Usando la co-composicionalidad para el aprendizaje de la subcategorización sintáctico-semántica *

Pablo Gamallo Otero   Alexandre Agustini   Gabriel P. Lopes
Depto. de Informática, Univ. Nova de Lisboa, Portugal
gamallo,agustini,gpl@di.fct.unl.pt

Resumen: El análisis sintáctico precisa de léxicos expandidos con información sobre subcategorización. Este artículo describe un método no supervisado capaz de adquirir restricciones sintácticas y semánticas a partir de corpora restrictos. Se dará especial atención al papel de la co-composición en el proceso de aprendizaje de este tipo de información.

Palabras clave: Extracción de información semántica, Adquisición de restricciones de selección, Corrección del análisis sintáctico, Léxico Generativo

Abstract: Natural language parsing requires extensive lexicons containing subcategorisation information for specific sublanguages. This paper describes an unsupervised method for acquiring both syntactic and semantic subcategorisation restrictions from corpora. Special attention will be paid to the role of co-composition in the acquisition strategy.

Keywords: Semantic Information Extraction, Selection Restrictions Acquisition, Parsing Improvement, Generative Lexicon

1 Introducción

Las gramáticas lexicalistas proyectan en las estructuras sintácticas los patrones de subcategorización codificados en el léxico. Estas gramáticas se sirven de léxicos con información específica sobre subcategorización a fin de restringir el análisis sintáctico. Los analizadores sintácticos necesitan esta información para reducir el número de análisis posibles y, con ello, resolver la ambigüedad estructural. En los últimos años, han ido apareciendo diferentes métodos para adquirir patrones de subcategorización a partir de corpora. Algunos de ellos tienen por objeto inducir la subcategorización sintáctica a partir de textos anotados (Brent, 1993; Briscoe y Carrol, 1997; Marques, 2000). Desgraciadamente, el recurso a la información sintáctica no es suficiente para resolver la ambigüedad estructural. Tomemos como ejemplo los dos siguientes análisis:

1. [pellar [SN la patata] [SP con un cuchillo]]
2. [pellar [SN la patata] [SP con una mancha]]

Los criterios puramente sintácticos no son suficientes para conseguir unir con_SP al verbo pelar en (1) y al SN la patata en (2). Supongamos que la dependencia entre el SP con un cuchillo y el verbo pelar es sugerida por el hecho de que este último subcategoriza expresiones de tipo con_SP. Esto no explica, sin embargo, el análisis (2), donde el SP con una mancha aparecerá ligado directamente al sintagma anterior: la patata. Para conseguir un análisis correcto, necesitamos recurrir a nuestro conocimiento del mundo, en particular, a nuestros conocimientos acerca de la acción de pelar, el uso de cuchillos y las propiedades de las patatas. En general, sabemos que los cuchillos sirven para pelar y que las patatas pueden tener manchas en la piel. De esta manera, para llegar a un análisis correcto, el analizador debe extraer del léxico, no sólo información acerca de la subcategorización sintáctica, sino también acerca de las preferencias semánticas.

Existen otros métodos que tienen como objetivo extraer restricciones de selección usando estrategias de adquisición supervisadas. El algoritmo de aprendizaje requiere que en el corpus de muestra estén anotados con etiquetas semánticas los nombres que aparecen en dependencias de tipo verbo-nombre, nombre-verbo, o verbo-prep-nombre. En algunos casos, la anotación es manual, y en otros con base en diccionarios y tesaurí construidos manualmente (Resnik, 1997; Framis, 1995).

También se están desarrollando estrate-
2.3.1.2. Herramientas de aprendizaje y enseñanza

La adquisición de las habilidades lingüísticas depende de la presencia de un conjunto de habilidades cognitivas, como la atención, la memoria y la ejecución de procedimientos. Estas habilidades se desarrollan a lo largo del desarrollo infantil y son esenciales para el aprendizaje de nuevas habilidades, como la lectura y la escritura.

