

DIRPI - um processo interdisciplinar e interinstitucional

Ana Luísa Leal¹ e Paulo Quaresma²

Departamento de Informática – Universidade de Évora

analu@di.uevora.pt , pq@di.uevora.pt

Resumo: Este artigo apresenta o Projecto de Cooperação Interinstitucional Interdisciplinar – DIRPI³, na área de linguística computacional. O projecto DIRPI, que incorpora outras investigações, é constituído pela interacção de grupos de pesquisas de instituições de países de Língua Portuguesa, são elas Universidade de Évora – Portugal e as Universidades Unisinos e PUCRS no Brasil. Além disso, neste trabalho, são apresentados os objectivos, os recursos e resultados, bem como, a importância dos projectos de cooperação entre diferentes instituições de pesquisa.

Introdução

No universo académico, o desenvolvimento de uma investigação científica é imprescindível, no sentido que é a pesquisa e os resultados obtidos pela sua execução que movimentam e fazem evoluir conceitos, dados e teorias apresentados em momentos específicos para um fim específico. É fato que a pesquisa científica se concretiza através da interacção de diferentes elementos, os quais não são excludentes em todo seu processo, são eles: o corpo docente e o discente, o grupo institucional, que acolhe pesquisadores e fornece material logístico para a evolução do trabalho; e as agências e órgãos de fomento e agências financiadoras, que disponibilizam recursos financeiros para que os projectos possam ser realizados.

¹ Ana Luísa V. Leal: bolsista de doutoramento CAPES – 2113/03-01 Universidade de Évora. Orientanda do Dr. Paulo Miguel Duarte Quaresma (Univ. Évora) e Co-orientanda da Dr^a Rove Chishman (Unisinos).

² Paulo Quaresma: Professor Associado do Departamento de Informática da Universidade de Évora – Portugal. Coordenador do Departamento de Informática da

³ DIRPI – Desenvolvimento e Integração de Recursos para Pesquisa de Informação. Projecto de Cooperação Internacional Brasil / Portugal. ICCTI/CAPES. (UE -Universidade de Évora e Unisinos - Universidade do Vale do Rio dos Sinos). Período 2002 – 2006.

Assim sendo, os projectos de cooperação entre instituições de nível superior, estejam elas em um mesmo país ou em países diferentes, representam uma possibilidade efectiva e significativa para a evolução das investigações de áreas afins. É a partir do intercâmbio entre instituições autónomas que se torna possível compartilhar informações técnicas e experiências diferenciadas entre grupos de pesquisas, os quais possuem particularidades e especificidades próprias adequadas à realidade social da qual fazem parte. A possibilidade de trocar experiências académicas e sociais, discutir e reestruturar resultados obtidos, bem como, reconhecer o manancial técnico particular a cada grupo de pesquisa é o que impulsiona a criação e a manutenção de projectos de cooperação interinstitucionais.

Os projectos de cooperação além de firmarem relações entre instituições diferenciadas, possibilitam agregar estudos interdisciplinares, ou seja, investigações que coadunam diferentes áreas. O DIRPI – Desenvolvimento e Integração de Recursos para Pesquisa de Informação – é um exemplo de projecto de cooperação interinstitucional, o qual abarca um projecto de doutoramento interdisciplinar, em que está sendo desenvolvido um estudo que relaciona as áreas Informática e Linguística, trata-se do projecto AuTema-Dis, apresentado sinteticamente neste artigo.

O principal objectivo desse trabalho é apresentar, a partir dos resultados obtidos pelo projecto DIRPI, a importância dos projectos de cooperação interinstitucionais e interdisciplinares, as conclusões obtidas a partir da interacção entre instituições e diferentes áreas, bem como, a necessidade de receber apoio de agências e órgãos de fomento, para a realização, evolução e continuidade nos processos de investigação académica.

