Un modelo para el control de dependencias en sistemas de T.A.

J.A.Alonso
Siemens S.A. CDS

Resúmen

El presente artículo presenta una propuesta de un modelo para formalizar las dependencias de control de concordancia en sistemas de traducción automática. El modelo expuesto se basa en el formalismo de la GPSG y se adapta a sistemas de T.A. basados en "transfer" y que utilicen descripciones estructurales (árboles) con nodos que contengan pares atributo-valor (rasgos) con diversa información morfológica, sintáctica y semántica. En particular, dicho modelo está siendo probado sobre el sistema METAL alemán-castellano.

En dependencias de control se engloban fenómenos como la concordancia entre constituyentes o los fenómenos de rección. Dado que las teorias lingüísticas más relevantes están pensadas sobre todo para análisis, los mecanismos proporcionados por ellas para tratar dichos fenómenos no sirven como tales para un aplicación en la que se deben generar árboles a partir de una información (análisis) previa, como es el caso en la traducción automática. Por lo tanto, es preciso crear nuevos mecanismos lingüísticos (o readaptar mecanismos ya existentes, en la medida de lo posible) con el fin de llegar a una teoría formal de la traducción automática.

UN MODELO PARA EL CONTROL DE DEPENDENCIAS EN SISTEMAS DE T.A.

Juan Alberto Alonso
Siemens S.A. (CDS)
Barcelona

1.- Introducción

El presente artículo intenta presentar una propuesta de un modelo para formalizar las dependencias de control en sistemas de traducción automática, centrándose sobre todo en las dependencias de control de concordancia. El modelo expuesto se adapta a sistemas de T.A. basados en "Transfer" y que utilicen descripciones estructurales [árboles] con nodos que contengan pares atributo-valor [rasgos] con diversa información morfológica, sintáctica y semántica. En particular, dicho modelo está siendo probado sobre el sistema METAL alemán-castellano.

En dependencias de control se engloban fenómenos como la concordancia entre constituyentes o los fenómenos de rección. Todo ello ha sido ampliamente tratado por diversos formalismos lingüísticos [GPSG, LFG, GB, etc.]. Sin embargo, dado que estas teorías están pensadas sobre todo para análisis, los mecanismos proporcionados por ellas no sirven como tales para una aplicación en la que se deben generar árboles a partir de una información [análisis] previa, como es el caso en la traducción automática. Por lo tanto, es preciso crear nuevos mecanismos lingüísticos (o readaptar mecanismos ya existentes, en la medida de lo posible) con el fin de llegar a una teoría formal de la traducción automática.

El modelo para el control de concordancias aquí presentado se basa en los mecanismos que la GPSG ofrece para dicho fenómeno [CAP y HFC] e intenta elaborar una nueva propuesta de los mismos.

Aunque en este artículo no se tratarán en detalle los fenómenos de rección, haremos referencia a ellos cuando sea necesario.

2.- El tratamiento de la concordancia en la GPSG

La GPSG utiliza como mecanismo básico para asegurar la concordancia entre constituyentes uno de los tres principios universales de instanciación de rasgos, en particular, el llamado "Control Agreement Principle" (CAP). Para asegurar la concordancia de elementos reflexivos, la GPSG hace uso de otro principio (el FFP o Foot Feature Principle, que además se encarga de realizar tareas de coindexación para constituyentes anafóricos).

La GPSG define el control entre categorías (entendiendo como control la relación de dependencia de un constituyente sintáctico respecto a otro a efectos de concordancia morfológica) en términos de tipo semántico de las mismas: una categoría controladora debe ser un argumento de la función cuyo funtor en la representación semántica es su categoría controlada. Parte además de la base de que sólo categorías nominales pueden ser controladores de concordancia (sobre categorías verbales, adjetivales o determinantes).

3.- Propuesta de un modelo pensado para T.A.

3.1.- Supuestos previos.

El tipo de sistema sobre el que puede trabajar el modelo propuesto debe utilizar árboles cuyos nodos estén compuestos de rasgos consistentes en pares atributo-valor. En el proceso de transferencia del árbol de análisis, determinados rasgos (léxicos) en los nodos ya transferidos pasan a tener nuevos valores, correspondientes a la lengua de destino.

