Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 35-42 recibido 21-12-11 revisado 10-03-12 aceptado 18-03-12

Inducción Gramatical Semisupervisada usando Información de Análisis Superficial *

Semisupervised Grammar Induction based on Text Chunking Information

Lourdes Araujo, Jesús Santamaría
UNED, NLP-IR Group
lurdes@lsi.uned.es, jsant@lsi.uned.es

Resumen: El análisis sintáctico de los textos es un proceso fundamental en el procesamiento del lenguaje natural que requiere disponer de la gramática correspondiente a la lengua considerada. La gramática puede obtenerse de un corpus anotado sintácticamente, pero tales corpora no existen para muchas lenguas. Esta razón ha provocado un interés creciente en los métodos no supervisados de inducción gramatical, que no requieren dichos corpora. Sin embargo, los resultados de estos métodos son menos precisos. Por este motivo nosotros hemos recurrido a información adicional menos costosa de obtener. Concretamente, en este trabajo estudiamos la forma de introducir el análisis sintáctico superficial para mejorar los resultados de la inducción gramatical no supervisada de un sistema basado en patrones léxicos. El análisis superficial o *chunking* identifica a los constituyentes de la oración, sin especificar su estructura interna. Los resultados han mostrado una mejora apreciable de los resultados a medida que se añaden distintos tipos de constituyentes.
Palabras clave: Inducción Gramatical, Métodos semisupervisados, Análisis superficial

Abstract: Syntactic analysis of texts requires the availability of the grammar underlying the language. The grammar can be obtained from corpora syntactically annotated, but such corpora do not exist for many languages. This has led to a growing interest in unsupervised grammar induction, which does not require such annotations, but provides less accurate results. Aiming at improving the accuracy of this kind of approach, we have resorted to additional information, which can be obtained more easily. Shallow parsing or chunking identifies the sentence constituents, but without specifying their internal structure. In this work we have investigated how the results of a pattern-based unsupervised grammar induction system improve as data on new kind of phrase are added.
Keywords: Grammar induction, Semisupervised Methods, Chunking

1. Introducción

Los analizadores sintácticos, que requieren la gramática subyacente a la lengua considerada, son un recurso necesario para muchos procesos como la extracción de información, la traducción automática, etc. La inducción gramatical (IG) es el proceso de extracción de una gramática a partir de una colección de textos. Los trabajos realizados en esta línea pueden clasificarse en tres grupos en función del grado de supervisión que requieren. Los sistemas de IG supervisada (IGS) requieren ejemplos anotados sintácticamente que no están disponibles en muchas lenguas. Los sistemas no supervisados (IGNS) sólo requieren textos planos[1], pero su rendimiento es menor que el de los sistemas supervisados. Finalmente, existe un enfoque intermedio, el semisupervisado (IGSS), que requiere cierto grado de supervisión, pero no textos con anotaciones de análisis sintáctico completo.

Se han desarrollado diversos trabajos dentro de los tres tipos de IG. Trabajos recientes en IGS (Petrov, 2010) han reportado resultados que se encuentran entre el 90 % y el 93 % cuando se aplican a la sección Wall Street Journal del Penn Treebank corpus (Marcus, Santorini, y Marcinkiewicz, 1994).

Algunas de las propuestas más recientes de IGNS han sido hechas por Klein y Manning (2005), Bod (2006) y Santamaría y Araujo

* Financiado por el proyecto Holopedia (TIN2010-21128-C02), y por el proyectos de la Comunidad Autónoma de Madrid MA2VICMR (S2009/TIC-1542).

[1] En la literatura se consideran métodos no supervisados aquellos que requieren textos etiquetados con categorías léxicas, siempre que no requieran anotaciones sintácticas.

