Explorador Diacrónico

Abrir demostrador

Este sistema permite procurar e ver os cambios léxicos de decenas de miles de palabras do castelán ao longo do tempo, concretamente no eixo temporal 1900-2009, utilizando como fonte de datos as representacións semánticas construídas cos n-gramas de Google en español (45 mil millóns).

O usuario busca por unha palabra e un período de tempo e o sistema devolve o sentido da palabra en cada ano do rango buscado. O sentido dunha palabra represéntase polo conxunto de palabras máis similares en termos semánticos e distribucionais. Por exemplo, a palabra "cancro" está estreitamente vinculada en 1910 con "tuberculosis" e "sífilis", pero en 1960 os termos máis próximos son "tumor" e "carcinoma".

A entrada do sistema é unha estrutura de datos na que as palabras están asociadas mediante graos de similitude (Coseno) con outras palabras e por ano. Estes datos foron xerados recientemente polo equipo ProLNat@Ge (Pablo Gamallo, Marcos García) a través de técnicas e módulos de Procesamento de Linguaxe Natural. Especificamente, efectuamos o procesamento semántico de 45 mil millóns de n-gramas, dispoñibles despois do escaneo de máis de 1 millón de libros do proxecto "Google Books". O procesamento semántico consiste en transformar os n-gramas en matrices distribucionais 'palabra-contexto'. Xerourse unha matriz por ano, onde cada palabra é un vector de contectos. Finalmente, calcúlase a similitude entre vectores (palabras) e elíxese, para cada palabra, as 20 máis similares por ano. En total, xerouse unha estrutura de datos de máis de 300MB, que é a entrada do demostrador.

AUTORES

Investigadores
Pablo Gamallo Otero
Marcos Garcia González
Iván Rodríguez Torres

SOFTWARE

Explorador Diacrónico - Explora a semántica do léxico dende 1900 ata a actualidade