língua, mas tendo como fonte documentos multilíngües, descreve um cenário não muito distante de nossa realidade. A tecnologia envolvida requer necessariamente algum processo de tradução.

A tradução automática (TA) foi uma das primeiras aplicações-alvo da Computação, e até hoje permanece como um grande desafio. Tradutores robustos e livres de erro não existem nem mesmo para as línguas mais estudadas, como o inglês. Seus principais problemas transcendem questões tecnológicas e são inerentes a questões lingüísticas e culturais. A grande demanda atual por esses sistemas, no entanto, torna populares mesmo os sistemas de baixa qualidade. O caráter multilíngüe da internet tem exigido tradutores de pares de línguas nunca antes tratados pela academia. Como o desenvolvimento de um sistema tradicional de TA (envolvendo a construção de gramáticas e dicionários bilíngües) é altamente custoso, uma das duas alternativas deve ser seguida: (a) o uso da técnica de tradução baseada em uma interlíngua [Dorr *et al.* 2000], reduzindo de $O(2^n)$ para $O(n)$ o esforço de desenvolvimento de tradutores entre n diferentes línguas; ou (b) o uso de técnicas estatísticas guiadas por grandes corpora de textos bilíngües automaticamente alinhados, e por modelos probabilísticos de realização de línguas [Och *et al.* 2004]. A primeira alternativa tem sido adotada e abandonada alternadamente ao longo do tempo, dada a alta complexidade de se criar uma interlíngua capaz de representar o significado comum a todas as línguas naturais, interlíngua esta que, teoricamente, para ser capaz da façanha anterior, acaba sendo tão complicada e de difícil tratamento quanto qualquer língua natural. Já a segunda alternativa tem sido a opção da atualidade. Tradutores entre línguas distantes, como inglês e coreano, inglês e chinês, desenvolvidos por esse método[11], já estão disponíveis no Google. O custo real da produção desses tradutores é a formação do corpus de textos alinhados, com várias centenas de milhões de palavras, cujo processamento demanda muito tempo, mas é feito uma única vez. Sua limitação é a mesma de qualquer sistema baseado em corpus: os resultados sempre serão dependentes das características do corpus (tamanho, gênero, domínio, etc.). Desenvolvimentos rápidos dessa natureza podem representar a chave para o salto quantitativo necessário para colocar o Português em posição de igualdade com outras línguas. No entanto, as primeiras iniciativas de tradução estatística com o Português estão ainda em nível de projeto de pesquisa [Pardo 2006]. Para a tradução envolvendo línguas de naturezas diferentes, como línguas visuais-gestuais (LIBRAS, por exemplo), com grande impacto para a inclusão digital, os desafios são ainda maiores, e projetos dessa natureza já se fazem presentes na comunidade científica, como por exemplo SignNet[12] e FALIBRAS [Tavares *et al.* 2006].

[11] http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html
[12] http://sign-net.ucpel.tche.br/

A grande questão além da aquisição e criação de recursos lingüísticos sofisticados reside no uso de tais recursos, quer via construção de bases de dados para fins específicos, quer via apelo público, como o projeto Open Mind Common Sense. Sabe-se muito pouco sobre como o cérebro humano processa a língua ou como, mesmo artificialmente, pode-se se beneficiar dos recursos produzidos. Por exemplo, na tradução automática, sabe-se que a sintaxe tem um papel importante, pois reordenações dos elementos sentenciais são essenciais para a geração de uma boa tradução, mas o alcance da sintaxe é limitado por nossa incapacidade de enxergar como utilizá-la em um nível além do básico [Knight and Marcu 2005]. Sabe-se que textos em línguas diferentes são organizados de forma diferentes, pois os falantes têm preferências diversas de realização lingüística baseadas em suas histórias e culturas. Há muitos mecanismos para representar a organização textual, alguns considerados clássicos, como a *Rethorical Structure Theory* [Mann and Thompson 1987] e a *Grosz and Sidner Discourse Theory* (GSDT) [Grosz and Sidner 1986], mas se sabe muito pouco sobre como utilizá-los para, por exemplo, reorganizar textos traduzidos para que eles soem mais naturais para os nativos da língua em questão. Muito pouco foi feito nesse sentido (veja, por exemplo, [Marcu *et al.* 2002]), dada a complexidade da tarefa. Ainda se está muito longe do poder de inferência necessário para se lidar com a língua devidamente, mesmo que todos os recursos lingüísticos estejam à mão.

A disponibilidade, hoje tão difundida, de fontes multimídia, coloca texto e fala lado a lado. No entanto, o processamento da fala tem sido tratado em comunidade separada da comunidade de PLN, tradicionalmente ocupada com a língua escrita. Estabelecida na área de Processamento de Sinais, na Engenharia Elétrica, a área de Processamento de Fala (Análise e Síntese) compartilha com o PLN a interdisciplinaridade com a Lingüística. O resultado é que alguns recursos lingüístico-computacionais são comuns e, portanto, o desenvolvimento de uma área favorece a outra.

