RESOLUCIÓN DE LA ANÁFORA DISCURSIVA MEDIANTE UNA ESTRATEGIA DE INSPIRACIÓN VECTORIAL

Celia Rico Pérez
Departamento de Filología Inglesa
Universidad de Alicante
fing@vm.cpd.ua.es

Abstract

One of the main problems in anaphora resolution is how to find the correct antecedent for each anaphoric expression that appears in the discourse. As shown in a previous study (Rico, 1993), the type of knowledge required to select an antecedent comprises different sources of linguistic information (morphology, syntax, semantics and pragmatics) and it is only through the coordination of all these sources that we will be able to approach anaphora from an effective point of view.

The strategy for anaphora resolution that I present here allows the integration of the mentioned sources by means of a vector-like model. According to it, I consider that each anaphoric expression and discourse entity is a vector and therefore, can be subject to vector comparison techniques in such a way that closer vectors will indicate closer anaphoric links.

1. Introducción.

Los modelos computacionales que se han ocupado recientemente de la identificación de antecedentes anafóricos tienen todos un punto en común a pesar de que las estrategias empleadas sean siempre diferentes. Este punto de acuerdo consiste en la necesidad de coordinar diferentes fuentes de información lingüística para poder localizar correctamente un antecedente.

A este respecto, mostrábamos en un estudio anterior (Rico, 1993) que los tipos de conocimiento que determinan la selección de una entidad discursiva como antecedente surgen por igual de la información morfológica, sintáctica, semántica y pragmática contenida en

todos los elementos que componen el discurso. La conclusión inmediata a la que llegábamos es que no existe un único criterio que permita por sí mismo seleccionar el antecedente y, por tanto, no podemos tampoco predecir el tipo de información que actuará en cada relación anafórica.

De lo dicho se desprende que cualquier modelo que quiera explicar la anáfora en toda su extensión deberá tener en cuenta todas y cada una de las fuentes de conocimiento y buscar el modo de integrarlas en una única estrategia.

En la presente comunicación describimos un modelo de identificación de antecedentes que consigue coordinar la información morfológica, sintáctica, semántica y pragmática contenida en el discurso mediante una estrategia de inspiración vectorial.

2. Descripción general de la estrategia.

Exponemos en este apartado los fundamentos teóricos sobre los que construimos nuestra estrategia de identificación de antecedentes y comprobaremos en el apartado siguiente su validez para resolver el problema que nos ocupa.

2.1. Requisitos previos.

Uno de los primeros puntos sobre los que es necesario reflexionar antes de emprender el diseño de la estrategia consiste en decidir el modo en que pretendemos coordinar las fuentes de información. Tal como apunta Carter (1990), los modelos de resolución anafórica desarrollados hasta ahora han organizado el conocimiento de dos maneras posibles.

En primer lugar, pueden adoptar el enfoque *democrático*, que se caracteriza porque todas las fuentes de información tienen asignado el mismo papel en la selección del antecedente anafórico. Según este enfoque, las unidades lingüísticas o entidades conceptuales que pueden ser susceptibles de convertirse en antecedente pueden surgir, por igual, de aportaciones de la información morfológica, de la información sintáctica, de la información semántica o de la información pragmática. Son aproximaciones democráticas las desarrolladas por Alshawi (1987), Carbonell y Brown (1988) y Rich y LuperFoy (1988).

El segundo tipo de enfoque es el que Carter denomina *consultivo*, cuya característica principal es que únicamente una fuente de información está capacitada para proponer

antecedentes, mientras que la función del resto de fuentes de información es confirmar o rechazar los antecedentes propuestos. Se trata, por tanto, de un enfoque basado en la consulta a varias fuentes de conocimiento con el fin de ratificar la selección propuesta con anterioridad por otra fuente de información. Las aproximaciones consultivas parten fundamentalmente de los trabajos de Grosz (1977) sobre la teoría del foco del discurso y sus posteriores ampliaciones (Sidner, 1979; Grosz, Joshi y Weinstein, 1983; Grosz y Sidner, 1986) y se han implementado en sistemas como SPAR (Carter,1987), CLE (Alshawi,1992) o Brennan *et al* (1987).