ISSN 1135-5948

(2010). En el primer trabajo se usa el algoritmo de Expectation-Maximization (EM) para seleccionar el árbol más probable de una oración. Este trabajo alcanza una medida-F de $71{,}1\,\%$ para el corpus WSJ10. Este corpus, que consta de 7422 oraciones, está formado por oraciones de hasta 10 palabras de la sección Wall Street Journal del Penn Treebank. El sistema propuesto por Bod (2006) genera todos los árboles binarios para cada oración y elige aquel que puede obtenerse con el número menor de sustituciones de subárboles en los nodos del árbol mayor generado. Este sistema alcanza una medida-F de $82{,}9\,\%$ para el corpus WSJ10, pero genera una enorme cantidad de subárboles para cada oración lo que hace que el sistema no sea aplicable a oraciones largas. El sistema de IGNS de Santamaría y Araujo (2010) está basado en la detección estadística de determinados patrones de etiquetas léxicas que aparecen en las oraciones. Este sistema, que a diferencia de otros, considera cualquier tipo de árbol, no sólo los binarios, alcanza una medida-F del 80.50 % para el WSJ10 corpus incluyendo la oración como sintagma, como hacen el resto de los sistemas y de 75.82 % cuando no se incluye este sintagma, que siempre es correcto.

Finalmente, también hay un interés creciente en el enfoque semisupervisado (IGSS), aunque la mayor parte de los trabajos se han desarrollado para análisis de dependencias. Por ejemplo, Druck, Mann, y McCallum (2009) han propuesto un sistema para extraer gramáticas de dependencias usando conocimiento lingüístico a priori. Haghighi y Klein (2006) han propuesto un modelo para incorporar conocimiento en el sistema no supervisado de Klein y Manning (2005). En este caso, el conocimiento a priori se especifica de forma declarativa, introduciendo una serie de ejemplos canónicos de cada tipo de sintagma considerado.

El enfoque supervisado requiere corpus anotados con el análisis sintáctico completo, que no están disponibles en muchas lenguas, o para determinados tipos de textos. Por este motivo es interesante considerar el enfoque no supervisado. Sin embargo, a pesar de las paulatinas mejoras en el rendimiento de este tipo de sistemas, aún hay una diferencia importante entre su rendimiento y el de los sistemas supervisados. Para salvar esta diferencia, en este trabajo se propone introducir un pequeño grado de supervisión que no requiera el análisis completo de las oraciones, sino sólo un *análisis superficial*. En este tipo de análisis se identifican los sintagmas de la oración pero sin identificar su estructura interna.

En los experimentos realizados en este trabajo hemos utilizado directamente los sintagmas anotados en un corpus, para evitar la dependencia de los resultados de un analizador superficial específico. Sin embargo, se han obtenido resultados superiores al 95 % (Sha y Pereira, 2003), para algunos analizadores superficiales, aunque no están integrados en un sistema de libre acceso. Por ello esperamos que los resultados sólo se vieran ligeramente afectados si en lugar de anotaciones se utilizara un analizador superficial de alto rendimiento.

Hemos analizado el efecto de introducir distintos tipos de sintagmas: nominales (SN), verbales (SV) y preposicionales (SP), separada y conjuntamente. También se ha estudiado el efecto de introducir diferentes grados de supervisión. El grado de supervisión es la tasa de sintagmas anotados previamente respecto al número total de sintagmas. Se ha elegido como sistema no supervisado de base el propuesto por Santamaría y Araujo (2010) ya que los patrones que se extraen en este modelo pueden combinarse fácilmente con otros correspondientes a la supervisión introducida, porque no se restringe a árboles binarios y proporciona resultados competitivos.

El resto del artículo se organiza de la siguiente forma: La sección 2 describe el sistema de IGNS de partida. La sección 3 describe la forma en que se introduce la información de análisis superficial en el sistema. La sección 4 presenta y analiza los resultados, y finalmente la sección 5 resume las conclusiones del trabajo.

