INTEGRACIÓN Y CONOCIMIENTO
ISSN 2347 - 0658

N° 5
Vol. 2 Año 2016

PROJETO RITA: COOPERAÇÃO, PROCESSAMENTO DE LINGUAGEM E LINGUÍSTICA

Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul, Brasil
mfinatto@terra.com.br

Laura Alonso Alemany
Universidad Nacional de Córdoba, Argentina
alemany@famaf.unc.edu.ar

PRESENTACIÓN

Este artículo presenta los objectivos y algunos resultados alcanzados en el marco del proyecto "RITA - RIch Text Analysis through Enhanced Tools based on Lexical Resources". Este proyecto fue seleccionado en la Convocatoria de 2013 del Programa STIC-Amsud tendo involucrado hasta Diciembre 2016 instituciones Superiores de Pesquisa y Formación Docente de Universidades de Brasil, Argentina, Uruguay y Francia. El programa STIC-AmSud financia movilidad de investigadores al interior de proyectos de investigación reuniendo, al menos, un equipo de investigación francés y dos equipos de investigación de dos países latinoamericanos (Argentina, Brasil, Chile, Colombia, Ecuador, Paraguay, Perú, Uruguay y Venezuela) bajo el domínio investigativo de las Ciencias y Tecnologías de Informação Y Comunicação (STIC). Nuestro principal objectivo com el Projecto RITA fue la creación y fomento de una red de cooperación en pesquisa entre expertos informáticos y linguistas de essas instituciones que desarollan las temáticas de Processamiento de la Linguagem Natural, *Information Retrieval* o representación de contenidos en espanõl e portugués por medio de *corpora* disponibles en línea.

ORIGEM E OBJETIVOS DO PROJETO RITA

Este relato apresenta os objetivos propostos e algus dos resultados alcançados no âmbito do projecto "RITA - *RIch Text Analysis through Enhanced Tools based on Lexical Resources*". Este projeto foi selecionado na Chamada Internacional de 2013 do Programa STIC-Amsud, tendo envolvido, no Brasil, por parte da CAPES, até 31/12/2016, pesquisadores da Universidade Federal

do Rio Grande do Sul (UFRGS), com sede em Porto Alegre, e da Universidade Federal de São Carlos (UFSCar), cuja sede se localiza na cidade de São Carlos, interior do Estado de São Paulo. Também estiveram oficialmente envolvidos, até 31/07/2016, no Projeto RITA, pesquisadores de Universidades da Argentina – *Universidad Nacional de Córdoba*, Uruguai – UDELAR - *Universidad de la Republica*, e da França, *Université Paris Nanterre* e U*niversité Paris-Est Marne-la-Vallée*. Cada grupo universitário esteve sob os auspícios de seus respectivos órgãos de fomento nacionais, com orçamentos diferenciados conforme as suas disponibilidades financeiras. Importa salientar que a coordenação geral internacional do projeto esteve a cargo da Profa. Dra. Laura Alemany, responsável pela equipe argentina.

STIC-Amsud é um programa internacional regular de cooperação científico-tecnológica no qual participam a França, Argentina, Brasil, Chile, Colombia, Equador, Paraguai, Peru, Uruguai e Venezuela. O objetivo desse programa é, justamente, gerar e fortalecer as capacidades regionais da América do Sul – salientando-se, naturalmente, o âmbito do MERCOSUL - e sua cooperação com a França. Sua proposta é estabelecer redes em matéria de pesquisa e desenvolvimento no âmbito das Ciências e Tecnologias da Informação e da Comunicação. Para tanto, o programa gera convocatórias a partir das quais se selecionam e apóiam projetos de pesquisa e de desenvolvimento, com potencial de transferência de inovação em nível regional. Esses projetos, com duração máxima de dois anos, para candidatarem-se a financiamentos internacional e nacional, devem envolver pelo menos dois países da região da América do Sul, bem como uma ou várias equipes de cientistas franceses.

ESCOPO DE TRABALHO E PESQUISA

O principal objetivo do projeto RITA foi criar um quadro para integrar os recursos e capacidades de cada grupo nacional, especialmente no âmbito da pesquisa em Processamento da Linguagem Natural (PLN). O PLN, no Brasil, também é conhecido por *Linguística Computacional* ou *Processamento de Língua Natural* (para mais detalhes sobre a área no Brasil, vale consultar, por exemplo, Pardo *et al.* 2010).