3.2.- Dependencias de control

Las dependencias de control se pueden clasificar en dos grupos: dependencias de concordancia y dependencias de rección.

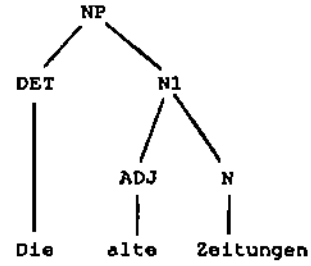
Las primeras implican una identidad de ciertos rasgos gramaticales entre ciertas categorías, por ejemplo, equivalencia de género y número entre núcleo nominal y determinante o adjetivo, o de número y persona entre sujeto y predicado (para el castellano).

Las segundas fuerzan la presencia de ciertos valores para ciertos rasgos gramaticales en determinadas categorías (regidas) dependiendo del valor de ciertos rasgos en otras categorías (rectoras). Así por ejemplo, el tiempo y aspecto de un predicado en una frase principal puede determinar o "regir" el tiempo y aspecto del predicado en la frase subordinada (p.ej. "*Quiero que vengas*"), o la preposición de un SP puede "regir" el caso del NP hermano.

En este artículo vamos a centrarnos principalmente en las dependencias de control de concordancia (DCC).

Decimos que en un árbol local existen DCC si existen categorías hijas en ese árbol cuya transferencia depende de pares atributo-valor provenientes de categorías (hermanas o externas al árbol local) que hayan sido previamente transferidas.

Por ejemplo, en el siguiente árbol de análisis correspondiente a la frase alemana "Die alten Zeitungen"



la traducción de los nodos ADJ y DET al castellano depende de los rasgos de género y número del nodo N, una vez haya sido transferido [Zeitung [GD F] --> Periódico [GD M]: Los periódicos viejos.

3.3.- El problema de la concordancia en T.A.

Un sistema de traducción automática debe construir, a partir de las descripciones estructurales de las oraciones en el idioma de partida, descripciones estructurales equivalentes capaces de generar oraciones gramaticales en el idioma de llegada. Las relaciones de concordancia existentes en el idioma de partida pueden verse alteradas en el idioma de llegada debido a varios factores, como puedan ser la diferencia de rasgos léxicos para ciertas categorías entre las dos lenguas (por ejemplo, un nombre alemán masculino como *Brief* puede tener como traducción un nombre en castellano que sea femenino: *carta*) o diferentes relaciones de control entre constituyentes (un adjetivo en posición predicativa debe concordar en género y número con el sujeto de la oración en castellano [*el estudiante es alto/la estudiante es alta*], pero no en alemán [*Der Student/Die Studentin ist gross*]).

En el proceso de la traducción debemos asegurarnos de que los constituyentes que son controladores efectivamente "controlan" la concordancia con sus controlados. En resumen, debemos elaborar un mecanismo que asegure el cumplimiento de estas dependencias de control.

Esto es particularmente vital en sistemas como METAL, donde la fase de transferencia se realiza transfiriendo los nodos del árbol de análisis de arriba a abajo, y efectuando un tráfico de pares atributo-valor entre determinados nodos. Es el orden en el que se transfieran los nodos hijos de un árbol local, y el tráfico de rasgos entre nodos los que aseguran que la concordancia de la frase de salida será correcta. Actualmente, tanto el orden de transferencia de nodos como el tráfico de rasgos son responsabilidad del lingüista que escribe las reglas gramaticales. La presencia de un mecanismo basado en este modelo que automatice este proceso facilitará la escritura de reglas y hará el sistema más seguro al evitar errores causados por fallos en las reglas gramaticales.

4.- Fundamentos del modelo para DCC.

A continuación definiremos una serie de conceptos que se utilizarán a lo largo de la exposición.

4.1.- Categorías controladoras y controladas.

