Creado el 09/05/2014
Última versión: 09/05/2014
M. Antònia Martí i Mariona Taulé, Universitat de Barcelona

Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Enfoques y recursos básicos/NIVELES DE ANÁLISIS DEL LENGUAJE

Pragmática 3

Esta es la última parte de los materiales relativos a los niveles de análisis del lenguaje semántico y pragmático. En este caso nos centramos en la correferencia.

Correferencia

La correferencia es un fenómeno que se da a nivel del discurso y contribuye a establecer la cohesión del mismo.

Cuando en un texto (oral o escrito, narrativo o dialogado) dos o más entidades comparten un mismo referente en el discurso, decimos que son correferentes. En el fragmento corresponde a los elementos marcados con el mismo color y variable.

Un texto como ejemplo de correferencia. Las variables y colores marcan el objeto de las referencias. El símbolo [Ø] la elipsis.

> **Juanjo Rodríguez**$_{(1)}$ estudió **Ingeniería Informática**$_{(2)}$ en la **Universitat Politécnica de Catalunya**$_{(3)}$ y al **[Ø]**$_{(1)}$ acabar **[Ø]**$_{(2)}$ **[Ø]**$_{(1)}$ decidió quedarse para preparar **su**$_{(1)}$ tesis doctoral sobre lenguaje natural, a la vez que **[Ø]**$_{(1)}$ daba clases. Todo hacía suponer que **su**$_{(1)}$ futuro estaba en la **UPC**$_{(3)}$ pero, en el 2011, **Juanjo**$_{(1)}$ recibió un e-mail de **Facebook**$_{(4)}$ en el que **le**$_{(1)}$ **[Ø]**$_{(4)}$ invitaban a entrar en **su**$_{(4)}$ proceso de selección de personal.

Las entidades en el discurso pueden aparecer expresadas de dos modos:

1. de forma plena, como es el caso de Juanjo Rodríguez, Juanjo, Facebook, Ingeniería Informática, Universitat Politécnica de Catalunya (en el ejemplo).
    a. En este caso decimos que las entidades tienen un modo de referir definido.
2. O bien las entidades en el discurso pueden aparecer expresadas de forma abreviada: bien mediante un pronombre, bien mediante la elipsis.
    a. En el ejemplo tenemos el posesivo 'su' y el pronombre personal 'le' referidos a Juanjo Rodríguez y el posesivo 'su' referido a Facebook que ejemplifican la referencia pronominal.
    b. Y tenemos elipsis en los sujetos de 'acabar', 'decidió' 'daba clases' (Juanjo Rodríguez) o en el sujeto de 'invitaban' (los directivos de Facebook).
    c. En los casos en los que se da una referencia abreviada, sea pronominal o elíptica, tenemos referencia anafórica.

Cuando la correferencia se da entre dos (o más) formas plenas, tenemos correferencia definida; cuando se da entre una forma plena y una o más formas abreviadas, tenemos correferencia anafórica o simplemente anáfora.

- Correferencia definida
    - Juanjo Rodríguez y Juanjo,
    - Universitat Politécnica de Catalunya y UPC
- Correferencia anafórica
    - Juanjo Rodríguez y su
    - Juanjo Rodríguez y le
    - Juanjo Rodríguez y [Ø]
    - Ingeniería Informática y [Ø]
    - Facebook y [Ø]

La importancia de la correferencia en la comprensión y la producción del lenguaje es capital ya que nos permite identificar de qué estamos hablando en cada momento del desarrollo discursivo.

En el texto, la correferencia anafórica y definida nos permite identificar que:

- Juanjo Rodríguez es el sujeto de los estudios en Ingeniería informática,
- Juanjo Rodríguez acabó los estudios
- Juanjo Rodríguez daba clases
- Juanjo Rodríguez hizo la tesis sobre lenguaje Natural
- Juanjo Rodríguez recibe una propuesta de trabajo de facebook
- Juanjo Rodríguez es Juanjo
- Facebook propone a Juanjo/Juanjo Rodríguez entrar en el proceso de selección de personal

Es por ello que identificar las cadenas correferenciales en el discurso es un factor clave en muchas aplicaciones y técnicas de PLN. Si no se dispone de un buen mecanismo de detección de las cadenas correferenciales:

- podemos recuperar entidades erróneas o perder parte importante de la información;
- podemos resumir eliminando segmentos clave de un texto o
- podemos tergiversar totalmente el sentido de un texto en el caso de la traducción automática.

Actualmente la resolución automática de la correferencia es un tema clave en los buscadores de internet ya que permitiría recuperar más información y más precisa.

Los sistemas de resolución automática de la correferencia, distinguen la resolución de la correferencia anafórica o de la anáfora de la resolución de la correferencia definida.

En el primer caso, el problema consiste en hallar el antecedente de una forma pronominal o de una elipsis, normalmente un sintagma nominal. La información que suelen usar los sistemas de resolución automática de la anáfora es la concordancia de género, número y persona (cuando ello es posible) o la proximidad en el texto. En el ejemplo, consistiría en

hallar los antecedentes de las formas pronominales ‘su’ y ‘le’ (Juanjo Rodríguez o Facebook) y de los sujetos elípticos.

La resolución de la correferencia definida consiste en identificar aquellos elementos del texto que se refieren a una misma entidad del discurso.

Los sistemas de resolución automática de la correferencia suelen recurrir a la identidad total o parcial de formas (en el ejemplo, Juanjo Rodríguez y Juanjo).

La resolución de relaciones de correferencia cuando las entidades no son idénticas o parcialmente semejantes en la forma como ocurre con ‘Universitat Politécnica de Catalunya’ y ‘UPC’ es más compleja y se suele recurrir a la información del contexto, de la estructura sintáctica o argumental cuando es posible, o bien a listas de entidades con sus posibles variantes. Este último recurso es muy común cuando se trata de nombres propios.

La resolución automática de la referencia anafórica y definida es crucial en muchas de las aplicaciones del procesamiento del lenguaje:

- Resulta imprescindible para hallar la traducción correcta de las formas pronominales en los sistemas de **traducción automática**, en especial cuando estas distinguen género y número como es el caso del español.
- En **extracción de información** permite identificar los documentos o fragmentos de texto que hacen referencia a una determinada entidad del discurso.
- Los sistemas de **resumen automático** y de **pregunta/respuesta** mejoran notablemente la selección de las frases relevantes si se tienen en cuenta las relaciones correferenciales.

Un tema que actualmente despierta mucho interés es la resolución de la correferencia entre documentos (*cross-document co-reference resolution*). Esta técnica se centra en la extracción de oraciones que hacen referencia a una misma entidad (p.e. José Ignacio Wert) en una colección de documentos, como por ejemplo Internet. Para ello es necesario disponer de mecanismos de desambiguación ya que una misma entidad (en este caso, un mismo nombre propio) puede tener más de un referente.