PLN, dito *grosso modo*, conforme Evers, Finatto (2016), pode ser entendido como uma subárea da Inteligência Artificial, ramo da Ciência da Computação, que reúne métodos formais para analisar textos – normalmente escritos, gerar frases escritas em uma língua natural e também descrever ou sistematizar conteúdos expressos em textos ou em acervos textuais. O objetivo final do PLN, nesse sentido, pode ser pensado como o de capacitar computadores para que possam "entender" e "redigir" textos em uma língua natural. Nesse "entender", estão as capacidades de, automaticamente, reconhecer um contexto de significação, fazer a análise sintática, semântica, léxica e morfológica de frases em textos, criar resumos, extrair informação, interpretar sentidos e até "aprender" noções ou significados de palavras ou de expressões fazendo uso de padrões depreendidos de textos processados.

Para exemplificar algo que a pesquisa PLN gerou e que usamos cotidianamente, podemos citar os sistemas de tradução automática ou mesmo os corretores ortográficos que usamos nos

nossos computadores ou telefones celulares. Qualquer pessoa que use uma ferramenta de processamento de texto, perceberá que há ali, embutido, um "programa" que destaca desvios ortográficos e gramaticais e até propõe correções – algumas equivocadas, mas outras muito adequadas. Os primeiros corretores ortográficos, criados no âmbito do PLN, funcionavam pela comparação simples de uma lista de palavras extraídas do texto que se digitava com uma lista de palavras (dicionário de palavras do programa) corretamente grafadas. Essa era e ainda é uma tarefa bem simples, que não demandava processamento complexo.

Essas ferramentas de PLN – como os corretores, hoje, tornaram-se muito mais sofisticadas e são capazes de detectar desvios relacionados não só à ortografia, como à morfologia (formação de plurais) e à sintaxe (ausência de um verbo ou falta de concordância), apontar problemas de pontuação e até sugerir palavras que podem ser mais adequados ao tipo de texto que se está produzindo (acadêmico, jornalístico, entre outros).

Pois bem, no âmbito de pesquisas em PLN, de diferentes graus de complexidade, conforme tentamos situar o nosso leitor, buscamos um cenário em que fosse posível integrar as *expertises* e os interesses de cada grupo universitário de pesquisa, salientando-se a cooperação entre especialistas de Computação e especialistas em Linguística. A Linguística, também dito de um modo muito simples, é uma ciência que se ocupa do funcionamento das línguas humanas, sendo estudada principalmente em cursos de Letras no Brasil. Seu mais importante precursor é Ferdinand Saussure, um estudioso genebrino, cujas lições nos legaram, em 1916, uma obra fundamental da Linguística, intitulada *Curso de Linguística Geral*. A Linguística Moderna hoje é constitída por uma série de variadas concepções e teorias sobre o que sejam as línguas e a faculdade da linguagem humana, das mais estruralistas às mais cognitivistas ou às centradas nos grupos sociais de falantes.

Reunindo especialistas de PLN e de Linguística de diferentes países e núcleos acadêmicos, nosso objetivo comum era alcançar um reforço ao nível dos tratamentos computacionais no nível sintático-semântico do Espanhol – com destaque para o espanhol sul-americano - e do Português do Brasil, considerando-se também experiências pré-existentes para o tratamento do Francês e ou para o contraste dessas línguas. Esse foco específico deu-se porque a história das pesquisas em PLN tem produzido muitos recursos e técnicas para línguas como Inglês, e bem menos para outras línguas. Desse modo, os objetivos secundários do projeto RITA foram promover a pesquisa justamente com línguas menos exploradas, como as línguas do MERCOSUL. Nesse sentido, buscamos:

a. explorar e desenvolver ferramentas para análise sintático-semântica integrada do Espanhol e do Português;
b. explorar e comparar de métodos para criar e enriquecer analisadores sintático-semântico usando métodos de aprendizado de máquina (*machine learning*) supervisionado e sem supervisão;
c. desenvolver de métodos para integrar léxicos enriquecidos em diferentes aspectos no processo de análise sintático-semântica do Espanhol e do Português;
d. desenvolver de métodos genéricos para extrair expressões multipalavra (*multiword*

expressions) típicas e recorrentes de determinados textos e obter a sua integração na análise sintático-semântica;

e. desenvolver de métodos genéricos para identificação dos argumentos verbais e seus papéis semânticos e temáticos (isto é, identificar "sujeitos" e "objetos" que recorrentemente acompanham os verbos de um texto), integrar essa identificação à análise sintático-semântico.