Llamaremos categoría controladora a cualquier categoría que establezca DCCs sobre otros constituyentes. En principio, adoptaremos la hipótesis de trabajo (cf. Gazdar85) de que sólo las categorías nominales pueden ser controladoras, y no se utilizará ninguna información sobre tipo semántico para definir relaciones de control, como propone la GPSG.

Las categorías controladas serán, por otra parte, aquéllas que dependan de categorías controladoras.

Categorías libres serán aquellos que no sea vean afectados por dependencias de control de concordancia.

4.2- Núcleos de Concordancia y de Rección

Dentro de un árbol local se define su núcleo de concordancia como aquél nodo hijo que:

i) Sea una categoría controladora.
ii) Sea una categoría controlada, en caso de no haber controlador local.

En caso de árboles locales libres (es decir, en los que no existen nodos controladores ni controlados) mantendremos la suposición de que dichos árboles no tienen NC.

El concepto de núcleo de concordancia se complementa con el de "núcleo de rección" (NR) (la categoría rectora dentro de un árbol local).

Cabe destacar la diferencia de concepto entre lo que se entiende por núcleo en la mayoría de teorías lingüísticas (incluyendo la GPSG) y la definición de núcleo que aquí se presenta. Esta es una definición de núcleo totalmente "ad-hoc", pensada para los fenómenos que se quieren tratar. Es por ello que no siempre tiene porqué interesarnos definir el núcleo de un árbol local como el nodo hijo con iguales valores de rasgos V y N, y un valor de BAR igual o menos al del nodo raíz.

4.3.- Tipos de rasgos.

4.3.1.- Rasgos de Control

Se definen como rasgos de control aquellos rasgos pertinentes para el control de concordancia entre categorías. La lista de rasgos de control es la unión de los rasgos de control pertinentes para cada categoría controlada y es obviamente dependiente del lenguaje específico que se esté tratando. Así, por ejemplo, para categorías controladas verbales [+V,-N], los rasgos de control en castellano serán NU (número) y PS (persona). Ver sección 5 para una lista completa de rasgos de control en castellano.

En este aspecto, el modelo se aparta considerablemente de la GPSG, donde se exige que la unificación de la totalidad de rasgos que componen la categoría controladora con el valor AGR de la categoría controlada.

4.3.2.- Rasgos de Rección

Se definen como rasgos de rección aquellos rasgos pertinentes para el control de rección entre categorías rectoras y regidas. En este artículo no se entrará en más detalles sobre este punto.

4.3.3.- Rasgos nucleares.

Se definen como rasgos nucleares el conjunto de rasgos resultante de la unión entre el conjunto de rasgos de control y el conjunto de rasgos de rección.

4.4.- Principios de instanciación de rasgos.

Estableceremos dos principios que aseguren el tráfico correcto de los rasgos de control y rección en los momentos adecuados durante el proceso de la transferencia. Como formalismo, utilizaremos la siguiente representación:

F(C)(R)|T

donde F es una función que devuelve los valores de los rasgos R para una categoría C en un instante T.

Asimismo, utilizaremos la notación:

F(C1)(R)|T ==> F(C2)(R)|T

para indicar que los rasgos de la categoría C2 toman el valor que tenga la categoría C1 para los rasgos pertenecientes al conjunto R que estén presentes en ella, todo ello en el instante T.

La introducción del factor temporal en los principios es otra de las particularidades del modelo. De hecho, lo importante es distinguir entre el valor de los rasgos en nodos que no hayan sido todavía transferidos o que en efecto lo hayan sido ya, por lo que una notación igualmente válido para nuestros propósitos podría ser:

F(C1)(R) ==> F(C2)(R')

donde los valores de R y R' pueden corresponder a rasgos de nodos antes (R) o después (R') de haber sido transferidos.

Dado que METAL funciona con un tipo de transferencia secuencial, en el que los nodos se van transfiriendo uno tras otro, y es por tanto fácil establecer relaciones temporales (frente a otros sistemas como

EUROTRA, en los que al utilizar lenguajes declarativos como Prolog y utilizar a fondo técnicas de unificación, el proceso de transferencia de nodos no es fácil de ver como secuencial), haremos uso de la notación con el parámetro T en nuestros principios.