La adquisición de habilidades lingüísticas puede ser influenciada por factores tanto internos como exteriores. Internos, como la genética, la edad y el género, pueden influir en el desarrollo de habilidades lingüísticas, mientras que factores externos, como la exposición a diferentes culturas y idiomas, pueden también ser relevantes.

El aprendizaje de nuevas habilidades lingüísticas es un proceso que requiere tiempo y práctica. Los individuos con una mayor exposición a diferentes idiomas y culturas suelen tener una mejor capacidad para adquirir nuevas habilidades lingüísticas.

El estudio de la adquisición de habilidades lingüísticas ha permitido identificar diferentes patrones de desarrollo, como la adquisición de habilidades básicas y la adquisición de habilidades avanzadas. Los individuos tienden a adquirir habilidades lingüísticas de manera progresiva, pasando primero por la adquisición de habilidades básicas y luego por la adquisición de habilidades avanzadas.

El aprendizaje de nuevas habilidades lingüísticas es un proceso que requiere tiempo y práctica. Los individuos con una mayor exposición a diferentes idiomas y culturas suelen tener una mejor capacidad para adquirir nuevas habilidades lingüísticas.

El estudio de la adquisición de habilidades lingüísticas ha permitido identificar diferentes patrones de desarrollo, como la adquisición de habilidades básicas y la adquisición de habilidades avanzadas. Los individuos tienden a adquirir habilidades lingüísticas de manera progresiva, pasando primero por la adquisición de habilidades básicas y luego por la adquisición de habilidades avanzadas.
Usando la co-composicionalidad para el aprendizaje de la subcategorización sintáctico-semántica

- \[ \lambda x^1(\text{de}; x^1, \text{república}^1) \] este es un contexto de subcategorización que requiere un núcleo que designe alguna organización, institución, cargo o función de la república.

Nuestra noción de subcategorización engloba por tanto no sólo preferencias sintácticas sino también semánticas.

La segunda hipótesis en la que basamos nuestro método se refiere al proceso de construcción de clases de contextos de subcategorización. Postulamos, en particular, que dos contextos de subcategorización pueden ser considerados semánticamente similares si aparecen en el corpus con la misma distribución. Tomemos, por ejemplo, los siguientes contextos:

\[
\begin{align*}
\lambda x^1(\text{de}; x^2, \text{república}^1) & \\
\lambda x^2(\text{de}; x^2, \text{estado}^1) & \\
\lambda x^3(\text{de}; x^2, \text{incumbrir}^1, x^3) &
\end{align*}
\]

Dado que todos ellos parecen compartir las mismas preferencias semánticas, tienden a tener una distribución de palabras muy similar. Las palabras que aparecen con más frecuencia en estos contextos representan la clase con la que se puede definir extensionalmente sus preferencias semánticas. Por ejemplo, las palabras ministro, presidente, asamblea, ..., que se distribuyen regularmente en estos contextos, pueden ser servir para construir la clase extensional con la que se describe sus restricciones de selección.

A diferencia de los métodos no supervisados más utilizados en la extracción de restricciones de selección, nuestro modelo no busca identificar palabras similares en base a la hipótesis distribucional de Harris. Según esta hipótesis, las palabras que ocurren en los mismos contextos de subcategorización son semánticamente similares. Tal como hemos explicado en la Introducción, esta concepción acerca de la semejanza entre palabras no parece poder tomar en cuenta el fenómeno de la polisemia. Por ello, el principal objetivo de nuestro método será el de medir la semejanza, no entre palabras, sino entre contextos de subcategorización (Faure y Nédellec, 1998; Faure, 2000).