2. Inducción Gramatical no Supervisada basada en Patrones

Santamaría y Araujo (2010) han desarrollado un sistema de inducción no supervisada basado en la idea de que las categorías léxicas tienden a desempeñar determinados papeles en la estructura del árbol sintáctico. De acuerdo con esto, las categorías léxicas se clasifican en un pequeño número de clases, cuyos componentes tienden a aparecer en determinadas posiciones del árbol de análisis. Este comportamiento no es específico de una determinada lengua, sino que se observa en muchas de ellas. El conjunto de etiquetas léxicas se clasifica en una serie de clases de etiquetas en función de ciertas heurísticas basadas en las frecuencias de aparición de dichas etiquetas. Después, esta clasificación se utiliza para analizar cualquier oración mediante un procedimiento determinista que selecciona los constituyentes y los niveles en que aparecen en el árbol de análisis en función de las clases de etiquetas.

Consideremos algunos ejemplos de árboles de análisis para ilustrar las clases de etiquetas léxicas que se consideran en el modelo. Al examinar el ejemplo que aparece en la figura 1 podemos ver que algunas secuencias de etiquetas léxicas o constituyentes aparecen en el nivel más bajo del árbol (están exclusivamente compuestas de etiquetas léxicas, sin etiquetas sintácticas que den lugar a otros constituyentes). Este es el caso de (DT, NNP) y (DT, NN), donde DT es un determinante, NNP un nombre propio singular, y NN un nombre singular. Otras etiquetas léxicas como VBZ (verbo presente de singular, 3^a persona) y TO dan lugar a un nuevo nivel en el árbol de análisis ya que después de ellas aparece un nuevo subárbol. El modelo intenta capturar estas ideas, identificando las clases de etiquetas que corresponden a un patrón estructural determinado.

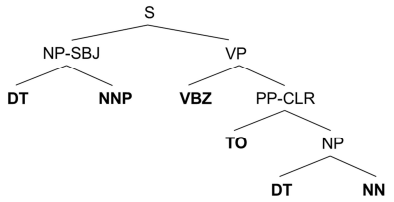

Figura 1: Árbol de análisis de la oración *The U.S.S.R. belongs to neither organization* procedente del Penn Treebank.

El procedimiento para identificar las distintas clases de etiquetas léxicas se basa en la detección del *constituyente seguro* (CS):

CS es la secuencia más frecuente de dos etiquetas léxicas en el corpus.

La hipótesis es que al menos para el constituyente más frecuente el número de apariciones supere con seguridad al número de secuencias que aparecen por azar. En el corpus WSJ10, CS es DT NN*, donde NN* representa a cualquier tipo de nombre (NN, NNS, NNP, NNPS).

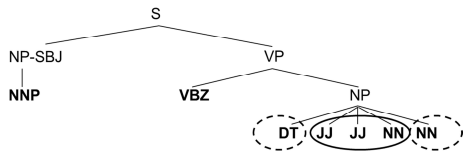

Figura 2: Árbol de análisis de la oración *Telerate provides an electronic financial information network* procedente de Penn Treebank.

El modelo considera las siguientes clases de etiquetas léxicas, cuya identificación se basa en el comportamiento de las etiquetas respecto al CS:

- *Extensores*: son etiquetas léxicas con una tendencia a aparecer entre las etiquetas que definen al CS. Este es el caso de la etiqueta JJ (adjetivo) que aparece entre DT y NN en la figura 2.
- *Separadores*: son etiquetas léxicas con una tendencia estadística a aparecer fuera del CS. Este es el caso de VBZ en el árbol de análisis de las figuras 1 y 2, y de TO en la figura 1.
- *Subseparadores*: son etiquetas léxicas que no tienen una tendencia definida a aparecer dentro o fuera del CS. Generalmente aparecen delimitando constituyentes, pero sin dar lugar a un nuevo nivel en el árbol de análisis como hacen los separadores. Este es el caso de la etiqueta POS (terminación de posesivo) que aparece en la figura 3.

