a folha

Boletim da língua portuguesa nas instituições europeias

http://ec.europa.eu/translation/portuguese/magazine

Número especial

W-PROPOR2016 — Tomar, 13 de julho de 2016

Corpora e ferramentas para o processamento de *corpora*

A União Europeia, pela sua própria natureza, tem um vasto acervo de publicações multilingues, cobrindo essencialmente a legislação, a jurisprudência e a administração europeia. A maior parte destes recursos está disponível para o público e uma boa parte está também disponível em formato bi- ou multilingue. Este acervo é há muito explorado na administração europeia através de memórias de tradução — memórias interinstitucionais Euramis — e de há uns poucos anos para cá também através da tradução automática, através do serviço MT@EC. Este serviço de tradução automática está atualmente disponível não só para todas as instituições e demais órgãos da União Europeia mas também para as administrações nacionais dos Estados-Membros e de alguns países do Espaço Económico Europeu. Ora, em todos os Estados-Membros da UE e demais países participantes existem muitos recursos linguísticos ainda inexplorados, pelo menos para a tradução automática. Acrescentados aos *corpora* paralelos da UE, esses recursos podem contribuir para melhorar consideravelmente a qualidade geral da tradução automática e ajudar nas tarefas de tradução em todas as administrações públicas participantes e nos serviços administrativos em linha para o público em toda a União Europeia.

O Mecanismo Interligar a Europa (conhecido pela sigla inglesa CEF — Connecting Europe Facility) financia projetos destinados a inserir os elos em falta na cadeia da energia, dos transportes e das infraestruturas digitais. A tradução automática (CEF.AT, *CEF Automated Translation*) é um dos seus elementos constitutivos e terá por base tecnologias de tradução automática essencialmente estatísticas, que «aprendem» a traduzir a partir de traduções humanas. Com efeito, para efetuar essa aprendizagem, são necessários *corpora*, em especial *corpora* paralelos, em quantidade, qualidade e diversidade nos temas que se pretende abranger. A compilação e a disponibilidade desses recursos proporcionará também oportunidades de investigação de ferramentas novas ou melhoradas, de preferência disponíveis para toda a sociedade. A fim de organizar a colheita desses recursos, a Coordenação de Recursos Linguísticos Europeus (conhecida pela sigla inglesa ELRC — European Language Resource Coordination) organizou uma série de seminários em todos os países participantes no Mecanismo Interligar a Europa. Seguir-se-ão projetos concretos de identificação e recolha de recursos.

A Direção-Geral da Tradução (DGT) da Comissão Europeia desenvolveu e gere o serviço de tradução MT@EC, um sistema de base estatística que usa o ecossistema de código aberto Moses, ele próprio fruto em grande medida da investigação financiada pela União Europeia. Será o MT@EC que vai inicialmente assegurar os serviços do CEF.AT. O Departamento de Língua Portuguesa (DLP) da DGT desempenhou um papel precursor na adoção da tradução automática estatística pela DGT, devido a

três dos seus tradutores terem criado um pacote de instalação do *Moses — Moses for Mere Mortals* — com o qual se testou e demonstrou a capacidade desta tecnologia. Por isso está muito empenhado em que a iniciativa de recolha e utilização de novos recursos a nível nacional seja coroada de êxito. Quanto mais e melhores recursos existirem para a tradução e a língua portuguesa, mais e melhores serviços poderá o DLP prestar, mais e melhores serviços poderão prestar os tradutores que usam a língua portuguesa em geral, e mais se contribuirá para uma maior internacionalização e projeção da língua portuguesa.

O presente *workshop* está integrado na PROPOR2016, a 12.ª edição da Conferência Internacional para o Processamento da Língua Portuguesa. Para além dos objetivos mais práticos descritos acima, esta é também uma forma de prosseguir o objetivo mais vasto de alargar o espaço de interação entre a DGT e o seu Departamento de Língua Portuguesa com a comunidade universitária lusófona e a administração pública portuguesa nas áreas da tradução e da língua portuguesa.

Seguindo as regras da PROPOR2016, os artigos selecionados para o presente *workshop* são publicados em inglês no formato exigido para a conferência. Decidiu-se, no entanto, traduzir os resumos e as palavras-chave dos artigos para português, tendo sido pedido aos respetivos autores que revissem as traduções e apresentassem as correções que entendessem por bem fazer. Para além da presente publicação n'«a folha», os artigos serão igualmente disponibilizados no repositório em linha da Universidade de Lisboa.

Resta-nos agradecer a todos os que tornaram possível a realização deste *workshop* e esperamos que o tal espaço de interação em que se insere saia reforçado para benefício da língua portuguesa dentro e fora das fronteiras lusófonas.