4.4.1.- Principio de Rasgos de Control [PRC]

El PRC exige que:

(i) Para todo árbol local cuyo núcleo de concordancia NC sea una categoría controladora se debe cumplir la siguiente equivalencia:

F(Cnc)(Rc)|Ti ==> F(Ci)(Rc)|Ti

donde Cnc es la categoría núcleo (de concordancia) controladora y Ci es la categoría o categorías controladas por Cnc, Rc son los rasgos de control correspondientes a cada Ci en particular, y Ti representa el instante en el que Cn ya se ha transferido y la(s) categoría(s) Ci aun no lo ha(n) sido.

(ii) Si el árbol local no tiene controlador (local) y por lo tanto su núcleo NC es una categoría controlada, se debe cumplir la siguiente equivalencia:

F(Crz)(Rc)|Ti ==> F(Ci)(Rc)|Ti

donde Crz es la categoría raíz del árbol local y Ci es la categoría o categorías controladas por alguna Cnc externo al árbol local, Rc son los rasgos de control correspondientes a cada Ci en particular, y Ti representa el instante en el que Cnc ya se ha transferido y la(s) categoría(s) Ci aun no lo ha(n) sido.

PRC(ii) asegura el tráfico de rasgos de control hacia abajo, de nodos raíz a núcleos de concordancia en caso de árboles locales sin controlador.

PRC(i) asegura además del tráfico hacia abajo, el orden de transferencia y el tráfico de rasgos horizontal, de controlador a controlado, para árboles locales con controlador.

4.4.2.- Principio de Rasgos Nucleares [PRN]

Debemos contar con algún principio que asegure el tráfico de rasgos desde los núcleos a los nodos padre, una vez los nodos léxicos han sido ya transferidos.

La aplicación de este principio presupone la existencia de otro principio, similar al PRC del punto anterior, que gobierne el comportamiento de las dependencias de control de rección (Principio de Rasgos de Rección, PRR). En principio, supondremos que un árbol local puede tener tanto núcleos de concordancia (NC, obligatoriamente) como núcleos de rección (NR, opcionalmente y que ambos pueden referirse a nodos distintos (en un árbol local dominado por un nodo S, el NC sería el NP sujeto, mientras que el NR sería el VP)

El PRN exige que para todo árbol local no libre se cumplan las siguientes equivalencias:

F(Cnc)(Rn)|Tnc ==> F(Crz)(Rn)|Tnc
F(Cnr)(Rn)|Tnr ==> F(Crz)(Rn)|Tnr

donde Crz, Cnc y Cnr son la categoría raíz, la categoría núcleo de concordancia y la categoría núcleo de rección (si existe) del árbol local, respectivamente, Rn son los rasgos nucleares, y Tnc/Tnr son los instantes posteriores al momento de la transferencia de las categorías núcleos Cnc y Cnr, respectivamente.

En resumen, el principio de rasgos nucleares asegura que los valores de los rasgos englobados en dicho conjunto coinciden en la categoría raíz con los de los núcleos de un árbol local después de haber transferido éstos a la lengua de llegada (tráfico de rasgos vertical, hacia arriba, después de la transferencia).

5.- Datos sobre DCCs para el castellano.

5.1.- Categorías Controladoras

[+N,-V,-INDEX, (BAR=2 ==> ROL = SUBJ)]

Es decir, sólo son controladores de concordancia las categorías nominales, y en caso de que sea una proyección máxima (un NP dominado por S) se exige además que presente el ROL de sujeto.