3 Adquisición de la subcategorización

Con el fin de evaluar las dos hipótesis que acabamos de formular, fue elaborado un método de adquisición de subcategorización sintáctica y semántica a partir de un corpus parcialmente analizado. Este método se compone de dos procesos secuenciales. El primero tiene como objetivo extraer contextos posibles de subcategorización (que llamamos "contextos candidatos"), mientras que el segundo identifica, entre los contextos candidatos, aquéllos que son correctos, a fin de agruparlos en clases semánticas de subcategorización. A continuación, estos dos procesos serán definidos de una manera más precisa.

3.1 Extracción de contextos candidatos

Se etiqueta primero el texto por medio de "tags" morfosintácticos (Marques, 2000), y luego se analiza parcialmente (Rocio, de la Clergerie, y Lopes, 2001) en secuencias de "chunks" básicos (SN, SP, SV, ...). Seguidamente, por medio de heurísticas simples de tipo "asociación a la derecha", asignamos estructuras sintácticas temporales: por ejemplo, consideramos que un chunk tiene de estar relacionado sintácticamente con el chunk que se encuentra inmediatamente a su izquierda. Tomando en cuenta la primera hipótesis formulada en 2, asumimos que dos chunks ligados sintácticamente forman una dependencia binaria susceptible de dividirse en dos contextos de subcategorización. Dado que las heurísticas estructurales empleadas no toman en cuenta las dependencias a distancia, es evidente que pueden surgir varios tipos de errores sintácticos.2 Por este motivo, los contextos de subcategorización identificados en esta primera etapa se consideran sólo contextos candidatos. Finalmente, también postulamos que el conjunto de palabras asociado a cada contexto candidato representa lo que podría ser una clase semántica. Por ejemplo, la frase

\[
\text{emanou de facto da lei}
\]

(emanó de hecho de la ley)

da lugar a las dos siguientes dependencias:

\[
(\text{obj de, emanar}^1, \text{facto}^1) \ (\text{de, facto}^2, \text{les}^1)
\]
a partir de las que se pueden generar cuatro posibles contextos de subcategorización:

\[
\begin{align*}
\lambda x^1(\text{obj de, emanar}^1, x^1) & \\
\lambda x^2(\text{de, facto}^1, x^2) & \\
\lambda x^3(\text{de, facto}^2, x^2) &
\end{align*}
\]

2 Los errores pueden ser causados, no sólo por las limitaciones de estas heurísticas, sino también por las propias incorrecciones del corpus, la incompletitud del diccionario, palabras incorrectamente etiquetadas, etc.
Sin embargo, estos cuatro contextos candidatos son incorrectos. El complemento preposicional de facto representa una locución adverbial interpolada entre el verbo y su verdadero complemento, da lei. Veremos más adelante cómo se adquiere la información sobre subcategorización, y también cómo ésta se utiliza para reconocer las dependencias incorrectas. El algoritmo de aprendizaje se basa esencialmente en el cálculo de la semejanza entre los conjuntos de palabras asociados a los contextos de subcategorización candidatos.

3.2 Generar clases de contextos de subcategorización

De acuerdo con la segunda hipótesis formulada en la sección 2, dos contextos de subcategorización con una distribución de palabras semejante deben contener las mismas restricciones de selección. En torno a esta idea, vamos a desarrollar el proceso de generación de clases de subcategorización semántica. Este proceso consta de dos tareas: “filtración” y “clustering”.

3.2.1 Filtración

A cada contexto de subcategorización se le asocia un conjunto de palabras con sus respectivas frecuencias, llamado “conjunto contextual”. Nuestro objetivo es reunir conjuntos contextuales similares en conjuntos cada vez más grandes que representen extensionalmente clases de subcategorización semántica. El problema reside en que la mayor parte de los conjuntos contextuales contiene palabras que fueron mal etiquetadas o que provienen de estructuras sintácticas incorrectas. El objetivo de la filtración es, precisamente, retirar estas palabras de los conjuntos contextuales. El proceso de filtración se realiza de la siguiente manera.