Texto e fala aparecem juntos nos mais variados cenários, mas optamos por citar um, que envolve vários dos desafios que terão de ser enfrentados no futuro próximo: o da computação ubíqua. Nesse cenário, várias questões presentes em cada um dos 5 Grandes Desafios da SBC[13] são colocadas e, entre elas, o tratamento computacional da língua falada e escrita figura entre as mais relevantes. Tratar a língua nesse cenário envolve encarar, se não todos, vários dos problemas ainda em aberto aqui trazidos. É importante que pesquisas multidisciplinares sejam incentivadas no sentido de acelerar o desenvolvimento de tecnologias que representem a solução para os problemas apontados, com especial foco na língua portuguesa.

3. Áreas correlatas e desafios de domínio conexo

Recentemente o processamento da língua natural tem se relacionado com outras áreas de importância científica e tecnológica para o desenvolvimento da sociedade. Exemplos podem ser citados entre as ciências da vida e da saúde - Biologia, Bioinformática e Medicina - e na área de infra-estrutura tecnológica, como aquela necessária, por

[13] http://www.sbc.org.br/index.php?language=1&subject=8&content=downloads&id=231

exemplo, para o governo eletrônico, sendo freqüente o inter-relacionamento entre o processamento de textos e a definição de conceitos principais e terminologia específicas das áreas (por meio de representação de conhecimento na forma de ontologias, como por exemplo em [Silva *et al*. 2006]). Motivos da presença do processamento da língua natural nessas áreas são, além da grande quantidade de informação armazenada de forma textual, o conhecimento altamente especializado e a necessidade de busca ou reaproveitamento desse conhecimento. Com esses exemplos, podemos ressaltar também a importância do processamento da língua para outras áreas estratégicas, como a Educação. As áreas de Informática na Educação e Educação a Distância, e o desenvolvimento de programas sofisticados que facilitam o aprendizado, são temas de pesquisa em evolução. Sendo a língua natural o meio mais usado para a transmissão de conhecimento, é fácil deduzir que o domínio da tecnologia para o processamento da língua é também relevante nesse cenário.

Uma das áreas onde o aparato do processamento da língua tem sido aplicado recentemente é a Biologia Molecular, entre outros domínios de biociência computacional [Hirschman *et al.* 2005]. A análise semântica profunda de documentos possibilita a extração de informação de domínio especializado, em uma área rica em conhecimento expresso em língua natural. Outros domínios relacionados à saúde também estão interessados em fazer uso de técnicas do processamento da língua para obter informações armazenadas de forma não estruturada. Por exemplo, podemos citar o desafio internacional lançado para extração de informação de textos contendo dados clínicos[14]. Além da capacidade de processamento e desenvolvimento dessas áreas através de dados organizados e anotados (tal como os dados biológicos são indispensáveis para as áreas relacionadas à Biologia), existe ainda outra forma de buscar conhecimento lingüístico-computacional, esta através do processamento de vastas coleções de textos puros (sem anotação). Nesse tipo de trabalho a modelagem matemática estatística é fundamental, e a capacidade de se tratar grandes volumes de dados também se faz necessária. Considerando a necessidade do tratamento de bases lingüísticas extensas, com ou sem anotação, e a importância do processamento da língua para o processamento da informação em geral, é que o processamento da língua natural também tem sido enquadrado como *Data Science*, juntamente com áreas como Bioinformática e Visão Computacional[15]. Nesse espectro se situa o processamento da língua como um problema a ser tratado juntamente com o que se conhece como "dados multimídia complexos".

Estudos de análise de texto também têm sido empregados em análises do mercado financeiro [Koppel and Shtrimberg 2004]. Uma nova área de pesquisa relacionada ao processamento do discurso tem sido a análise de sentimentos, emoções e subjetividade em textos. Alguns trabalhos procuram relacionar o impacto no mercado financeiro de notícias divulgadas na imprensa.

E não seria justo deixar de destacar aqui a própria Lingüística, cujos avanços mais recentes se valem, em boa parte, dos trabalhos em computação com a análise de

[14] http://www.computationalmedicine.org/challenge/index.php
[15] Data Sciences Summer Institute http://mias.uiuc.edu/mias/summer_institute, Data Sciences Journal http://dsj.codataweb.org/

grandes volumes de texto, e é aliada permanente na construção de soluções para o processamento da língua.

Pode-se ainda citar duas iniciativas correlatas importantes: projetos[16] aprovados pelo programa Institutos do Milênio. Na Lingüística, o projeto de construção do Dicionário Histórico do Português do Brasil (Séculos XVI, XVII e XVIII) e na Matemática um projeto de caráter interdisciplinar de promoção da Matemática e seu relacionamento com diversas áreas, entre elas a modelagem da língua natural.