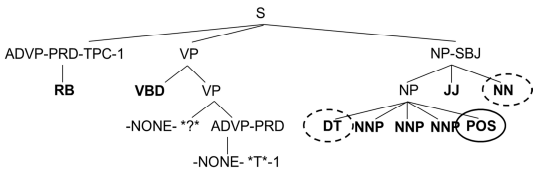

Figura 3: Árbol de análisis de la oración *So did the Federal Reserve Board's industrial-production index* procedente de Penn Treebank.

Vamos a introducir alguna notación para especificar como se identifican los extensores, separadores y subseparadores. Sea #(E1, ..., En) el número de apariciones de la secuencia de etiquetas (E1, ..., En). El *lado determinante* de una etiqueta E se define como la etiqueta izquierda (CSi) o la derecha (CSd) del constituyente seguro, con la mayor diferencia en el número de apariciones de E a sus dos lados. Por ejemplo, si la diferencia de E inmediatamente a la derecha y a la izquierda de DT (etiqueta izquierda del CS) es mayor que la diferencia de apariciones de E a ambos lados de NN (etiqueta derecha de CS), entonces CSi es el lado determinante de E y se denota *ds(E)*. Introducimos también el predicado *sim* para denotar la similitud entre el número de apariciones de una secuencia de dos etiquetas y las de la secuencia en orden inverso, es decir:

$$sim(E_1, E_2) = \text{ true iff}$$
$$\#(E\_1, E\_2)/\#(E\_2, E\_1) \in [3/4, 4/3]$$

Entonces, una etiqueta E se considera un separador si el lado determinante es el izquierdo,

Separadores	MD, PRP, IN, RB, RBR, CC, TO, VB, VBD, VBN, VBZ, VBP, VBG, EX, LS, RP, UH, WP, WRB, WDT
Sub-separadores	DT, PDT, POS, SYM, NN, NNS, NNP, NNPS

Cuadro 1: Separadores y Sub-separadores obtenidos para el corpus WSJ10.

CSi, y el número de apariciones de E a ambos lados de CSi no es similar, apareciendo E con más frecuencia a la izquierda de CSi (fuera de CS), o si el lado determinante es el derecho, CSd, y el número de apariciones de E a ambos lados de CSd no es similar, apareciendo E con más frecuencia a la derecha de CSd (fuera de CS de nuevo).

Una etiqueta se considera un subseparador si el lado determinante es el izquierdo, CSi, y el número de apariciones de E a ambos lados de CSi es similar, o si el lado determinante es el derecho, CSd, y el número de apariciones de E a ambos lados de CSd es similar.

En los casos restantes se considera que la preferencia de la etiqueta E es a estar dentro de los constituyentes, formando parte de ellos, y por tanto se clasifica como extensor.

El cuadro 1 muestra el conjunto de separadores y subseparadores obtenidos aplicando el procedimiento descrito al corpus WSJ10.

Los sub-separadores pueden formar grupos hacia su izquierda o hacia su derecha. La preferencia de cada uno de ellos por una u otra dirección se determina comparando el número de apariciones de la secuencia más frecuente formada por el sub-separador y una etiqueta léxica a su izquierda y por el sub-separador y una etiqueta léxica a su derecha. Después se toma como dirección preferente la correspondiente a la secuencia más frecuente. De acuerdo con este criterio DT, PDT, SYM tienden a aparecer a la izquierda del sintagma, mientras POS, NN, NNS, NNP y NNPS lo hacen a la derecha.

La clasificación de las etiquetas en separadores, sub-separadores y extensores permite definir un procedimiento de análisis determinista basado en el comportamiento esperado de cada etiqueta léxica de la oración. Este procedimiento sigue los siguientes pasos:

- Se identifican los separadores. Para la oración del ejemplo *The(DT) computers(NNS) were(VBD) crude(JJ) by(IN) today(NN) 's(POS) standards(NNS)* los separadores aparecen en negrita:

  [DT NNS **VBD** JJ **IN** NN POS NNS]

- Se divide la oración de acuerdo a los separadores. El primer separador que es un verbo, si hay alguno, se utiliza para dividir la oración en dos partes.