En primer lugar, a cada conjunto contextual se le asocia una lista de conjuntos similares. Cuanto mayor sea el porcentaje de palabras que comparten dos conjuntos, mayor es su grado de semejanza. Entre los diferentes coeficientes estadísticos que fueron usados para extraer listas de conjuntos semejantes, los mejores resultados se consiguieron con una versión con pesos del coeficiente Jaccard. Atribuimos pesos a las palabras en función de su dispersión (peso global) y de su frecuencia relativa (peso local) en cada contexto o conjunto contextual. La dispersión de la palabra (peso global) \( disp \) toma en cuenta el número de contextos asociados a una palabra dada y su frecuencia en el corpus. El peso local se calcula tomando en cuenta la frecuencia relativa \( fr \) del par palabra/contexto. El peso \( P \) de la palabra \( pal \) con respecto al contexto \( cntx \) se calcula de esta manera:

\[
P(pal_i, cntx_j) = \log_2(fr_{ij}) \times \log_2(disp_i)
\]

donde

\[
fr_{ij} = \frac{frecuencia de pal_i con cntx_j}{\sum_i frecuencia de pal_i con cntx_j}
\]

\[
disp_i = \frac{\sum_j frecuencia de pal_i con cntx_j}{\text{numero de contextos con pal}_i}
\]

Finalmente, la medida Jaccard de semejanza \( SJ \) entre dos contextos \( m \) y \( n \) es el resultado de \( 3 \):

\[
SJ(cntx_m, cntx_n) = \frac{\sum_{i,j}(P(cntx_m, pal_i) \times P(cntx_n, pal_i))}{\sum_{i,j}(P(cntx_m, pal_i) + P(cntx_n, pal_i))}
\]

Para cada conjunto contextual (y su correspondiente contexto de subcategorización), seleccionamos una lista de contextos considerados semejantes. Por cada par de conjuntos semejantes, extraemos sólo las palabras que comparten. Es decir, seleccionamos la intersección de palabras comunes de cada par de conjuntos contextuales que fueron considerados semejantes. Cada intersección representa una clase semántica homogénea, que llamamos “clase básica”. Tomemos un ejemplo. En nuestro corpus, el contexto estadísticamente más parecido a \( [\lambda x^1 (de, violação, x^1)] \) es el contexto: \( [\lambda x^1 (dobj, violar, x^1)] \). Ambos contienen las siguientes palabras:

- sigilo
- princípios
- preceito
- plano
- norma
- lei
- estatuto
- disposto
- disposição
- direito

Esta clase básica no contiene palabras como \( vez, flagrantemente, obrigação, interesse \), que aparecen asociadas por error a uno de los dos contextos. El proceso de filtración consigue, por tanto, eliminar este tipo de errores. Esta clase básica puede ser considerada semánticamente homogénea pues las

\( ^{3}\text{comun restricte el sumatorio a las palabras comunes a } m \text{ y } n. \)
palabras componentes designan todas el mismo tipo de palabras: todas se refieren a documentos legales. Una vez que las clases básicas han sido generadas, entra en juego el algoritmo de clustering, que toma como punto de partida estas clases con el fin de producir otras más generales.

### 3.2.2 Clustering conceptual

Usamos un proceso de clustering aglomerativo (bottom-up), para ir agregando progresivamente las clases básicas en nuevas clases derivadas. Los detalles de este proceso aparecen en (Gamallo, Agustini, y Lopes, 2001). La Figura 1 ilustra dos clases básicas generadas por dos pares de contextos de subcategorización similares. \([\text{CONTX}_i]\) representa un par de contextos de subcategorización similares. Estos dos contextos semejantes tienen en común la clase básica: preceito, lei, norma (precepto, ley, norma). Por otro lado, \([\text{CONTX}_j]\) representa otro par de contextos similares, que comparten las palabras preceito, lei, direito (precepto, ley, derecho). Ambas clases básicas fueron generadas por el proceso de filtración descrito en la sección anterior. El proceso de clustering las junta en una clase más general constituida por preceito, lei, norma, direito. Al mismo tiempo, los dos pares de contextos, \([\text{CONTX}_i]\) y \([\text{CONTX}_j]\), también se reúnen en un nuevo conjunto de contextos \([\text{CONTX}_k]\). Este proceso de generalización nos permite inducir información sintáctica que no aparece explícitamente en el corpus. En particular, inducimos que la palabra norma puede aparecer en contextos de subcategorización representados por \([\text{CONTX}_j]\), y también que direito puede ocurrir en los contextos representados por \([\text{CONTX}_k]\).