4. Considerações finais

Ao longo de várias décadas, desde o surgimento dos primeiros computadores, a compreensão de línguas naturais tem representado um grande desafio. Mesmo em recortes para domínios e cenários muito particulares, a compreensão ainda é um grande empecilho. Acostumados a encarar a compreensão como uma tarefa inatingível, os pesquisadores têm proposto soluções paliativas que, ao final, mostraram-se valiosos avanços em diferentes cenários.

Este artigo abordou a problemática do Processamento da Língua Natural e do Processamento da Língua Portuguesa no cenário da Ciência da Computação. Trouxemos uma motivação inicial para o assunto – as questões sociais e sócio-econômicas no acesso participativo ao conhecimento. Situamos os problemas em discussão na atualidade e os desafios de pesquisa que se apresentam em aberto, sempre os remetendo ao cenário do processamento de nossa língua, o Português, e associando com as referências de maior relevância na área. Finalizando, traçamos uma correspondência com outras áreas de domínio conexo a esse tema de pesquisa, deixando saliente a importância que os avanços no processamento da língua representam para estas áreas. Reconhecida a importância e o valor da informação e do conhecimento para o desenvolvimento da sociedade, não se pode negligenciar a capacidade de processar o conhecimento disponibilizado em língua natural. O tratamento computacional da informação em língua natural, a modelagem matemática da língua e a aquisição do conhecimento lingüístico podem representar inestimável auxílio em termos do acesso universal à informação, e são desafios não apenas da área de PLN, mas de toda a Ciência da Computação.

Referências

Abreu, S. C. *et al.* (2007) Summit: um corpus anotado com informações discursivas visando sumarização automática. In: V TIL, 2007, Rio de Janeiro. Congresso da SBC, 2007 (a ser publicado).

Abreu, S. C.; Vieira, R. (2006) Learning Portuguese Discourse-new References. In: IFIP 19th World Computer Congress, TC-12 IFIP AI 2006 Stream. Berlin: Springer, 2006. v. 217. p. 267-276.

[16] http://www.cnpq.br/programasespeciais/milenio/projetos/2005/26.htm e http://www.cnpq.br/programasespeciais/milenio/projetos/2005/15.htm.

Aluísio, S.M. *et al.* (2003) The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: Proceedings of Corpus Linguistics, Vol. 16, pp.14-21.

Arcoverde, J.M.A.; Nunes, M.G.V.; Scardua, W. (2006) Using noun phrases for local analysis in automatic query expansion. Cross Language Evaluation Forum – CLEF 2006. Alicante, ES. Taller Digital, pp.1-4.

Bick, E. (2000). The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aarhus: Aarhus University Press.

Chaves, M.S.; Strube de Lima, V.L. (2004) Looking for Similarity between Ontological Structures. In: Branco, A.; Mendes, A.; Ribeiro, R. (Org.). Language Technology for Portuguese: shallow processing tools and resources. Lisboa: Edições. Colibri, v.1 pp.1-14.

Chklovski, T. and Pantel, P. (2004) VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona.

Cimiano, P., Hotho, A., Staab, S. (2005) Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research (JAIR) 24: 305-339.

Coelho J. C. B. *et al.* (2006) Resolving Nominal Anaphora. In: PROPOR - Workshop for Processing of Portuguese Language, Itatiaia. Lecture Notes in Artificial Intelligence 3960. Berlin: SPRINGER. pp. 160-169.

Dias-da-Silva, B.C.; Di Felippo, A.; Hasegawa, R. (2006) Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations. In: PROPOR - Workshop for Processing of Portuguese Language, Itatiaia. Lecture Notes in Artificial Intelligence 3960. Berlin: SPRINGER. pp.120-130.

Dorr, B.J.; Jordan, P.W.; Benoit, J.W. (2000). A Survey of Current Paradigms in Machine Translation. In: M. Zelkowitz (Ed.) Advances in Computers, Vol.49, pp.1-68. Academic Press, London.

Fellbaum, C. (Ed.) (1998) Wordnet: an electronic lexical database. Cambridge, MIT Press, 1998.

Feng-Yang Kuo *et al.* (2004) An investigation of effort-accuracy trade-off and the impact of self-efficacy on Web searching behaviors. Decision Support Systems, v.37 n.3, pp.331-342.

Gamallo, P. *et al.* (2005) Using Syntax-based methods for extracting semantic information. Linguistica Computazionale, Pisa-Roma, v.XXII, n.IV, pp.201-229.

Gonzalez, M.A.I.; Strube de Lima, V.L. (2004) Redefinig traditional lexical semantic relations with Qualia information. Palavra (PUCRJ), Rio de Janeiro - RJ, v.12, pp.25-36.

