Programação do 59º seminário do GEL

59º SEMINÁRIO DO GEL - 2011	
REFERÊNCIA	ARIANI DI FELIPPO. TERMINOLOGIA BASEADA EM CORPUS: A EXTRAÇÃO DE CANDIDATOS A TERMO SEGUNDO A ABORDAGEM ESTATÍSTICA NO PROJETO TERMINET. IN: SEMINÁRIO DO GEL, 59., 2011, PROGRAMAÇÃO... BAURU (SP): GEL, 2011. ACESSO EM: DD.MMM.AAAA
TÍTULO:	TERMINOLOGIA BASEADA EM CORPUS: A EXTRAÇÃO DE CANDIDATOS A TERMO SEGUNDO A ABORDAGEM ESTATÍSTICA NO PROJETO TERMINET
AUTOR(ES):	ARIANI DI FELIPPO
RESUMO	NESTE TRABALHO, DESCREVE-SE O PROCESSO DE EXTRAÇÃO AUTOMÁTICA DE CANDIDATOS A TERMO (EAT) SEGUNDO A ABORDAGEM ESTATÍSTICA REALIZADO NO ÂMBITO DO PROJETO TERMINET. NO REFERIDO PROJETO, PROPÔS-SE UMA METODOLOGIA SEMIAUTOMÁTICA BASEADA EM CORPUS (ISTO É, CONJUNTO DE TEXTOS EM FORMATO DIGITAL CONSTRUÍDO PARA UM FIM ESPECÍFICO) PARA O DESENVOLVIMENTO DE BASES DE DADOS LEXICAIS TERMINOLÓGICAS NO FORMATO WORDNET (TERMINETS). EM UMA TERMINET, OS TERMOS (ISTO É, SIGNO LINGUÍSTICO QUE REPRESENTA CERTO CONCEITO NO INTERIOR DE UM DOMÍNIO) DA CLASSE DOS NOMES QUE PERTENCEM A UM DOMÍNIO ESPECÍFICO ESTÃO ORGANIZADOS EM FUNÇÃO DE SEUS CONCEITOS SUBJACENTES, OS QUAIS SÃO CODIFICADOS EM CONJUNTOS DE FORMAS SINÔNIMAS (SYNONYM SET OU SYNSET). POR EXEMPLO: O SYNSET {CAR, AUTO, AUTOMOBILE, MACHINE, MOTORCAR} DA WORDNET DE PRINCETON, BASE DE LÍNGUA GERAL DO INGLÊS NORTE-AMERICANO, CODIFICA O CONCEITO CAR. OS SYNSEYTS QUE REPRESENTAM OS CONCEITOS NOMINAIS DEVEM SER ORGANIZADOS PELAS RELAÇÕES DE HIPERONÍMIA E HIPONÍMIA. DENTRE AS ETAPAS DE CONSTRUÇÃO DE UMA TERMINET PREVISTAS PELO PROJETO, DESTACA-SE A EAT. DADO UM CORPUS, A EAT CONSISTE NA IDENTIFICAÇÃO E SELEÇÃO DE UNIDADES LEXICAIS COM POTENCIAL ESTATUTO DE TERMO POR MEIO DE UM FERRAMENTAL COMPUTACIONAL. A EAT PODE SER FEITA COM BASE EM TRÊS DIFERENTES ABORDAGENS: LINGUÍSTICA, ESTATÍSTICA E HÍBRIDA. NO PARADIGMA LINGUÍSTICO, BUSCA-SE IDENTIFICAR OS CANDIDATOS A TERMO COM BASE EM CONHECIMENTO LINGUÍSTICO. A EXTRAÇÃO DE CANDIDATOS SIMPLES É FEITA COM BASE EM DOIS TIPOS DE CONHECIMENTO LINGUÍSTICO: CATEGORIA SINTÁTICA OU NUCLEARIDADE SINTAGMÁTICA. PARA A EXTRAÇÃO DE CANDIDATOS COMPLEXOS, UTILIZAM-SE FREQUENTEMENTE TRÊS TIPOS DE INFORMAÇÃO LINGUÍSTICA: PADRÃO MORFOSSINTÁTICO, EXPRESSÃO INDICATIVA