  [[DT NNS] [**VBD** [JJ [**IN** [NN POS NNS]]]]]

  Cada separador se utiliza para generar dos sintagmas: uno compuesto de la secuencia de etiquetas entre el separador y el siguiente separador, y otro que incluye al separador y la secuencia de etiquetas hasta el final de la parte de la oración en que se encuentra.

- Se identifican los sub-separadores, que aparecen subrayados en la oración:

  [[DT NNS] [**VBD** [JJ [**IN** [NN POS NNS]]]]]

- Se divide la oración de acuerdo a los subseparadores formando grupos de acuerdo a su dirección de agrupamiento preferente.

  [[DT NNS] [**VBD** [JJ [**IN** [[NN POS] NNS]]]]]

En este punto se ha obtenido el árbol de análisis de la oración.

3. Introducción de Información de Análisis Superficial

El análisis superficial o *chunking* divide a las oraciones en segmentos sin solapamientos. Los sistemas de análisis superficial asignan a las oraciones una estructura más plana que los de análisis completos. Generalmente identifican los constituyentes para una determinada profundidad del árbol.

La mayor parte de los analizadores superficiales se basan en modelos estadísticos (Church, 1988; Sang, 2002; Sha y Pereira, 2003) que se aplican para obtener un autómata finito (Pla, Molina, y Prieto, 2000; Araujo y Serrano, 2008) o un conjunto de reglas (Bourigault, 1992; Voutilainen, 1993; Ramshaw y Marcus, 1995) que permiten identificar a los constituyentes. En muchos casos se centran en un tipo particular de sintagma, los nominales, aunque las técnicas son aplicables también al resto. Algunos de estos sistemas han alcanzado valores tanto de precisión como de cobertura que están por encima del 95 %.

Estas consideraciones nos han llevado a diseñar un método de análisis que aprovecha la información de análisis superficial para mejorar los análisis proporcionados por un sistema de inducción gramatical no supervisada. En el sistema semisupervisado (IGSS) al que da lugar esta combinación de información se supone que los sintagmas de un determinado tipo y a una determinada profundidad del árbol de análisis se han identificado previamente. La idea del sistema semisupervisado es aplicar los sintagmas identificados como restricciones al proceso de generación del análisis completo que realiza el sistema no supervisado. El sistema de IGNS se aplica para generar el árbol sintáctico correspondiente a cada uno de los sintagmas identificados. Uno de estos subárboles es el de la posición superior del árbol de análisis y se obtiene sustituyendo los sintagmas identificados por etiquetas especiales que representan a alguna de las clases de etiquetas léxicas con las que trabaja el sistema no supervisado. El proceso de análisis semisupervisado sigue los siguientes pasos:

1. Reemplazar en la oración los sintagmas del tipo (nominales, verbales, etc) y nivel que se esté considerando, por etiquetas léxicas especiales que representan a las clases de etiquetas léxicas que considera el sistema de IGNS: separadores, subseparadores, etc.

2. Aplicar el sistema de IGNS para generar el árbol de la oración resultante del paso anterior.

3. Aplicar el sistema de IGNS para generar el árbol correspondiente a los sintagmas identificados previamente.

4. Reemplazar las etiquetas léxicas especiales del árbol generado en el paso 2, por los correspondientes subárboles generados en el paso 3.

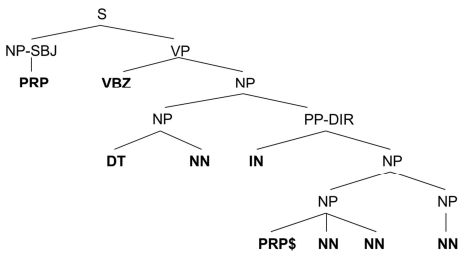

Figura 4: Árbol de análisis de la oración *It has no bearing on our work force today* procedente de la parte Wall Street Journal del Penn Treebank.