La estrategia que presentamos aquí se encuadra dentro de los enfoques democráticos porque pretende coordinar todas las fuentes de conocimiento de tal manera que se consiga un tratamiento simultáneo de toda la información, evitando, de este modo, la selección anticipada de un antecedente que quizá sea rechazado más tarde.

Como consecuencia del enfoque adoptado, nuestra estrategia cumplirá los siguientes requisitos:

a) Efectuará un tratamiento simultáneo de toda la información lingüística.

b) Asignará distintos grados de relevancia a cada fuente de información en función del contexto lingüístico, sin incidir de antemano sobre determinadas fuentes de información. De este modo, evitaremos que otras fuentes queden relegadas a un segundo plano en la selección del candidato correcto.

c) Comparará en términos de igualdad cada antecedente posible y la expresión anafórica. Nuestra intención es que todas las entidades discursivas tengan las mismas posibilidades de ser consideradas como antecedente, con lo cual evitaremos que se establezcan relaciones de preferencia *a priori* y no excluiremos ningún candidato.

2.2. La estrategia: el producto escalar.

Se ha utilizado en los modelos conexionistas una operación estándar, el producto escalar, que nos ayudará a tratar la información en los términos expuestos en el apartado anterior. El producto escalar es una operación tomada del álgebra y se define como:

$$P1 \cdot P2 = \sum_{k}^{n} P1_k x P2_k$$

donde P1 y P2 son vectores que tienen *n* elementos, $P1_k$ es el elemento *k* de P1 y $P2_k$ es el elemento *k* de P2.

Como vemos, el producto escalar opera con vectores y el resultado que se obtiene tras su aplicación es un número o escalar. Este número permite fijar la posición relativa de los dos vectores ya que facilita el calculo del ángulo que separa ambos vectores[1].

El producto escalar se ha utilizado con éxito anteriormente en sistemas como PARROT (Sutcliffe, 1989), un prototipo construido con el objetivo de ensayar un mecanismo que produjera paráfrasis de historias o secuencias relacionadas con hechos y situaciones típicas de una casa. En este sistema, el producto escalar es la operación básica y permite comparar el significado de diferentes conceptos. Para ello, Sutcliffe parte de la idea de que los significados de las palabras pueden considerarse como vectores, de manera que dos palabras con el mismo significado corresponderán a vectores muy próximos y, por el contrario, dos palabras de significado diferente serán dos vectores muy distantes.

A la vista de los resultados obtenidos para el analisis semántico, pensamos que el producto escalar puede servir igualmente para la identificación de antecedentes anafóricos. Así pues, nuestra estrategia se centra en la definición de las relaciones anafóricas en forma vectorial, de manera que las expresiones anafóricas y las entidades discursivas están representadas por vectores y, por tanto, para localizar el antecedente correspondiente a una determinada expresión anafórica examinaremos cada uno de los vectores de las entidades discursivas y seleccionaremos el que esté mas próximo.

---

[1] Para saber qué posición relativa tiene un vector respecto del otro, necesitamos conocer el ángulo θ que separa *u* de *v*. Calcularemos θ mediante la siguiente fórmula:

$$u \cdot v = \|u\| \|v\| \cos\theta$$

donde

|u| es el módulo de *u*, esto es, su longitud
|v| es el módulo de *v*

3. Aplicación del producto escalar a la identificación de antecedentes.

Tal como queda expuesto en el apartado anterior, la aplicación del producto escalar a la identificación de antecedentes anafóricos exige que tanto las entidades discursivas como las expresiones anafóricas sean consideradas como vectores. Sólo así será posible utilizar el producto escalar como método de comparación entre posibles candidatos a antecedente.

Así pues, para codificar la información lingüística en forma vectorial procedemos del siguiente modo:

1- Definimos un conjunto de atributos anafóricos que son aplicables de manera general tanto a entidades discursivas como a expresiones anafóricas. Este conjunto de atributos contiene toda la información lingüística necesaria para consolidar una relación anafórica y está basado en Rico (1993).