5.2.- Categorías Controladas y sus Rasgos de Control

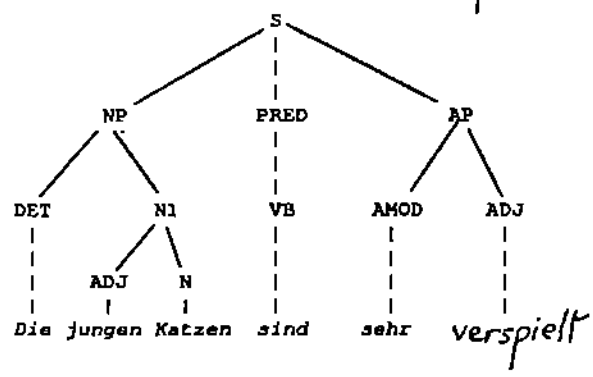
[-N,+V]	<NU,PS>	(categorías verbales)
[+N,+V]	<GD, NU>	(categorías adjetivales)
[+N,-V,+ANAPH]	<CA, GD, NU, INDEX>	(anáforas)
[DET]	<GD, NU>	(determinantes)

Son categorías controladas las categorías verbales (en número y persona), las categorías adjetivales (en género y número), las categorías nominales anafóricas (en caso, género, número e índice de coindexación con su antecedente) y los determinantes (en género y número)

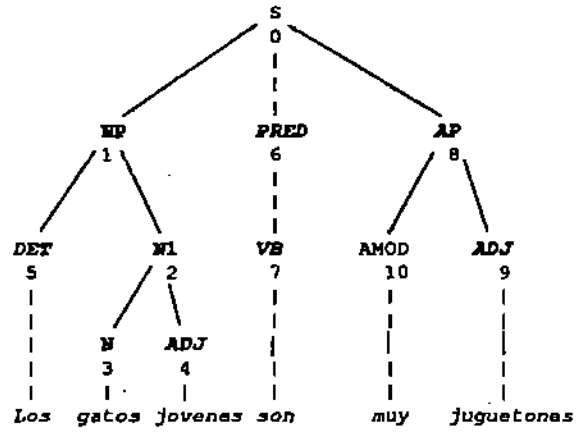
El modelo asume que un mecanismo independiente asigna previamente valores de INDEX iguales a las anáforas y a sus antecedentes.

6.- Aplicación del modelo a un caso real: METAL.

Veamos un ejemplo de la aplicación del modelo expuesto. La siguiente figura representa el árbol de análisis que hace METAL para la frase alemana "*Die jungen Katzen sind sehr verspielt*".



El árbol de transferencia correspondiente en castellano es:



En este último árbol se han marcado en negrita los nodos controladores, y en negrita cursiva los nodos controlados, según lo expuesto en el punto 5. Debajo de cada nodo hay un número que indica el orden de transferencia de los nodos.

Cabe señalar que en esta frase se producen los dos fenómenos expuestos en 3.2: "*Katze*" es femenino en alemán, mientras que su traducción, "*gato*", es masculino en castellano. Por otra parte, el adjetivo en posición predicativa "*juguetones*" debe concordar en género y núcleo con "*gatos*", cosa que no ocurre en alemán con "verspielt".

El nodo léxico "*gatos*" controla al adjetivo "*jovenes*" en género y número, su proyección N1 controla al determinante "*los*", también en género y número. La proyección de N1, NP, controla a PRED, en número y

persona, y a AP, en género y número.

La aplicación de PRC(i) asegura que el orden de transferencia sea el correcto (primero nodos controladores y después, nodos controlados) y que los rasgos de control pertinentes para cada nodo controlado se copien antes de transferirlos una vez transferido el controlador.

La aplicación de PRN asegura que los rasgos de control se suben desde los nodos léxicos hasta las cabezas de las proyecciones.

Por último, PRC(ii) asegura que los rasgos de control se bajan desde las proyecciones máximas de las categorías controladas (PRED y AP) hasta los nodos léxicos que son sus núcleos ("*son*" y "*juguetones*", respectivamente).

7.- *Bibliografía*

[Gazdar85]: GAZDAR, KLEIN, PULLUM & SAG. Generalized Phrase Structure Grammar, 1985.

[Isabelle86]: ISABELLE, MACKLOVITCH, Transfer aand MT Modularity, Proceedings del Coling '86, 1986.

[Pollard87]: POLLARD, SAG. Head-Driven Phrase Structure Grammar, Standford University,1987.

[Shieber86]: A simple reconstruction of GPSG, Proceedings del Coling '86, 1986.