ALGUNS RESULTADOS

Como se pode imaginar, em termos de desafios, nossos objetivos foram ambiciosos, de modo que, para dar conta deles, no tempo de duração projeto, buscamos aproveitar ao máximo o que cada equipe do Projeto RITA já tivesse produzido, além de buscar experiências e resultados obtidos em diferentes centros do mundo. Nesse sentido, em 2014, já organizamos um primeiro *workshop* para compartilhamento de experiências. Esse *workshop* ocorreu durante uma dos mais importantes eventos de PLN relacionados ao processamento do Português, o PROPOR (*The International Conference on Computational Processing of Portuguese*), que ocorreu em São Carlos - SP, de 6 a 9 de outubro de 2014, tendo sido promovido justamente também por colegas do projeto RITA pesquisadores da UFSCar. Nosso *workshop* foi um evento conexo ao PROPOR e foi denominado **ToRPorEsp** - *Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish.*Um interesse especial foi o de facilitar o acesso a tecnologias e recursos que são específicos para Português e Espanhol.

Listamos, a seguir, os títulos dos trabalhos apresentados no nosso **ToRPorEsp**, salientando que a língua oficial das apresentações no PROPOR é o Inglês (a íntegra dos trabalhos pode ser conferida em <https://sites.google.com/site/torporesp/program/proceedings>) e que recebemos trabalhos de pesquisadores de Portugal, Noruega, Uruguai, Argentina e Brasil, que também apresentaram seus trabalhos no PROPOR:

- ✓ *The CINTIL and LX companion collections of language resources and tools for Portuguese*. António Branco, João Silva, Francisco Costa, Sara Silveira, Patricia Gonçalves and João Rodrigues
- ✓ *Desarrollo de un parser HPSG estadístico para el español.* Luis Chiruzzo and Dina Wonsever
- ✓ *Improving the Verb Lexicon of OpenWN-PT*. Valeria De Paiva, Alexandre Rademaker, Claudia Freitas and Livy Real
- ✓ *Enriquecendo o Córpus CSTNews – a Criação de Novos Sumários Multidocumento*. Márcio Dias, Thiago Pardo, Maria Lucia Castro Jorge, Alessandro Garay, Carla Chuman, Cláudia Barros, Erick Mazieiro, Fernando Nóbrega, Jackson Souza, Marco Cabezudo, Marina Delege, Naira Silva, Paula Cardoso, Pedro Balage, Roque Lopes, Vanessa Marcasso, Ariani Felippo and Maria Nunes
- ✓ *Lexical Resources for the Identification of Causative Relations in Portuguese Texts*. Brett Drury, Paula Cardoso, Janie Thomas and Alneu de Andrade Lopes.

- ✓ *Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT*. Hugo Gonçalo Oliveira
- ✓ *Filling the gap: inserting an artificial constituent where a subject is omitted in Portuguese*. Nathan Hartmann, Magali Duran and Sandra Aluísio
- ✓ *Extração de paráfrases em português a partir de léxicos bilíngues: um estudo de caso*. Paulo César Polastri, Helena De Medeiros Caseli and Eloize Rossi Marques Seno
- ✓ *Extending NomLex-PT using AnCora-Nom*. Livy Real, Valeria De Paiva and Alexandre Rademaker
- ✓ *O Tratamento de Marcadores Discursivos em uma Ferramenta de Apoio à Escrita Acadêmica em Português Para Nativos de Espanhol*. Lianet Sepúlveda Torres, Magali Sanches Duran and Sandra Maria Aluísio
- ✓ *Nos bastidores da Gramateca: uma série de serviços*. Alberto Simões and Diana Santos
- ✓ *O Corpus CSTNews e sua Complementaridade Temporal*. Jackson Souza and Ariani Di Felippo
- ✓ *Aprendizado de Máquina Sem-Fim para Indução Automática de Léxico Bilíngue*. Thiago Vieira and Helena Caseli
- ✓ *Towards a Phonetic Brazilian Portuguese Spell Checker*. Lucas Vinicius Avanço, Magali Sanches Duran and Maria Das Graças Volpe Nunes
- ✓ *Building a Corpus for Named Entity Recognition using Portuguese Wikipedia and DBpedia*. Cristofer Weber and Renata Vieira
- ✓ *Uma análise do perfil de entropia das estruturas sintáticas do português*. Marcely Zanon Boito, Luiza Hagemann, Rodrigo Wilkens and Aline Villavicencio