2- Establecemos los grados de aplicación de cada atributo en función de la información lingüística que se extrae del antecedente y la expresión anafórica. Codificamos los grados de aplicación mediante el empleo de valores numéricos, de manera que si un atributo se aplica en mayor grado que otro el valor correspondiente será más alto.

3- Consideramos que el conjunto de atributos que definen una entidad discursiva concreta, es decir, los valores numéricos asignados, constituye el vector que representa esa entidad. De igual modo, el conjunto de atributos y valores numéricos asignados a una expresión anafórica constituyen el vector de esa expresión anafórica. Así, el vector contendrá todas las fuentes de información codificadas con valores numéricos.

Una vez que la información está codificada en forma de vectores, el producto escalar actuará como operación básica para comparar los vectores correspondientes a todas las entidades discursivas y el vector correspondiente a la expresión anafórica. El resultado que obtendremos será una lista de las entidades ordenadas según su mayor o menor proximidad con el vector correspondiente a la expresión anafórica.

Veremos a continuación cómo actúa nuestra estrategia en casos reales.

4. Identificación de antecedentes.

Para probar la validez de nuestra estrategia construímos un prototipo que utiliza el producto escalar como operación básica para la identificación de antecedentes. El prototipo consta de diferentes módulos de analisis que aportan toda la información lingüística necesaria para establecer las relaciones anafóricas, así como otros módulos que se centran en la implementación del producto escalar como operación básica. En el cuadro I mostramos la arquitectura general de nuestro prototipo.

Los módulos de analisis morfológico y sintáctico son adaptaciones de los analizadores morfológico y sintáctico desarrollados por otros autores para otras aplicaciones. Así, en el caso del modulo morfológico nos basamos en el desarrollado por Gal *et al* (1991) y con respecto al modulo sintáctico adoptamos el desarrollado por Amores (1992). El resultado que obtenemos tras el analisis es una estructura de pares atributo-valor donde queda reflejada toda la información morfológica y sintáctica que utilizaremos mas tarde para establecer relaciones anafóricas.

El resto de módulos ha sido construido específicamente para nuestra aplicación y siempre con el objetivo de resolver la anáfora discursiva. Así, gracias al modulo de identificación de unidades anafóricas podremos extraer del enunciado todas las entidades discursivas y las expresiones anafóricas. A continuación, asignamos a estas unidades los rasgos semánticos que les corresponden según una jerarquía de rasgos semánticos previamente definida (Rico, 1994).

El modulo de asignación de rasgos pragmáticos actúa inmediatamente después de la asignación semántica. Los rasgos pragmáticos que aplicamos son dos: distancia clausal y distancia oracional. Ambos rasgos nos permitirán establecer dos restricciones clave para identificar los antecedentes correctos. Por una parte, la restricción de los pronombres reflexivos para seleccionar como antecedente sólo aquellas entidades discursivas que se encuentren dentro de la misma cláusula que el pronombre; por otra parte, la necesidad de buscar el antecedente de un pronombre no reflexivo fuera de la cláusula donde éste aparece.

En el siguiente modulo, identificación de relaciones anafóricas, tienen lugar diferentes

Módulos	Input - Output
	enunciado
Análisis morfológico	
	unidades léxicas
Análisis sintáctico (LFG)	
	estructura-c estructura-f
Identificación de unidades anafóricas	
	entidades discursivas expresiones anafóricas
Análisis semántico	
	rasgos semánticos
Análisis pragmático	
	rasgos pragmáticos
dentificación e relaciones anafóricas	pares atributo-valor
Asignación de valores numéricos	
	vectores
Normalización	
Aplicación del producto escalar	antecedentes localizados

Cuadro I

procesos que expondremos a continuación mediante el análisis, a modo de ejemplo, de los enunciados *The children bought he cakes* y *They ate them.*