Os organizadores,

António Branco
Universidade de Lisboa

Hilário Leal Fontes
Direção-Geral da Tradução — Comissão Europeia

1. Recursos multilingues

CM2News: Construção de um *corpus* de resumos multilingues multidocumentos

Ariani Di-Felippo
Núcleo Interinstitucional de Linguística Computacional (NILC), Universidade de São Paulo em São Carlos
Departamento de Letras, Universidade de São Carlos

Resumo: O presente artigo descreve a construção em curso do CM2News, um *corpus* semanticamente anotado destinado à investigação sobre resumos multilingues multidocumentos. O *corpus* é composto por 20 grupos de notícias em inglês e em português do Brasil e por um conjunto de resumos multidocumentos manuais e automáticos. Todos os textos de partida têm uma anotação semântica ao nível lexical. Alguns grupos de notícias têm também anotação ao nível da frase, bem como alinhamentos de textos e resumos manuais. O *corpus* é um dos resultados obtidos no contexto do projeto Sustento, o qual pretende gerar conhecimento linguístico para o resumo multidocumentos. No presente artigo descreve-se em pormenor a conceção do *corpus* e as tarefas de anotação manual.

Palavras-chave: *corpus*, recursos multilingues, resumo multidocumentos.

(ler artigo completo em inglês na pág. 1 da ata)

Recursos linguísticos e ferramentas de processamento de linguagem natural do grupo NLX da Universidade de Lisboa

António Branco, João Silva, Francisco Costa, João Rodrigues, Pedro Martins, Eduardo Ferreira, Filipe Nunes, Sérgio Castro, Steven Neale, Catarina Carvalheiro, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Andreia Querido, Rita de Carvalho, Marisa Campos, Nuno Rendeiro, Catarina Correia, Patrícia Gomes, Diana Amaral e Rita Valadas Pereira
Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa

Resumo: No presente artigo, apresentamos muitos dos recursos linguísticos e ferramentas de processamento de linguagem natural criados e disponibilizados na Universidade de Lisboa pelo NLX — Grupo da Fala e Linguagem Natural. Estes foram elaborados, ao longo dos anos, para apoiar o desenvolvimento de um vasto leque de aplicações de processamento de linguagem natural, nomeadamente a tradução automática.

Palavras-chave: português, tecnologias da linguagem, processamento de linguagem natural, recursos linguísticos, ferramentas para o processamento de linguagem natural.

(ler artigo completo em inglês na pág. 9 da ata)

Recursos linguísticos para a extração de dados e o processamento semântico — PLN da PUCRS

Renata Vieira, Daniela do Amaral, Sandra Collovini, Evandro Fonseca, Artur Freitas, Larissa Freitas, Roger Granada, Lucas Hilgert, Lucelene Lopes, Daniela Schmidt, Bernardo Severo e Marlo Souza
Pontifícia Universidade Católica do Rio Grande do Sul — Porto Alegre
Cassia Trojahn
Université Toulouse — Jean Jaurès & Institut de Recherche en Informatique de Toulouse

Resumo: No presente artigo, apresentamos uma panorâmica dos recursos linguísticos desenvolvidos no laboratório de processamento de linguagem natural (PLN) da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), colocando-os à disposição da comunidade de investigadores.

Palavras-chave: extração de informação, processamento semântico, recursos linguísticos

(ler artigo completo em inglês na pág. 17 da ata)

2. Recursos para tarefas específicas

Processamento de *corpora* com tratamento específico das EM através da ferramenta mwetoolkit e modelos semânticos distribucionais

Silvio Cordeiro[1,2], Carlos Ramisch[2], Marco Idiart[3] e Aline Villavicencio[1]
[1] Instituto de Informática, Universidade Federal do Rio Grande do Sul
[2] Aix-Marseille Université, Centre national de la recherche scientifique, Laboratoire d'Informatique Fondamentale de Marseille
[3] Instituto de Física, Universidade Federal do Rio Grande do Sul

Resumo: As expressões multipalavras (EM) são parte integrante da língua, sendo a sua importância já há muito reconhecida. No entanto, as suas características heterogéneas são um desafio para tarefas e aplicações computacionais, incluindo a tradução automática. No presente artigo discutimos como tratar as EM utilizando o mwetoolkit, uma plataforma independente da

língua para tarefas relacionadas com as EM. Em especial, concentramo-nos em três tarefas: 1) processamento de *corpora* para identificação do tipo de EM a partir de *corpora*; 2) identificação de elementos e anotação de *corpora*; e 3) construção de modelos semânticos distribucionais para deteção de composicionalidade com base em *corpora*. O mwetoolkit proporciona uma plataforma uniforme para criar recursos de EM, discutindo-se a sua utilização para o processamento de EM em inglês e português.