Depois do **ToRPorEsp**, realizamos uma série de missões de trabalho – entre docentes pesquisadores e missões de estudos entre pesquisadores pós-graduandos. Na parte brasileira, pela CAPES, pudemos contar com uma bolsa de doutorado-sanduíche de 12 meses, entre 2014 e 2015, para que uma doutoranda em Linguística da UFRGS pudesse realizar estudos junto ao *Laboratoire d'Informatique Gaspard-Monge, Université Paris-Est Marne-la-Vallée* relacionados a sistemas para tratamento do português do Brasil. Além desse apoio importante, em 2016, pudemos contar com o mesmo tipo de bolsa para uma doutoranda, também em Linguística, da UFSCar, realizar estudos contrastivos Espanhol-Português junto à Universidad Nacional de Córdoba, Argentina.

Também realizamos, em Porto Alegre, pela UFRGS, o **Segundo *Workshop* do Projeto RITA**, nos dias 17 e 18 de março de 2016, tendo contado com os apoios financeiros suplementares do Programa de Pós-Graduação em Letras e apoio de infraestrutura do Programa de Pós-Graduação em Computação da UFRGS. Nesse encontro, cada um dos centros universitários nacionais do Projeto RITA se fez representar, seja ao vivo, seja via recursos de videoconferência. A temática do encontro girou em torno das experiências em torno do assunto "Processamento do português e do espanhol, bases para um dicionário de homógrafos do português do Brasil e do espanhol rio-platense" e foram convidados estudantes de Linguística e de PLN interessados no tema.

Nesse encontro, também contamos com a participação de uma doutoranda em Linguística da Universidade de Federal de Santa Catarina (UFSC), que finalizava seu trabalho sobre aspectos da tradução de obras argentinas e uruguaias para o português do Brasil a partir de técnicas de contraste estatístico de padrões de correspondências tradutórias. Por sua vez, equipe do Uruguai do Projeto RITA, coordenada pela Profa. Dra. Ailá Rosá, da UDELAR, também nos brindou com a apresentação de seus trabalhos relacionados ao tema da construção de um dicionário computacional de homógrafos Espanhol-Português. Além disso, tivemos uma apresentação de trabalho realizado conjuntamente entre uma doutoranda de Linguística da UFRGS e um doutorando em Ciência da Computação do *Laboratoire d'Informatique Gaspard-Monge* em torno da detecção automática de erros de escrita em português produzidos por aprendizes do Português do Brasil que fossem hispanofalantes.

Ao final do **Segundo *Workshop* do Projeto RITA**, em uma reunião de encerramento, foram delineadas propostas de novos trabalhos em conjunto a serem apresentados em eventos e submetidos a publicações até dezembro de 2016.

PERSPECTIVAS

O projeto RITA se encerra oficialmente no Brasil em 31/12/2016, com paoio da CAPES, com uma série de resultados positivos, especialmente em termos de trabalhos conjuntos e do conhecimento e reconhecimento mútuo de diferentes pesquisadores e de suas conexões, especialmente os do âmbito do MERCOSUL. Com certeza, os núcleos do Brasil, Argentina e Uruguai puderam aprender muito uns com os outros e vivenciar diferentes realidades da pesquisa em PLN em parceria com as pesquisa em Linguística no âmbito latino-americano e europeu. Todavia, foi preciso que soubéssemos aprender a lidar, criativamente, com reduções repentinas de verbas em algumas equipes, prazos desiguais e, especialmente, com interesses focais diferentes da pesquisa de cada universidade e de cada país ou região. Ainda assim, acreditamos que todas as equipes envolvidas, da França, do Brasil, da Argentina e do Uruguai, puderam reforçar seus laços em torno do progresso para o tratamento computacional do Espanhol e do Português.

REFERÊNCIAS

▪ Pardo, T., Gasperin, C., Caseli, H. y Nunes. M. (2010). Computational Linguistics in Brazil: An Overview. In the Proceedings of the *NAACL-HLT Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pp. 1-7. June 1-6, Los Angeles, CA/USA. Disponível em: http://www.aclweb.org/anthology/W10-1601

▪ Evers, A. Finatto, M. (2016). Linguística de Corpus, Léxico-Estatística Textual e Processamento de Linguagem Natural: perspectivas para estudos de vocabulário em produções textuais. *Revista GTLEX*, *1*(2), pp. 271-295. Disponível em: http://www.seer.ufu.br/index.php/GTLex

Recibido: 01 de octubre de 2016 - **Aceptado**: 27 de octubre de 2016