

APROXIMACION A LA MORFOLOGIA DE DOS NIVELES.

Kimmo Koskeniemi

En 1983 el profesor Koskeniemi definió el modelo computacional de morfología de dos niveles. Este modelo ha tenido una gran aceptación en años posteriores y se distingue por las siguientes características:

- Es un modelo general aplicable a cualquier lengua.
- Es válido tanto para el análisis como para la generación.
- Separa claramente el conocimiento lingüístico y el algoritmo, como consecuencia la implementación para cualquier lengua es sencilla ya que el programa es el mismo.
- Separa claramente el nivel superficial de la palabra a analizar y el nivel léxico o profundo que es el que se representa en el sistema de diccionario (sistema léxico). Así se evita el almacenar distintas formas del mismo morfema debido a cambios morfofonológicos.
- Utiliza un sistema de reglas de aplicación paralela en contraste con los sistemas de reglas de reescritura utilizados en fonología generativa, con lo que el sistema es conceptual y computacionalmente más sencillo.

La diferencia esencial con la morfología generativa es que no hay estados intermedios entre las dos representaciones. Las reglas no ejecutan nada, solo establecen correspondencias entre los dos niveles. Así el reconocimiento de las palabras se reduce a encontrar una representación léxica válida correspondiente a una forma de superficie. Inversamente la generación parte de la representación léxica conocida y busca representaciones de superficie que se correspondan con ella.

Los elementos básicos de la morfología de dos niveles son dos: las reglas y el sistema léxico. Mientras que las reglas sirven para describir las diferencias entre los dos niveles, el sistema léxico define el conjunto de morfemas, clasificandolos según los posibles encadenamientos entre ellos. Consiste en un conjunto de subléticos y en las clases de continuación que regulan las secuencias posibles de raíces y afijos.

El profesor Koskeniemmi presentó un esquema de un procedimiento para la extracción de las reglas de dos niveles. El objetivo de esta aproximación fue aclarar la controvertida cuestión de la naturaleza de la representación morfológica profunda. Presentó una interpretación concreta para los morfofonemas en su estructura profunda, estudiando un método para llegar a una elección más objetiva de esta representación. Además esbozó las líneas maestras para el establecimiento de reglas fonológicas apropiadas.

En la actualidad este formalismo ha sido aplicado de forma extensiva para 5 lenguas (finlandés, inglés, alemán, árabe y euskara) y de forma experimental para una veintena de lenguas diversas.

Nicoletta Calzolari

Istituto di Linguistica Computazionale
di Pisa

Lexicografía Computacional

El procesamiento del lenguaje aplicado a corpus textuales a gran escala requiere la construcción de léxicos computacionales que contengan cientos de miles de entradas. Cada entrada debe contener de manera explícita información de todo tipo, incluida la semántica.

En este marco, se considera fundamental la reusabilidad de material lexicográfico ya existente. El concepto de reusabilidad apareció por vez primera en el Workshop realizado en Grosseto (1986) y financiado por la CE, donde se planteó la necesidad de diseñar herramientas lingüísticas reusables, multifuncionales y multilingües.

El concepto de reusabilidad debe entenderse en dos sentidos: a) para la extracción y uso de información léxica presente de manera explícita o implícita en fuentes de información léxica diversa (MRDs, Bases de Datos terminológicas, corpus textuales etc.) como ayuda para construir léxicos computacionales a gran escala y reusables en el sentido de b, y b) - para construir léxicos computacionales destinados a diversos tipos de usuarios (diferentes sistemas de PLN, lexicógrafos, lingüistas, usuarios no especializados).

Los léxicos computacionales actuales responden a uno de estos dos tipos. La necesidad de disponer de recursos léxicos que responda a estas características se pone de manifiesto en el impulso dado por la CE a este tema, que se ha traducido en la puesta en marcha de diversos proyectos como son Aquilex (Esprit BRA), Eurotra 7 Multilex (Esprit) y Genelex (Eureka) y la subvención de otro como TEI (Text Encoding Initiatives) y Survey on Linguistic Resources.

El proyecto Esprit Aquilex puede considerarse como un prototipo del primer grupo, mientras que otros proyectos, como Eurotra-7 como un prototipo del segundo.

El proyecto Aquilex

El objetivo de la investigación en el marco de Aquilex se centra en el desarrollo de técnicas y metodologías para extraer información, tanto explícita como implícita, de tipo sintáctico-semántico, a partir de diccionarios en soporte magnético (MRDs) para construir el componente léxico de sistemas de PLN.