Gonzalez, M.A.I. (2005) Termos e relacionamentos em evidência na recuperação de informação. Tese de Doutoramento. UFRGS.

Gonzalez, M.A.I.; Strube de Lima, V.L.; Lima, J.V. (2006) Tools for nominalization: an alternative for lexical normalization. In: PROPOR - Workshop for Processing of Portuguese Language, Itatiaia. Lecture Notes in Artificial Intelligence 3960. Berlin: SPRINGER. pp.100-109.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. Computational Linguistics, Vol. 12, N. 3.

Hirschman, L. *et al.* (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005, 6 (Suppl 1):S1 doi: 10.1186/1471-2105-6-S1-S1. http://www.biomedcentral.com/1471-2105/6/s1/s1.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In: Proceedings of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas.

Knight, K. and Marcu, D. (2005). Machine Translation in Year 2004. In: Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp.18-23. Philadelphia, PA.

Koppel, M., Shtrimberg, I. (2004) Good News or Bad News? Let the Market Decide. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text. Palo Alto: AAAI Press. pp. 86-88.

Leite, D.S. et al. (2007) Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In: Proceedings of the Workshop on TextGraphs-2 Graph-Based Algorithms for Natural Language Processing (associado ao HLT/NAACL 2007), Rochester, USA. v. 1. pp. 17-24.

Mani, I. (2001). Automatic Summarization. John Benjamins Pub. Co. Amsterdam.

Mann, W.C., Thompson, S.A. (1987). Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190.

Manning, C. and Schütze, H. (1999) Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press.

Marcu, D., Carlson, L. and Watanabe, M. (2000). The Automatic Translation of Discourse Structures. In: Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2000), Seattle, Washington.

Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, Vol.19, N. 2, pp.313-330.

Martins, R. T.; Hasegawa, R.; Nunes, M.G.V. (2003) Curupira: a functional parser for Brazilian Portuguese. In: PROPOR - International Workshop on the Computational Processing of Portuguese, Faro. Lecture Notes in Computer Science 2721 Berlin: SPRINGER.

Nunes, M.G.V. *et al.* (1996) (In Portuguese) Development of a parser for Brazilian Portuguese. In: Proceedings of the 2nd Workshop on Computational Processing of Written and Spoken Portuguese. Curitiba: CEFET-PR. pp.71-80.

Och, F.J. *et al.* (2004). A Smorgasbord of Features for Statistical Machine Translation. In the Proceedings of HLT/NAACL.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In: PROPOR - International Workshop on the Computational Processing of Portuguese, Faro. Lecture Notes in Computer Science 2721. Berlin: SPRINGER. pp. 210-218.

Pardo, T.A.S.; Nunes, M.G.V. (2006). DiZer – an Automatic Discourse Analyzer for Brazilian Portuguese. In: Proceedings of the V Best MSc Dissertation/PhD Thesis Contest – CTDIA. Ribeirão Preto-SP, Brazil.

Pustejovsky, J. (1996) The Generative Lexicon, MIT Press. Cambridge, MA.

Rino, L.H.M. *et al.* (2004). A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA. Lecture Notes in Artificial Intelligence 3171. Berlin: SPRINGER. pp.235-244.

Santos, C.N. (2005) Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro. Dissertação de Mestrado. IME-RJ.

Silva, J. P. M. *et al.* (2006) Exploring molecular networks using MONETontology. Genetics and Molecular Research, v. 5, n. 1, pp. 182-192.

Singh, P. *et al.* (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In: Proceedings of the First Int. Conf. on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems. Lecture Notes in Computer Science. Heidelberg: Springer-Verlag.

Staab, S., Studer, R. (Eds.) (2004) Handbook on Ontologies. International Handbooks on Information Systems, Springer Verlag.

Strube de Lima, V., Abrahão, P.R.C., Paraboni, I. (1997) Approaching the dictionary in the implementation of a natural language processing system: toward a distributed structure. International Informatics Series 8. Baeza-Yates, R. (Ed.) Fourth South American Workshop on String Processing, WSP'97, Valparaíso – Chile. Carleton University Press, Ottawa – Canada.

Tavares, O.L. *et al.* (2006) O Sistema Falibras-MT como Ferramenta de Apoio Pedagógico. In: Anais do IV Congresso Ibero-Americano Sobre Tecnologias de Apoio a Portadores de Deficiência, Vitória. v. II. pp. CO-109-CO-112.

Voorhees, E. (1999) Natural Language Processing and Information Retrieval. In: Pazienza, M.T. (Ed.) Information Extraction: Towards Scalable, Adaptable Systems, Lecture Notes in Artificial Intelligence 1714. Berlin: SPRINGER.

Vossen, P. (2004) Ontologies. In: Mitkov, R. (Ed.) The Oxford handbook of Computational Linguistics. Oxford: Oxford University Press. pp.464-82.