En primer lugar, tomamos la primera expresión anafórica localizada, que en este momento tiene codificados como una lista de pares atributo-valor todos sus rasgos morfológicos, sintácticos, semánticos y pragmáticos. En los enunciados seleccionados, la expresión anafórica *they* obtendrá la siguiente representación:

```
[oracion : 2,cláusula : 1,subj : [pred : they,
gen : indefinido,pers : terc,num : plur,
cara:no,animate:yes, count:yes,edible:no,
human:yes,object:no,distancia1:dif_cla,
distanci2:dif_orac]]
```

A continuación, es necesario, como apuntamos en el apartado 3, que convirtamos toda esta información en un vector. El predicado encargado de realizar esta tarea es el siguiente:

```
convierte(Expresion(N,[H|T]),Expresion(E,[V1|V2])):-
      inlist(pred:E,[H|T]),
      convierte_valor([H|T],[V1|V2]).
```

El primer argumento de este predicado, (Expresion(N,[H|T])) es la expresión anafórica con todos sus rasgos y el segundo argumento, (Expresion(E,[V1|V2]), es el vector resultante.

La conversión en vectores se realiza mediante la asignación de valores numéricos, consultando para ello unas tablas que contienen las correspondencias numéricas[2] para cada par atributo-valor. Mostramos aquí la tabla correspondiente a los valores para la información morfológica.

```
valor(restrictivo,gen:neutro,[3,0,0]).
```

---

[2] Con relación a la asignación de valores numéricos, apuntaremos que, si bien los valores se han determinado manualmente, no existe arbitrariedad en ello puesto que su asignación es el fruto de un trabajo exhaustivo sobre el corpus descrito en Rico (1993). El método empleado consiste en asignar unos valores originales, experimentar los resultados obtenidos con ellos y ajustar cada valor en función de los errores cometidos en la selección del antecedente. Sería deseable sin embargo, que en versiones posteriores de sistema, se probara con métodos conexionistas que asignaran automáticamente estos valores.

Agradecemos, en este sentido, las sugerencias de Julio Gonzalo, del Instituto de Ingeniería de Conocimiento de la Universidad Autónoma de Madrid así como los comentarios del Dr. Gabriel Amores del Departamento de Lengua Inglesa de la Universidad de Sevilla.

```
valor(restrictivo,gen:masculino,[0,3,0]).
valor(restrictivo,gen:femenino,[0,0,3]).
valor(restrictivo,gen:indefinido,[3,3,3]).
valor(restrictivo,num:sing,[3,0]).
valor(restrictivo,num:plur,[0,3]).
valor(restrictivo,num:colectivo,[3,3]).
valor(restrictivo,pers:prim,[3,0,0]).
valor(restrictivo,pers:seg,[0,3,0]).
valor(restrictivo,pers:terc,[0,0,3]).
```

Así pues, el vector que corresponde a la expresión anafórica anterior es:

```
ea(they,[0,0,4,0,3,3,3,0,0,3,0,3,0,2,2,0,2,0,0,2,2,0,0,2,0,1,0]
```

Una vez que hemos seleccionado la primera expresión anafórica y hemos asignado los valores numéricos correspondientes para construir su vector debemos identificar el antecedente correcto para la expresión anafórica.

Para llevar a cado esta tarea, el programa considera una por una todas la entidades discursivas y calcula la distancia real que las separa de la expresión anafórica que se está tratando en ese momento. Esto se consigue gracias al siguiente predicado:

```
calcula_distancia(ea(_,[O,Cla|_])):-
      call(edis(N1,[O2,Cla2|T2])),
      oracion(O,O2,O3),
      cláusula(Cla,Cla2,Cla3),
      append(Cla3,O3,Dist),
      append([O2,Cla2|T2],Dist,T3),
      convierte(edis(N1,T3),edis(n2,T4)),
      normaliza(edis(N2,T4), fail.
```

Esta regla toma la primera entidad discursiva que existe en la base de datos interna de Prolog y a continuación comprueba cuál es la distancia oracional y clausal de la entidad discursiva con respecto a la expresión anafórica. Después convierte la entidad discursiva en un vector.

