Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan

Asahara, Masayuki; Maekawa, Kikuo; Imada, Mizuho; Kato, Sachi; Konishi, Hikari

Alexandria: The Journal of National and International Library and Information Issues, ISSN 0955-7490, Vol. 25, n. 1/2, 2014, p. 129-148

En 2011, el National Institute for Japanese Language and Linguistics (NINJAL) puso en marcha un proyecto de compilación para desarrollar una colección web para investigación lingüística que alcance diez billones de palabras para el año 2016. El proyecto se divide en cuatro categorías: Colección de Páginas, Anotación Lingüística, Publicación y Preservación. Para Colección de Páginas, se emplean rastreadores web para recopilar textos en la web, rastreándose 100 millones de páginas cada tres meses, y reteniendo diversas versiones de textos por períodos de tres meses. Para Anotación Lingüística, los corpus web para estudios lingüísticos contienen información lingüística anotada. Para mejorar la facilidad de uso de estos recursos lingüísticos, se llevan a cabo tareas de normalización tales como eliminación de etiquetas, división silábica de palabras, análisis sintáctico de dependencia, y registro de la estimación. Para Publicación, se publican listas de palabras y datos n-gram basados en el corpus de texto rastreado y anotado. Además, las aplicaciones se están desarrollando para permitir la búsqueda de patrones morfosintácticos en los corpus de los diez billones de palabras. Para Preservación, se conservan las páginas web

rastreadas en orden cronológico como archivos web, principalmente para apoyar el estudio de los cambios lingüísticos en curso. En este trabajo presentamos el diseño básico de las cuatro categorías. Además, se presenta el estado actual del corpus utilizando estadísticas básicas de los datos de rastreo y se discute la importancia de la eliminación de duplicados de frases.

Traducción del resumen de la propia publicación