Procesamiento del Lenguaje Natural, num. 28 (2002), pp.127-128

Resolución y generación de la anáfora pronominal en español e inglés en un sistema interlingua de Traducción Automática*

Jesús Peral
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante. Alicante, noviembre de 2001
jperal@dlsi.ua.es

Resumen: Este trabajo presenta una aproximación interlingua de un sistema de Traducción Automática (TA) que permite la generación de la anáfora pronominal en el idioma destino. Nuestro sistema mejora otras propuestas desarrolladas hasta la fecha ya que permite generar la anáfora interoracional, detectar las cadenas de correferencia y generar los cero pronombres españoles, aspectos que apenas han sido considerados por otros sistemas. Se ha realizado una evaluación de la generación de los pronombres personales en tercera persona en español e inglés obteniendo unas precisiones de 80,39% y 84,77% respectivamente.
Palabras clave: Traducción Automática, interlingua, resolución y generación de la anáfora pronominal

Abstract: This paper presents an interlingua approach of a Machine Translation (MT) system that allows the pronominal anaphora generation into the target language. Our system improves other proposals presented so far due to the fact that it is able to generate intersentential anaphora, to detect coreference chains and to generate Spanish zero pronouns, issues that are hardly considered by other systems. The generation of third person personal pronouns into Spanish and English has been evaluated obtaining a precision of 80.39% and 84.77% respectively.
Keywords: Machine Translation, interlingua, pronominal anaphora resolution and generation

En este trabajo se ha presentado una aproximación de un sistema interlingua de Traducción Automática (TA) que lleva a cabo la resolución de la anáfora pronominal originada por los pronombres personales de tercera persona en español e inglés y su posterior generación en el idioma destino.

El sistema, al que hemos denominado AGIR (*Anaphora Generation with an Interlingua Representation*), utiliza una serie de fuentes de información lingüísticas (léxicas, morfológicas, sintácticas y semánticas) en los módulos de análisis y generación del mismo. Concretamente, estas fuentes de información se utilizan en las etapas de análisis sintáctico y resolución de problemas lingüísticos que conducen a la última etapa del módulo de análisis, la representación interlingua del texto origen. El módulo de generación recibe como entrada la representación interlingua (en la que se incluye la información lingüística) y, tras las etapas de generación sintáctica y morfológica, realiza la correcta generación de la anáfora pronominal en el idioma destino (español o inglés).

Gracias al uso de información independiente del dominio, AGIR es capaz de resolver y generar las anáforas pronominales en textos no restringidos de cualquier dominio.

Las principales aportaciones de este trabajo son las siguientes:

- La propuesta de un sistema interlingua de TA. La representación interlingua tiene las siguientes características:
  - ◇ Representa el texto completo utilizando la cláusula (en vez de la oración) como unidad básica de la representación.
  - ◇ Contiene las relaciones existentes entre las distintas entidades del texto mediante los enlaces correspondientes.
  - ◇ Representa las unidades léxicas interlingua del texto utilizando su sentido correcto en WordNet.
  - ◇ Se obtiene tras realizar un análisis sintáctico parcial del texto origen lo que

* Tesis Doctoral presentada por Jesús Peral Cortés, dirigida por Antonio Ferrández Rodríguez. Esta investigación ha sido parcialmente financiada por el Ministerio de Ciencia y Tecnología a través del proyecto núm. TIC2000-0664-C02-01/02

ISSN 1135-5948

garantiza su aplicabilidad sobre textos no restringidos de cualquier dominio.

- La construcción de los correspondientes módulos de análisis para inglés y español en AGIR que implican la realización de las siguientes tareas:

  ◇ Análisis léxico y morfológico.

  ◇ Análisis sintáctico parcial.

  ◇ Resolución de problemas lingüísticos. En esta tarea nos hemos centrado en la detección de los pronombres *it* pleonásticos en inglés y en la resolución de la anáfora pronominal originada por los pronombres personales (para español e inglés) y los cero pronombres españoles.

  ◇ Obtención de la representación interlingua. Esta etapa es la última del módulo de análisis y en ella se obtiene la representación interlingua global del texto.

- La construcción de los correspondientes módulos de generación para inglés y español en AGIR. En estos módulos nos hemos centrado exclusivamente en la generación de la anáfora pronominal y cero pronombres en el idioma destino. Para llevar a cabo esta tarea se ha desarrollado un estudio profundo sobre las diferencias (*discrepancias*) entre español e inglés en cuanto al tratamiento de los pronombres. Este estudio nos permitirá realizar una correcta generación de las expresiones anafóricas en el idioma destino.

  En el módulo de generación se llevan a cabo dos tareas:

  ◇ Generación sintáctica. En esta etapa se tratan las discrepancias sintácticas entre español e inglés. Básicamente se han estudiado dos:

  - Pronombres pleonásticos. Concretamente, se ha llevado a cabo la identificación y clasificación de los pronombres *it* pleonásticos en inglés. Estos pronombres no son anafóricos y, por lo tanto, no se deben generar en español.
  - Cero pronombres. En AGIR se ha desarrollado el primer estudio para el español que permite la detección, resolución y generación en inglés de los cero pronombres españoles con función de sujeto.

  ◇ Generación morfológica. En esta etapa se tratarán y resolverán las discrepancias de número y género existentes entre los pronombres en español y en inglés:

  - Discrepancias de número. Se producen por las diferencias del número gramatical entre palabras de distintos idiomas (en este caso español e inglés) que expresan el mismo concepto. Particularmente, estas palabras serán referidas por un pronombre en singular en el idioma origen y por un pronombre plural en el idioma destino o viceversa.
  - Discrepancias de género. Se originan por las diferencias morfológicas existentes entre el español y el inglés en el tratamiento de los pronombres personales. El español marca el género gramatical de los pronombres personales y distingue entre las formas masculina y femenina de los mismos mientras que el inglés, en general, no lo hace (principalmente para los pronombres es plural).

- La evaluación global del sistema AGIR. Las tareas evaluadas fueron las siguientes:

  ◇ Detección de los pronombres *it* pleonásticos en la que se obtuvo una precisión de 88,75%.

  ◇ Detección de los cero pronombres españoles en la que se obtuvieron unas precisiones de 98,17% y 80,58% en la detección de verbos con su sujeto omitido y verbos con su sujeto no omitido respectivamente.

  ◇ Resolución de la anáfora pronominal en inglés en la que se obtuvieron precisiones de 86,61% y 76,81% en corpus con y sin información semántica respectivamente.

  ◇ Resolución de la anáfora pronominal y cero pronombres en español en las que se alcanzaron precisiones de 82,19% y 81,38% respectivamente.

  ◇ Generación de la anáfora pronominal en español y en inglés en las que se obtuvieron precisiones de 80,39% y 84,77% respectivamente.