En el proceso de clustering extendemos, por tanto, una clase básica (o preferencia semántica) en dos direcciones: aumentamos el número de palabras que constituyen la clase (o preferencia), al mismo tiempo que incrementamos el número de contextos que impiden esa clase (o preferencia).

#### 3.2.3 Representación de las palabras polisémicas

Las palabras polisémicas aparecen en diferentes clusters. Por ejemplo, la Tabla 1 sitúa la palabra traballo (trabajo) en tres clases diferentes. Cluster_1 contiene palabras que designan objetos temporales. Estas palabras aparecen, por tanto, en contextos de subcategorización que exigen argumentos temporales: e.g., \([\lambda x^1(\text{de};\text{suspensão})]\) y \([\lambda x^1(\text{em}; x^4, \text{curso})]\). Cluster_2 agrega palabras que describen el resultado de una acción. Este significado es relevante en contextos como: \([\lambda x^1(\text{obj}_{\text{por}};\text{receber}^4, x^1)]\). En efecto, el motivo por el que se recibe dinero no es la acción de trabajar, sino el producto realizado a través del trabajo. Finalmente, Cluster_3 ilustra el significado “tarea” o “función”. Este sentido se activa en contextos tales como: \([\lambda x^1(\text{de}; \text{inspector}^4)]\), o \([\lambda x^1(\text{obj};\text{desempenhar}^4, x^1)]\).

### 4 Aplicación y evaluación

Las clases adquiridas en la etapa de aprendizaje serán utilizadas por el analizador sintáctico. Para ello, enriquecemos primero el léxico con información sobre subcategorización sintáctico-semántica. Una vez actualizado el diccionario léxico, realizamos un segundo ciclo de análisis sintáctico, con el fin de verificar y corregir las dependencias que fueron identificadas en el primer ciclo.

#### 4.1 Actualización del léxico

La Tabla 2 muestra tres entradas léxicas enriquecidas con información sobre subcategorización sintáctico-semántica. Cada entrada contiene dos listas de información: los
Tabla 2: Dictionary entries

- **abono**
  - \([\lambda x^2(\text{de}; x^4, \text{abono}^0)] = \)
  - [aplicación caso fixação montante pagamento título]
  - \([\lambda x^2(\text{de}; \text{abono}^0, x^7)] = \)
  - [ajuda despesa pensão quantia remuneração subsidio suplemento valor vencimento]
  - \([\lambda x^2(\text{obj de}; x^4, \text{abono}^0)] = \)
  - [conceber conter definir determinar fixar manter prever]
- **emana**
  - \([\lambda x^2(\text{obj de}; \text{emana}^4, x^7)] = \)
  - [alínea artigo código decreto diploma disposição estatuto legislação lei norma regulamento]
  - \([\lambda x^2(\text{obj de}; \text{emana}^4, x^7)] = \)
  - [administración autoridade comissão conselho direcção estado governo ministro tribunal órgão]
- **presidente**
  - \([\lambda x^2(\text{de}; \text{presidente}^4, x^7)] = \)
  - [asamblea câmara comissão conselho direcção estado empresa gestão instituto região republica secção tribunal órgão]
  - \([\lambda x^2(\text{de}; x^4, \text{presidente}^0)] = \)
  - [cargo categoria função lugar remuneração vencimento]
los que (do\textsuperscript{b}y; \textit{editar\textsuperscript{a}}, \textit{artigo\textsuperscript{a}}) es una relación binaria. La dependencia es confirmada. En caso contrario, la dependencia no puede ser confirmada.