Una vez que la distancia ya ha sido asignada a todas las entidades discursivas, los pares atributo- valor de todas ellas están ya completos y se puede proceder a la comparación de estas entidades con la expresión anafórica para así localizar el antecedente correcto. Esta comparación se lleva a cabo mediante el producto escalar que, como ya definimos, constituye la base central de la estrategia de identificación de antecedentes.

El resultado final es una lista ordenada de las entidades discursivas en función de si

de si sus vectores están más o menos próximos al vector correspondiente a la expresión anafórica.

En los enunciados que tomamos como ejemplo el resultado final sería el siguiente:

```
ANTECEDENTE DE   pred:them
      pred:cake (0.8643)
   Antecedentes rechazados:
      pred:child (0.8371)
ANTECEDENTE DE   pred:they
      pred:child (0.9883)
   Antecedentes rechazados:
      pred:cake (0.7041)
```

5. Resultados.

Con respecto a los resultados obtenidos en la aplicación de la estrategia descrita, veremos a continuación una serie de casos donde el antecedente se ha seleccionado correctamente:

a) Selección del antecedente correcto mediante criterios gramaticales.

b) Selección del antecedente correcto mediante criterios sintácticos.

c) Selección del antecedente correcto mediante criterios semánticos.

5.1. Selección del antecedente mediante criterios gramaticales.

Tomamos como ejemplo el siguiente texto:

Peter bought the fish with Mary.
She ate it.

Al aplicar el producto escalar como estrategia para comparar los vectores resultantes de las entidades discursivas y las expresiones anafóricas, conseguimos los siguientes valores:

```
ANTECEDENTE DE pred:it
      pred:fish (0.852459016393443)
   Antecedentes rechazados:
      pred:peter (0.652123598746764)
      pred:mary (0.61594731974624)
```

Como vemos, el producto escalar ha seleccionado como antecedente del pronombre *it* la entidad discursiva *fish* ya que su vector correspondiente está más próximo que el resto.

5.2. Selección del antecedente mediante criterios sintácticos.

Los ejemplos que mostramos en este apartado contienen relaciones anafóricas donde el antecedente se localiza gracias a que ocupa la misma posición sintáctica que su expresión anafórica.

a) Mary played with the girl.
Peter also played with her.

Obtenemos el resultado siguiente:

```
ANTECEDENTE DE pred:her

      pred:girl (0.982142857142857)

   Antecedentes rechazados:

      pred:mary (0.956100492239443)
      pred:peter (0.804361815109743)
```

Sin embargo, cuando cambiamos el texto la referencia anafórica también cambia:

b) Mary played with the girl.
She also played with Peter.

En este texto el antecedente seleccionado es:

```
ANTECEDENTE DE pred:she

      pred:mary (0.98529411764706)

   Antecedentes rechazados:

      pred:girl (0.956100492239443)
      pred:peter (0.803115707754045)
```

5.3. Selección del antecedente mediante criterios semánticos.

Comprobaremos aquí si nuestra estrategia puede decidir qué antecedente es el correcto cuando el verbo impone restricciones de selección.

Los ejemplos que analizaremos son los siguientes:

a) A cake was on the table.
Mary ate it.

b) A cake was on the table.
Mary had painted it.

En (a) el verbo, *ate*, impone la restricción de que su objeto directo debe tener el rasgo semántico de «comestible» y, por tanto, sólo *cake* puede ser antecedente ya que *table* no tiene este rasgo.

En (b) el verbo, *had painted*, necesita un objeto directo que tenga el rasgo semántico de «cosa» y, por tanto, sólo *table* puede ser antecedente ya que *cake* no tiene este rasgo.

Veamos cómo se resuelve esta situación aplicando el producto escalar.