O Projecto DIRPI

O projecto DIRPI – Desenvolvimento e Integração de Recursos para Pesquisa de Informação – tem como objectivo desenvolver ferramentas e recursos para o tratamento computacional da Língua Portuguesa, uma pesquisa na área do Processamento de Linguagem Natural (PLN). Como mencionamos, trata-se de um projecto de Cooperação Internacional Brasil/Portugal – CAPES⁴ – ICCTI em que pesquisadores brasileiros e portugueses buscam, através de seus estudos, consolidar e integrar os resultados obtidos nas áreas de Inteligência Artificial e Processamento da Linguagem Natural.

Elementos integrantes da parte Brasileira (instituição coordenadora):

Unisinos – Universidade Do Vale do Rio dos Sinos – PPGLA – Centro de Ciências da Comunicação e PIPCA – Centro de Ciências Exactas e Tecnológicas), com os projectos: Projecto DIXIT; Projecto ANACORT (Unisinos e Pucrs); Projecto MLXML (Unisinos); APSCO (Unisinos); Projecto COMMON-REFs (Unisinos, Pucrs/ProTeM-CC-Inria), coordenados pela prof^a. Dr^a. Renata Vieira;

PUCRS – Pontificia Universidade Católica do Rio Grande do Sul, (instituição participante), coordenada pela prof^a. Dr^a Vera Lúcia S. de Lima.

Elementos integrantes da parte Portuguesa:

DI-EU – Departamento de Informática da Universidade de Évora;

CENTRIA – Centro de Inteligência Artificial, com os projectos: Projecto ABC – Acesso Inteligente a Bases de Conhecimento; Projecto Cooperativo e o Projecto PGR, coordenados pelo Prof. Dr. Paulo Quaresma – UE e UNL (Universidade Nova de Lisboa).

⁴ CAPES: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – www.capes.gov.br

No âmbito do projecto DIRPI, os temas seleccionados para desenvolvimento de pesquisas são:

- Construção e uso de etiquetadores para textos em Língua Portuguesa: trabalho desenvolvido com etiquetadores morfossintáticos, visando à construção de desambigüadores de etiquetas de carácter sintático;
- Construção de analisadores sintáticos: trabalho desenvolvido para analisar sintacticamente (análise parcial) texto não-restrito. Diagnóstico de correcção de falhas, que permite criar agentes de recuperação, buscando melhor resultado;
- Construção de sistemas para resolução de co-referência: trabalho desenvolvido para o tratamento de co-referências de descrições definidas em textos da língua portuguesa; construção de esquemas padronizados para a anotação de informação linguística e semântica em corpus;
- Gestão de Diálogos para a Interação com Sistemas de Pesquisa de Informação: projecto desenvolvido, visando ambiente de programação lógica para modelagem de agentes computacionais com a capacidade de agir de maneira activa, cooperativa e inteligente em diálogos;
- Sistemas de Pesquisa de informação em textos da língua portuguesa: projecto desenvolvido a partir do resultado do projecto PGR5, que permitiu a construção de um sistema de pesquisa de informação para textos em Língua Portuguesa; tal sistema permite pesquisas em diversas formas de palavras; extracção automática de informações relevantes existentes no texto; investigação semântica de páginas da WWW e busca de informações;
- Recuperação de Informações através do léxico gerativo, ferramentas de etiquetagem, reconhecimento e classificação de SN, extracção automática de relações semânticas em corpora, construção e manipulação de Thesaurus, entre outras;
- Construção de Corpora Anotados: projecto desenvolvido para realizar anotação morfológica, anotação sintáctica parcial e anotação semântica em

⁵ PGR : Projecto PGR – Acesso Selectivo aos Pareceres da Procuradoria-Geral da República.

três tipos distintos de corpora; alinhamento de corpus paralelo para extracção, construção de dicionários bilingues.

O projecto de DIRPI tem dado suporte à comunhão de projectos e áreas afins, assim sendo, destacam-se trabalhos de investigação que estabelecem relações entre diferentes áreas. Da necessidade em se desenvolver as pesquisas supracitadas e considerando a interacção entre instituições universitárias, foram promovidos dois doutoramentos, os quais estão sendo realizados por estudantes brasileiros na Universidade de Évora e recebem orientação no trabalho por parte do Dr. Paulo Quaresma (Univ. Évora) e co-orientação por parte das Professoras Dr^a. Rove Chishman e Dr^a. Renata Viera (Univ. Unisinos). Os projectos individuais de doutoramento recebem apoio das agências de fomento CAPES e ALBAN.