OU NUCLEARIDADE SINTAGMÁTICA. OS PADRÕES MORFOSSINTÁTICOS SÃO SEQUÊNCIAS DE ETIQUETAS MORFOSSINTÁTICAS (P.EX.: [NOME+ADJETIVO]). AS EXPRESSÕES INDICATIVAS (OU PADRÕES LÉXICO-SINTÁTICOS), POR SUA VEZ, INTRODUZEM DEFINIÇÕES E OS TERMOS DEFINIDOS, POR EXEMPLO: “É UM TIPO DE”, ETC. NO PARADIGMA ESTATÍSTICO, OS CANDIDATOS SÃO EXTRAÍDOS COM BASE NA APLICAÇÃO DE MEDIDAS ESTATÍSTICAS. PARA A EXTRAÇÃO DE UNIDADES SIMPLES, UTILIZA-SE COMUMENTE A FREQUÊNCIA SIMPLES, QUE PODE SER ENTENDIDA COMO A QUANTIDADE DE VEZES QUE UMA PALAVRA (DEFINIDA COMO N-GRAMA OU TOKEN, ISTO É, SEQUÊNCIA DE CARACTERES SEPARADA POR ESPAÇOS EM BRANCO) OCORRE EM UM CORPUS. A EXTRAÇÃO DE UNIDADES COMPLEXAS, POR SUA VEZ, É FEITA COM BASE NAS MEDIDAS INFORMAÇÃO MÚTUA, LOG-LIKELIHOOD RATIO E COEFICIENTE DICE, POIS ESTAS BUSCAM IDENTIFICAR A ESTABILIDADE DE EXPRESSÕES SINTAGMÁTICAS. A EXTRAÇÃO HÍBRIDA CARACTERIZA-SE PELA COMBINAÇÃO DE CONHECIMENTO LINGUÍSTICO E ESTATÍSTICO. NO PROJETO TERMINET, METODOLOGIA PROPOSTA ESTÁ SENDO VALIDADA POR MEIO DA CONSTRUÇÃO DA WORDNET.EAD, BASE TERMINOLÓGICA DO DOMÍNIO DA EDUCAÇÃO A DISTÂNCIA EM PORTUGUÊS DO BRASIL. PARA A CONSTRUÇÃO DA WORDNET.EAD, A EXTRAÇÃO ESTATÍSTICA FOI FEITA POR MEIO DAS MEDIDAS DISPONÍVEIS NO PACOTE NSP (DO INGLÊS, N-GRAM STATISTICS PACKAGE), QUE É CONSTITUÍDO POR UM CONJUNTO DE PROGRAMAS QUE AUXILIA NA ANÁLISE DAS PALAVRAS EM ARQUIVOS NO FORMATO TXT. O NSP DEPENDE DE UMA SÉRIE DE ESPECIFICAÇÕES: (I) REGRAS DE FORMAÇÃO DE TOKENS; (II) REGRAS DE FORMAÇÃO DE NÃO-TOKENS; (III) LISTA DE STOPWORDS, (IV) TAMANHO DO TOKEN E (V) PONTO DE CORTE (ISTO É, FREQUÊNCIA ABAIXO DA QUAL OS CANDIDATOS EXTRAÍDOS SERÃO DESCONSIDERADOS). ALÉM DESSAS ESPECIFICAÇÕES, QUE DELIMITAM A FASE DE PRÉ-PROCESSAMENTO DO CORPUS, A EXTRAÇÃO PELO NSP ENGLOBOU UMA FASE DE PÓS-PROCESSAMENTO, QUE CONSISTIU NA LIMPEZA DAS LISTAS DE CANDIDATOS EXTRAÍDAS. NA SEQUÊNCIA, AS LISTAS LIMPAS DE CANDIDATOS PASSARAM POR UM PROCESSO DE

	VALIDAÇÃO. AS ESPECIFICAÇÕES DO PRÉ-PROCESSAMENTO, O NÚMERO DE CANDIDATOS EXTRAÍDOS, O PROCESSO DE VALIDAÇÃO DOS MESMOS E AS VANTAGENS E DESVANTAGENS DO MÉTODO ESTATÍSTICO UTILIZADO NO PROJETO TERMINET SERÃO DETALHADOS NESTE TRABALHO. (FAPESP 2009/06262, CNPQ 471871/2009-5).