Esta metodología se basa en la posibilidad de extraer de manera

automática la información que los MRDs contienen de manera implícita. Así, el diccionario se considera como una fuente de conocimiento básico y los objetivos fundamentales son la formalización de este conocimiento básico general en forma de conceptos y relaciones semánticas.

Entre los temas de investigación conectados con el objetivo último de adquisición de conocimiento cabe destacar: a- el diseño de programas para la extracción de los superordinados a partir de las definiciones, su desambiguación y la construcción de taxonomías; b- el diseño de analizadores para el análisis de las definiciones con el objetivo de extraer la información semántica implícita; c- el estudio de como representar la información semántica extraída (i.e.: en forma de conceptos, atributos, relaciones entre conceptos); d- el estudio de como relacionar las taxonomías y la representación conceptual; e- el diseño e implementación de software básico para la creación, acceso y procesamiento de bases de datos léxicas y una base de conocimiento léxico.

Las taxonomías y las plantillas (estructuras de rasgos en que se representa la información semántica) desarrolladas en el marco de Aquilex constituyen un primer grado de normalización y estandarización en la representación de la semántica y del conocimiento general a partir de diccionarios (unos diez en total) y de lenguas diversas (cuatro).

En el proyecto Aquilex colaboran las universidades de Cambridge, Dublín, Amsterdam, la Universidad Politécnica de Cataluña y el Instituto de Lingüística Computacional de Pisa. Se trabaja sobre diccionarios monolingües del inglés, holandés, italiano, castellano y sobre diccionarios bilingües inglés-holandés e inglés-italiano.

Como complemento indispensable de los trabajos lexicográficos debemos citar las bases de datos textuales. En este marco cabe destacar el programa TEI (Text Encoding Initiative) consistente en una iniciativa para definir criterios generales de codificación e intercambio de textos en soporte magnético. Se trata de un proyecto internacional de cuatro años de duración que tiene como objetivos: a)- especificar un formato uniforme para los textos en soporte magnético; b)- difundir las líneas generales de codificación acordadas; y c)- documentar los esquemas de documentación más importantes y desarrollar un metalenguaje en el que describirlos.

UK National Programmes in Natural Language Research

Karen Sparck Jones

Presentamos un resumen de la ponencia de la Dra. K. Sparck Jones elaborado a partir de la documentación que nos facilitó. Comité Redacción SEPLN.

La ponente presentó diversos programas de investigación y desarrollo en procesamiento del lenguaje natural en el Reino Unido: el Programa Alvey (1983-1988) y el Programa IEATP (Information Engineering Advanced Technology Programme) iniciado en 1988.

El programa Alvey ha tenido una duración de cinco años y ha supuesto una inversión de 350 millones de libras, de los cuales 130 han corrido a cargo de empresas privadas.

En este marco, el desarrollo de sistemas de procesamiento del lenguaje representa una inversión de 2,8 millones de libras, de los cuales 1,7 estaba a cargo de Alvey. La participación humana ha sido de 40 a 50 investigadores.

De entre los proyectos emprendidos en esta línea cabe citar el Japanese/English Machine Translation (UMIST, U. de Sheffield e ICL); el desarrollo de interfaces a bases de datos (U. de Cambridge); una interfaz a un sistema experto médico (ICRF); una interfaz a un sistema planificador (U. de Cambridge, U. de Edimburgo y BT). En la U. de Sussex se ha trabajado en la generación de textos a partir de planes y en la definición de formalismos para la representación de la información léxica.

De entre los proyectos sobre voz, cabe señalar el desarrollado en la U. de Cambridge y el STL sobre reconocimiento del habla continua y el acceso a bases de datos mediante voz desarrollado por la U. de Cambridge, BT y Logica.

Gracias al programa Alvey, en la actualidad se cuenta con una infraestructura considerable de recursos sobre estos temas y con una comunidad investigadora activa cualificada.

El proyecto IEATP dispone de un 50% menos de inversión y está orientado estrictamente al desarrollo de proyectos: análisis del contenido de textos hablados (U. de Lancaster y dos empresas privadas); un asistente de editores desarrollado por la U. de Edimburgo y la empresa Syntek; procesamiento de textos poéticos (U. de Sussex y la empresa Racal); resumen automático de textos (U. de Cambridge).

Los objetivos fundamentales de estos programas son evitar la duplicación de trabajos y promover la conexión entre los equipos de investigación.