Projecto *AuTema-Dis* – Automatização Temática do Discurso

O Projecto de doutoramento – “Da automatização da análise sintáctica à temática discursiva: Um estudo de relações retóricas em textos” – *AuTema-Dis* – está sendo desenvolvido pela doutoranda Ana Luísa Leal na Universidade de Évora, em Évora – Portugal e é orientado pelo Prof. Dr. Paulo Miguel Duarte Quaresma. O doutoramento em questão insere-se no âmbito do projecto DIRPI.

O projecto *AuTema-Dis* prevê uma intersecção de conhecimentos de áreas distintas, ou seja, Linguística e Informática. Trata-se da construção de uma arquitectura computacional que realiza a análise textual considerando as informações mais relevantes dispostas na superfície de um texto. O objectivo do trabalho é desenvolver um programa de computador que reconheça a informação principal em um determinado texto, considerando o resultado da análise sintáctica e a organização das estruturas e, por consequência, produzir automaticamente uma estrutura sintética em língua natural, por exemplo o Português, do que é apresentado no texto completo.

O AuTema-Dis, descrito de forma técnica, realiza automaticamente a análise sintáctica em um texto através do analisador sintáctico *Palavras*; o resultado dessa análise nos fornece as informações necessárias para elaborarmos regras para justificar a identificação dos segmentos, delimitarmos os seus limites e a sua estrutura. A estrutura do texto, considerando as características da automatização sintáctica, é apresentada em um sistema arbóreo, o que chamamos de DTS – árvore de dependência dos segmentos. Acreditamos que a estrutura em árvores – DTS pode auxiliar na elaboração da macroestrutura de um texto, pois as árvores representam hierarquicamente os segmentos, que representam concretamente as proposições. Assim sendo, o objectivo do autor pode ser reconhecido através da análise da macroestrutura oferecida por essas árvores.

Em relação à pesquisa em desenvolvimento no projecto AuTema-Dis, concluímos tratar-se de uma proposta extensiva relacionada a diferentes áreas de conhecimento, bem como, diferentes níveis cognitivos. Nós acreditamos ser possível desenvolver um sistema robusto que é capaz de articular essas áreas e esses níveis. Algumas das etapas propostas já se encontram concluídas e apresentam resultados satisfatórios. Alguns dos resultados obtidos fornecem-nos suporte necessário para dar continuidade ao estudo. Para finalizar, acreditamos fortemente que é possível usar a estrutura sintáctica de um texto para obter os segmentos e as suas proposições correspondentes, gerar árvores DTS, inferir as estruturas retóricas e obter a macroestrutura do texto.

O Projecto – MAQA – Uma abordagem Multi-Agentes para perguntas e respostas

O Projecto de doutoramento MAQA – está sendo desenvolvido pela doutoranda Cássia Trojahn dos Santos⁶ na Universidade de Évora, em Évora – Portugal e é orientado pelo Prof. Dr. Paulo Miguel Duarte Quaresma. O doutoramento em questão insere-se no âmbito do projecto DIRPI.

Na proposta, uma abordagem multiagente para um sistema de perguntas e respostas, para a língua portuguesa é apresentada. Esta proposta é composta por três módulos: (1) processamento do documento e da pergunta; (2) construção de ontologias; (3) geração da resposta. Cada módulo é composto por múltiplos agentes cooperativos que adoptam estratégias distintas para gerar seus resultados e cooperam para criar um resultado global. Esta abordagem permite usar diferentes estratégias e reduzir os erros introduzidos pelo uso individual dos métodos. A cooperação entre os agentes objectiva obter melhores soluções em cada etapa do processamento.