El resultado de este proceso depende de los tipos de sintagmas identificados previamente. Veamos un ejemplo de aplicación del procedimiento que se ha descrito. Supongamos que se están considerando únicamente sintagmas nominales. Entonces este tipo de sintagma se sustituye por una etiqueta especial que se corresponde con una de las clases de etiquetas léxicas consideradas. En el caso de los sintagmas nominales hemos elegido la clase de los subseparadores puesto que el núcleo de este tipo de sintagmas, los nombres (NN, NNS, NNP and NNPS), se han identificado como subseparadores por el sistema de IGNS. En consecuencia, hemos definido la etiqueta léxica especial SUBS para sustituir a los sintagmas nominales. Apliquemos el procedimiento de análisis que se ha descrito a la oración que aparece en la figura 4, considerando sólo los sintagmas nominales del nivel superior del árbol:

1. Se identifican los sintagmas nominales del nivel superior del árbol: SN = [DT NN IN PRP$ NN NN NN].

2. Los sintagmas nominales identificados en el paso previo se reemplazan por SUBS, dando lugar al árbol que aparece en la figura 5.

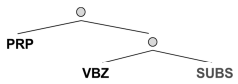

Figura 5: Árbol generado por el sistema de IG no supervisada para la secuencia de etiquetas léxicas obtenida después de reemplazar por la etiqueta léxica especial SUBS los sintagmas nominales del nivel superior en el árbol de la figura 4.

3. El sistema no supervisado también se aplica para generar el árbol de análisis del SN reemplazado. El resultado se muestra en la figura 6.

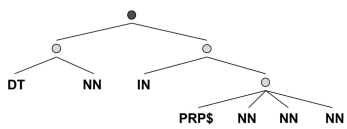

Figura 6: Árbol generado por el sistema no supervisado para la secuencia de etiquetas léxicas correspondiente al SN reemplazado en la figura 4.

4. Finalmente, la etiqueta especial SUBS se sustituye por el árbol de análisis encontrado para el constituyente en el paso anterior, obteniendo el árbol final de la figura 7.

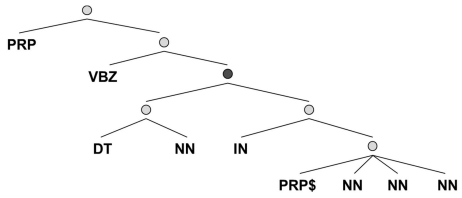

Figura 7: Árbol final obtenido por el sistema semisupervisado para la oración de la figura 4.

Para otros tipos de sintagmas el proceso es análogo. En el caso de los sintagmas verbales y preposicionales la etiqueta léxica especial que se utiliza para la sustitución es SEP y se trata como si fuera un separador. La razón es que el núcleo de los sintagmas verbales, el verbo (VB, VBD, VBN, VBZ, etc.), pertenece a la clase de los separadores, como puede observarse en el cuadro 1. Podemos observar también que la preposición (IN), el núcleo de los sintagmas preposicionales, también pertenece a la clase de los separadores.

4. Resultados

En nuestros experimentos hemos usado el corpus WSJ10, que ha sido el utilizado por el sistema de base no supervisado, lo que nos permite compararnos. Aunque el corpus está anotado sintácticamente, nosotros sólo hemos utilizado esta información con fines de evaluación, y para obtener los sintagmas que se supone que proporciona un analizador superficial, lo que nos permite no depender del rendimiento de un analizador particular. En la evaluación de los análisis obtenidos hemos utilizado las medidas más comunes en análisis sintáctico e inducción gramatical: cobertura, precisión y su media armónica, la medida-F.