La verificación semántica se basa en la hipótesis sobre la co-composición, introducida anteriormente (sección 2). Según esta hipótesis, dos palabras son sintácticamente dependientes sólo si una de estas dos condiciones se confirma: el núcleo requiere semánticamente al modificador, o el modificador requiere semánticamente al núcleo. Debido a las propiedades de la co-composicionalidad, basta apenas que se realice una de estas dos condiciones para poder afirmar que dos palabras son sintácticamente dependientes.

4.3 Evaluación de del procedimiento de resolución de dependencias

La Tabla 3 ilustra la evaluación manual de las correcciones propuestas por el sistema de diagnóstico. Evaluamos la precisión y la cobertura de las correcciones propuestas en torno a tres tipos de dependencias candidatas: \textit{prep}(N, V), \textit{dobj}(V, N), y \textit{iobj}_{\textit{prep}}(V, N). Llamamos precisión al porcentaje de verificaciones correctas con respecto al número de correcciones efectivamente realizadas. La cobertura indica la proporción de dependencias candidatas que fueron efectivamente verificadas.

<table>
<thead>
<tr>
<th>Dep. Candidata</th>
<th>Precisión (%)</th>
<th>Cobertura (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{prep}(N, V)</td>
<td>95.53</td>
<td>30.27</td>
</tr>
<tr>
<td>\textit{dobj}(V, N)</td>
<td>94.44</td>
<td>19.44</td>
</tr>
<tr>
<td>\textit{iobj}_{\textit{prep}}(V, N)</td>
<td>93.87</td>
<td>10.11</td>
</tr>
<tr>
<td>Total</td>
<td>94.61</td>
<td>19.94</td>
</tr>
</tbody>
</table>

Mientras que la precisión alcanza valores elevados (sobre 95%), la cobertura apenas logra el 20%. El motivo de esta baja cobertura se debe a que la información sobre subcategorización sólo puede ser adquirida cuando las palabras aparecen con cierta frecuencia en el corpus. Por consiguiente, la cobertura podrá ser incrementada a medida que la talla del corpus aumente.

5 Trabajo futuro

Dado que el método que hemos presentado todavía no ofrece propuestas para las dependencias a distancia, no es posible compararlo con los trabajos ya existentes que toman en cuenta este tipo de dependencias (Hindle y Rooth, 1993; Brill y Resnik, 1994). En nuestro enfoque, la resolución de dependencias a distancia sólo será posible una vez que se agoten las correcciones sobre las dependencias inmediatas. Estamos actualmente trabajando en la elaboración de nuevos ciclos de análisis centrados en la verificación y corrección de dependencias lejanas. Tomemos de nuevo la frase \textit{emanou de facto da lei}. Como ya fue dicho anteriormente, el sistema de diagnóstico confirmó que el \textit{SP} de \textit{facto} no puede ser ligado al verbo \textit{emanou}. En un nuevo ciclo de análisis, el sistema deberá verificar si el segundo \textit{SP} da lei puede ser directamente ligado al verbo. El sistema ejecutará n-ciclos de análisis, hasta agotar el número de candidatos posibles. Al final del proceso, el parser deberá proponer aquellas dependencias que mejor se adapten a la información disponible en el léxico.

Bibliografía


Briscoe, Ted y John Carrol. 1997. Automatic extraction of subcategorization from corpora. En 5\textsuperscript{th} Conference on Applied Natural Language Processing (ANLP97), Washington, DC, USA.


Faure, David y Claire Nédellec. 1998. Asium: Learning subcategorization frames and
restrictions of selection. En ECML98, Workshop on Text Mining.


Sekine, Satoshi, Jeremy Carrol, Sofia Ana-