A Figura 1 apresenta a arquitectura multi-agente proposta. O primeiro módulo, Processamento de documentos e pergunta, é responsável pelo processamento dos documentos e das perguntas. Os agentes deste módulo actuam em duas etapas. Primeiro, eles extraem a informação dos documentos e criam uma base de conhecimento. Segundo, os agentes processam a pergunta e criam a estrutura semântica das sentenças. Dois tipos de agentes compõem o primeiro módulo: agente de análise sintáctica e agente de análise semântica e pragmática. O agente de análise sintáctica é responsável pelo processamento das sentenças dos documentos e das perguntas, gerando a estrutura sintáctica destas sentenças (i.e., árvore sintáctica, representada em Prolog). O agente de análise semântica e pragmática transforma a saída do agente de análise sintáctica em outra colecção, onde cada documento ou pergunta tem uma representação semântica (i.e., estrutura de representação do discurso, DRS

⁶ Cássia Trojahn dos Santos: Bolsista de doutoramento Alban: E05D059374BR, Universidade de Évora. Orientanda do Dr. Paulo Miguel Duarte Quaresma (Univ. Évora) e Co-orientanda da Dr^a Renata Vieira (Unisinos).

[12]). A ontologia do domínio é considerada neste processamento. Na fase de processamento dos documentos, uma base de conhecimento contendo as instâncias da ontologia é construída, como resultado. Estes agentes adoptam diferentes estratégias (i.e., métodos linguísticos e de aprendizagem) para gerar os seus resultados e cooperam para criar um resultado final. Esta abordagem também é adoptada no segundo módulo da arquitectura.

No segundo módulo, Construção de ontologia, a ontologia do domínio é construída. Múltiplos agentes processam os documentos e cooperam para criar uma ontologia que representa o domínio. Cada agente é responsável por aplicar uma estratégia específica. Após, os resultados individuais são combinados em uma ontologia global.

Finalmente, o terceiro módulo, Geração da resposta, a resposta é formada: um conjunto de palavras e a identificação do documento onde a resposta foi encontrada. Este módulo é composto por um agente de geração de respostas, responsável por interpretar a pergunta na base de conhecimento através da unificação das entidades do discurso da pergunta com as entidades do discurso dos documentos.

Figura 1

Discussões e Resultados

A realização do Projecto DIRPI contou com financiamento da agência financiadora CAPES, possibilitando, desta forma, o desenvolvimento das metas descritas na proposta inicial, referidas neste artigo, bem como, a capacitação dos pesquisadores para realizá-las de maneira adequada aos requisitos descritos como objectivos.

Os componentes dos dois grupos brasileiros e português trabalharam de forma cooperativa através de reuniões expositivas, auxiliando na orientação e participação em bancas de trabalhos de conclusão de curso e dissertações de mestrados, em realização de tutoriais e publicações, assim como, participação conjunta em projecto de cooperação internacional, nomeadamente o Projecto Fundações Brasil-Portugal CAPES-ICCTI ao qual esta proposta veio dar continuidade.

Os vários projectos no âmbito do Projecto DIRPI foram desenvolvidos e concluídos em conformidade com as metas de formação descritas e as acções para atingi-las. No entanto, alguns dos estudos, como é o caso das duas propostas de doutoramento referenciadas neste artigo, encontram-se em desenvolvimento. Neste sentido, pretende-se submeter uma proposta de continuação no âmbito do Cytéd⁷.

Os grupos de envolvidos na cooperação, isto é, Brasil e Portugal, puderam se beneficiar da troca de experiências, conhecimentos e sistemas desenvolvidos por variadas equipas. As visitas de intercâmbio tornaram possíveis:

– A realização de cursos em áreas específicas relacionadas aos projectos, tais como, resolução de anáfora pronominal e resolução de descrições definidas na

⁷ CYTED: Programa Iberoamericano de Ciência y tecnología para el Desarrollo. www.cytéd.org

língua portuguesa, métodos estatísticos em processamento de linguagem natural, recuperação de informação e gestão de diálogos para recuperação de informação;

– Apresentação dos resultados obtidos pelas investigações em desenvolvimento: Compartilhamento e formação em ferramentas e recursos, que possibilitem a construção de corpora anotados (morfologicamente, sintacticamente e semanticamente) da Língua Portuguesa.