Palavras-chave: expressões multipalavras, identificação de elementos, deteção de composicionalidade.

(ler artigo completo em inglês na pág. 26 da ata)

ZAC: *Zero Anaphora Corpus*. Um *corpus* para a resolução da anáfora zero em português

Jorge Baptista[1,3], *Simone Pereira*[1,3] *e Nuno Mamede*[2,3]
1 Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
2 Universidade de Lisboa, Instituto Superior Técnico
3 Instituto de Engenharia de Sistemas e Computadores — Investigação e Desenvolvimento, Laboratório de Sistemas de Língua Falada

Resumo: O presente artigo descreve um *corpus* de textos em português do Brasil construído com vista à criação de um sistema de resolução de anáforas, o qual faz parte de um sistema mais vasto de processamento da linguagem natural (STRING). O *corpus* ZAC visa a resolução das chamadas anáforas zero, ou seja, uma relação anafórica em que a expressão anafórica (ou anáfora) foi reduzida a zero. O artigo analisa sucintamente as questões linguísticas no processo de resolução de anáforas zero e descreve em pormenor o processo de anotação, bem como os principais aspetos das relações anafóricas assim anotadas.

Palavras-chave: anáfora zero, *corpus*, português do Brasil, resolução de anáforas, processamento de linguagem natural

(ler artigo completo em inglês na pág. 38 da ata)

3. Para além da tradução automática

Recursos para tradução monolingue: um estudo de caso de simplificação do texto para português

Rodrigo Wilkens, Leonardo Zilio, Marco Idiart, Jorge Wagner Filho, Eduardo Ferreira, Luis Mollmann, Bianca Pasqualini e Aline Villavicencio
Instituto de Informática, Universidade Federal do Rio Grande do Sul

Resumo: A simplificação de textos pode ser considerada como uma atividade de tradução monolingue e para se obterem resultados precisos são necessários vários tipos de recursos. Por conseguinte, no presente artigo analisamos os recursos disponíveis para o português e o inglês. Entre estes, discutimos *corpora* simples e gerais, dicionários de palavras simples, tesauros, listas de expressões multipalavras e recursos com anotação semântica. A diferença em termos de quantidade e de cobertura de recursos construídos manualmente para as duas línguas revela o fosso que há ainda a colmatar para o português.

Palavras-chave: recursos lexicais, simplificação lexical, simplificação de textos, português

(ler artigo completo em inglês na pág. 46 da ata)

Construção de um *corpus* paralelo português do Brasil — língua de sinais brasileira utilizando dados de captura de movimento

José Mario de Martino e Paula D. Paro Costa
Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas
Ângelo Benetti
Centro de Tecnologia da Informação Renato Archer, Campinas
Luciana Aguera Rosa, Kate Mamhy Oliveira Kumada e Ivani Rodrigues Silva
Centro de Estudos e Pesquisas em Reabilitação Prof. Dr. Gabriel O. S. Porto, Universidade Estadual de Campinas

Resumo: A língua de sinais brasileira, ou Libras, é a língua reconhecida por lei federal como primeira língua da comunidade surda brasileira. No entanto, os surdos brasileiros ainda enfrentam sérios problemas de acesso aos serviços públicos ou mesmo para prosseguirem nos estudos, uma vez que a maior parte da informação, básica ou avançada, continua a estar disponível apenas em português do Brasil (PB) escrito. Em geral, o conhecimento do PB escrito por cidadãos surdos está longe de ser satisfatório. Neste contexto, a tradução automática de PB para Libras é uma abordagem promissora para ajudar as pessoas surdas a alavancar os seus conhecimentos e representa uma opção valiosa para reduzir barreiras de comunicação, especialmente nas situações em que não está disponível um intérprete de língua de sinais. O presente artigo descreve a nossa abordagem para a construção de um abrangente *corpus* paralelo PB-Libras. A abordagem combina uma metodologia baseada na tradução de livros escolares com uma descrição pormenorizada de sinais gestuais e de expressões faciais, com base em dados de captura de movimento. A metodologia também procura abordar os desafios de se trabalhar com uma língua de sinais que ainda carece de vocabulário escolar.

Palavras-chave: língua gestual, língua de sinais, língua de sinais brasileira, tradução automática, *corpus* paralelo, avatar sinalizador, avatar gestual, captura de movimento

(ler artigo completo em inglês na pág. 56 da ata)