Resultado del producto escalar para el enunciado (a):

```
ANTECEDENTE DE pred:it

    pred:cake (0.983606557377705)

  Antecedentes rechazados:

    pred:table (0.838372740765715)
    pred:mary (0.6473808187208)
```

El resultado del producto escalar para el enunciado (b) es:

```
ANTECEDENTE DE pred:it

    pred:table (0.975249922931546)

  Antecedentes rechazados:

    pred:cake (0.852459016393443)
    pred:mary (0.6473808187208)
```

6. Conclusiones.

Hemos descrito en esta comunicación una estrategia de identificación de antecedentes que consigue integrar todas las fuentes de información lingüística en un único patrón, tratándolas de modo simultáneo.

Para ello, hemos creado un modelo de interpretación anafórica que concibe las relaciones anafóricas en forma vectorial, de modo que podemos aplicar las técnicas de comparación de vectores para localizar los antecedentes.

De este modo, hemos codificado en forma de vectores toda la información lingüística contenida en los enunciados que componen el discurso y hemos obtenido como resultado un vector diferente por cada entidad discursiva y expresión anafórica que aparece en el texto.

A continuación, nuestra estrategia emplea el producto escalar como operación básica para efectuar la comparación de vectores y seleccionar el antecedente correcto. En este sentido, hemos comprobado en el último apartado de nuestra comunicación que los resultados obtenidos al aplicar el producto escalar son correctos.

Referencias bibliográficas

Alshawi, H. (1987): *Memory and Context for Language Interpretation*. Cambridge: Cambridge University Press.

Alshawi, H. y J. Van Eijck (1989): «Logical forms in the Core Language Engine» en *ACL Proceedings, 27th Annual Meeting.*

Alshawi, H. (1990): «Resolving quasi logical forms» en *Computational Linguistics* 16 (3): 133-144.

_____ (1992): *The Core Language Engine*. Cambridge: MIT Press.

Amores Carredano, J. G. (1992): *A Lexical-Functional Grammar Machine Translation System for Medical Abstracts.* Tesis doctoral. Universidad de Sevilla.

Brennan, S., M. Friedman y C. Pollard (1987): «A centering approach to pronouns» en *Proc. of the 25th Annual Meeting of the ACL*: 155-162.

Carbonell, J. G. y R. D. Brown, (1988): «Anaphora resolution: a multy-strategy approach» en *COLING 1*: 96-101.

Carter, D. M. (1987): *Interpreting Anaphors in Natural Language*. Chichester, UK: Ellis Horwood.

Carter, D. M. (1990): «Control issues in anaphor resolution» en *Journal of Semantics 7*: 435-454.

Gal, A., G. Lapalme, P. Saint-Dizier y H. Somers (1991): *Prolog for Natural Language Processing*. Chichester: J. Wiley & Sons Ltd.

Grosz, B. (1977): «The representation and use of focus in a system for understanding dialogs» en *IJCAI 5.*

Grosz, B., A. Joshi y S. Weinstein (1983): «Providing a unified account of definite noun phrases in discourse» en *Proc. of the 21st Annual Meeting of the Association for Computational Linguistics*: 44-50.

Grosz, B. y C. Sidner (1986): «Attentions, intentions and the structure of discourse» en *Computational Linguistics 12 (3)*: 175-204.

Rich, E. y LuperFoy, S. (1988): «An architecture for anaphora resolution» en *Proceedings of the Second Conference on Applied NLP*: 18-24.

Rico Pérez, C. (1993): «Estudio de la incidencia de diferentes fuentes de información en el establecimiento de relaciones anfóricas» en *Procesamiento del Lenguaje Natural*, 14.

Rico Pérez, C. (1994): *Aproximación estadístico-algebraica a la resolución de la anáfora en el discurso.* Tesis doctoral. Universidad de Alicante.

Sidner, C. (1979): «Towards a computational theory of definite anaphora comprehension in English discourse» en *Technical Report 537.* Cambridge, Mass: MIT Artificial Intelligence Laboratory.

Sutcliffe, R. (1989): *A Parallel Distributed Processing Approach to the Representation of Knowledge for Natural Language Understanding.* Tesis doctoral. Universidad de Essex.