Em termos mais globais, obteve-se a ampliação da colaboração existente actualmente entre as diversas entidades participantes nesta proposta, permitindo o reforço do número de investigadores dos dois países envolvidos nestas áreas de investigação. Em termos específicos, apresentamos abaixo o quadro descritivo em termos numéricos das actividades realizadas:

QUADRO DEMONSTRATIVO DO PROJECTO DIRPI – ANO 2002/2006 –

ATIVIDADE	ANO 2003/2004	ANO 2004/ 2005	ANO 2005/2006	TOTAIS
MISSÕES DE TRABALHO	07 (total de missões) 05 – em Portugal 02 – no Brasil	04 (total de missões) 02 – em Portugal 02 – no Brasil	06 (total de missões) 04 – em Portugal 02 – no Brasil	17 Missões
FERRAMENTAS E SOFTWARES	01 Software 02 Ferramentas			03 (Ferramentas e Softwares)
PARTICIPAÇÕES EM BANCAS	01 – Defesa no Brasil – Univ. Pucrs	01 – Defesa no Brasil – Univ. Unisinos	01 – Defesa no Brasil – Univ. Pucrs	03 Defesas
DISSERTAÇÕES DE MESTRADO (em andamento)	04 – Brasil (Unisinos e Pucrs) 02 – Portugal (Univ. Évora)	01 – Brasil (Univ. Unisinos) 01 – Portugal (Univ. Évora)	04 – Brasil (Univ. Unisinos) 02 – Portugal (Univ. Évora)	14 Dissertações desenvolvidas
DISSERTAÇÕES DE DOUTORAMENTO (em andamento)	01 – Univ. Évora	03 – Univ. Évora	04 – Univ. Évora	04 Dissertações de Doutoramento em andamento
DOUTORAMENTOS SOLICITADOS	02 – Univ. Évora	03 – Univ. Évora	01 – Univ. Évora	06 Doutoramentos solicitados 02 Doutoramentos obtidos com financiamento CAPES ⁸ e ALBAN ⁹
PUBLICAÇÕES REL.ÀS MISSÕES DE TRABALHO	22 – Publicações	17 – Publicações	17 – Publicações	56 – Publicações

⁸ CAPES: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil.

⁹ ALBAN: Programa Alban de bolsas de estudo de alto nível destinado à América Latina.

Considerações Finais e Cooperações futuras

O projecto DIRPI de cooperação entre UNISINOS/PUCRS (Brasil) e Universidade de Évora (Portugal) na área de Linguística Computacional promoveu o desenvolvimento e a integração de recursos para pesquisa de informação, através de trabalho cooperativo entre equipas interinstitucionais. No momento, o projecto encontra-se em fase final, em que estão sendo concluídas as actividades de formação de recursos humanos e consolidados os resultados em termos de publicações científicas em eventos.

De acordo com os resultados obtidos ao final do projecto DIRPI, vislumbra-se a possibilidade de novas cooperações, bem como, a conclusão de pesquisas que ainda se encontram em andamento. Neste sentido, apontar para uma proposta de continuação no âmbito do Cyted – Programa Iberoamericano de Ciência y Tecnología para el Desarrollo.

Bibliografia

Leal, Ana Luísa Varani. AuTema-Dis – Da automatização sintáctica à temática discursiva: Um estudo de relações retóricas em textos. Universidade de Évora. Tese de doutoramento. 2006. (em andamento).

Santos, Cássia Trojahn dos. Abordagem Multiagente para Sistemas de Perguntas e Respostas. Universidade de Évora. Tese de doutoramento, 2006 (em andamento).

Vieira, Renata. Relatório Técnico 1 do Projecto DIRPI. Unisinos. Porto Alegre, 2003/2004.

Vieira, Renata. Relatório Técnico 2 do Projecto DIRPI. Unisinos. Porto Alegre, 2005/2006.

Vieira, Renata e Quaresma Paulo. Projecto de DIRPI. Unisinos e Universidade de Évora. Porto Alegre, 2002.