Criação de Ferramentas e Recursos Computacionais de Processamento de Linguagem Natural para Estudos Cognitivo-Computacionais

Clei Antonio de Souza Junior[1],
Orientador: Marco Aurelio Pires Idiart[2]

[1] Ciência da Computação, UFRGS
[2] Instituto de Física, UFRGS

UFRGS PROPESQ
XXV SIC Salão Iniciação Científica
CET - Ciências Exatas e da Terra

1. INTRODUÇÃO

Para o melhor entendimento da evolução do vocabulário de crianças é preciso um estudo mais aprofundado de quais métricas mostram quais palavras são aprendidas primeiro.

Neste trabalho foram escolhidas algumas métricas psicolinguísticas de um corpus e realizadas análises estatísticas sobre os dados para buscar identificar variáveis do processo de aquisição lexical.

2. DADOS

1. Palavras vindas do conjunto de textos CHILDES [MacWhinney, 2000], da língua inglesa.
   - O CHILDES contém transcrições de diálogos de crianças.

Idade (anos)	Palavras
1	8756
2	39313
3	8478
4	7836

2. Informações sintáticas vindas de:
   - subcategorização verbal: Valex [Korhonen et al. 2006];
   - frames das palavras: FrameNet [Baker et al 1998] ;
   - verbos organizados em classes: VerbNet [Kipper et al. 2008];
3. Informações semânticas vindas de:
   - Conjuntos de palavras interligadas por relações léxico-semânticas: WordNet [Miller 1995];
4. Informações psicolinguísticas vindas de:
   - Frequência, idade de aquisição e familiaridade: MRC [Wilson 1988].

3. METODOLOGIA

Seleção dos dados onde a idade da criança é menor que quatro anos dos corpora longitudinais.

Corpora longitudinal é um conjunto de textos onde há dados da mesma criança com várias idades.

Anotação dos verbos, substantivos, adjetivos e advérbios com informações sintáticas, semânticas e psicolinguísticas.

Cálculo da correlação de Spearman para determinar quais informações são diferentes entre as idades. Onde as métricas tem correlação entre -0,3 e 0,3 (baixa correlação)

Construção de um banco de dados com essas palavras.

4. CONCLUSÕES

Resultados que mostram diferenças para:
- polissemia para quatro pares de idade (1 e 2 anos, 2 e 3 anos, 3 e 4 anos e 2 e 4 anos);
- métricas de antônimos (diretos e indiretos) dos adjetivos para o par de idades 1 e 2 anos.

Contribuição para a formação de uma base empírica de dados para outras análises.

As análises poderão ser usadas em trabalhos de classificação de crianças em idades a partir de seu vocabulário.

5. REFERÊNCIAS

MACWHINNEY, B. The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
MILLER, George A. WordNet: A Lexical Database for English. Communications of the ACM. Vol. 38, No. 11: 39-41, 1995.
FELLBAUM, Christiane. Wordnet: An Eletronic Lexical Database. Cambridge, MA: MIT Press, 1998.
BAKER, Collin F., FILMORE, Charles J., LOWE, John B. The Berkeley FrameNet Project. In Proceedings of COLING-ACL'98, pages 86–90, Montréal,Canada, 1998.
KIPPER, Karin; KORHONEN, Anna; RYANT, Neville; PALMER, Martha. A Large-scale Classification of English Verbs. In Language Resources and Evaluation Journal Vol. 42(1), pp. 21-40, Springer Netherland, 2008.
WILSON, M.D. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. Behavioural Research Methods, Instruments and Computers, 20(1), 6-11, 1988
KORHONEN, Anna; KRYMOLOWSKI; BRISCOE, Ted. VALEX Disponível em: http://www.cl.cam.ac.uk/users/alk23/subcat/lexicon.html .

25 anos

MODALIDADE DE BOLSA

AGRADECIMENTO:
PROBIC FAPERGS
Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
Projeto número: CNPq 551964/2011-1