La *precisión* viene dada por el número de paréntesis en el análisis a evaluar que se corresponden con los del árbol correcto y la *cobertura* mide el número de paréntesis del árbol correcto que están en el evaluado. Estas medidas tienen equivalentes para árboles cuyos nodos internos no están etiquetados, que son los considerados en este trabajo y en los sistema de IGNS que son la referencia de este trabajo. En estas medidas no se tienen en cuenta las etiquetas asignadas a los sintagmas. Los sintagmas que no puede ser erróneos (los de tamaño uno y los correspondientes a la oración completa) no se han incluido en las medidas.

Las definiciones de precisión (PSE) y cobertura (CSE) sin etiquetas de un corpus propuesto $P = [P_i]$ respecto a un corpus de referencia $G = [G_i]$ son:

$$PSE(P,G) = \frac{\sum_i |\text{paréntesis}(P_i) \cap \text{paréntesis}(G_i)|}{\sum_i |\text{paréntesis}(P_i)|},$$

$$CSE(P,G) = \frac{\sum_i |\text{paréntesis}(P_i) \cap \text{paréntesis}(G_i)|}{\sum_i |\text{paréntesis}(G_i)|}.$$

Finalmente, la medida-F sin etiquetas, FSE viene dada por la media armónica entre las dos medidas anteriores.

El sistema no supervisado subyacente alcanza una FSE de $75{,}82\,\%$ para el corpus considerado. En la evaluación del sistema este ha sido nuestro valor de referencia.

Hemos estudiado el efecto de usar la información del análisis superficial para distintos tipos de sintagmas separadamente, antes de evaluar el resultado de considerarlos simultáneamente. Para cada uno se ellos también se han investigado distintos niveles de supervisión.

4.1. El Efecto de los Sintagmas Nominales

En este caso la entrada son las secuencias de etiquetas léxicas correspondientes a los sintagmas nominales de cada oración, en diferentes niveles del árbol de análisis. Los sintagmas identificados previamente se sustituyen por la etiqueta SUBS, que es un sub-separador.

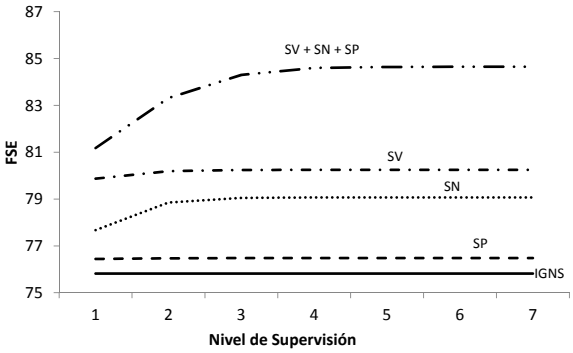

Figura 8: Resultados de FSE por nivel de supervisión usando la anotación de distintos tipos de sintagmas: nominales (SN), verbales(SV), preposicionales (SP) y los tres simultáneamente (SV+SN+SP). IGNS corresponde al valor de referencia dado por el sistema no supervisado.

La gráfica etiquetada SN en la Figura 8 muestra la FSE alcanzada usando la anotación de sintagmas nominales.

Muchos árboles de análisis no presentan algún tipo de sintagma en determinados niveles del árbol. Por ello, en esta fase que considera cada tipo de sintagma por separado, los niveles de supervisión se refieren al nivel del árbol en el que

el tipo de sintagma que se está considerando aparece. Por ejemplo, si los SNs sólo aparecen en los niveles 1 y 3 del árbol, el primer nivel de supervisión es el 1 del árbol, y el segundo es el 3 del árbol.

La supervisión introducida por los sintagmas nominales mejora apreciablemente los resultados del sistema no supervisado (IGNS). Incluso para el nivel de supervisión más bajo, es decir, considerando sólo los sintagmas correspondientes al primer nivel en que aparezca alguno (que se corresponde con un nivel de supervisión del 17,55 %) obtenemos una mejora del 2,43 % sobre el resultado de referencia del sistema no supervisado.

4.2. El Efecto de los Sintagmas Verbales

La gráfica etiquetada SV en la Figura 8 muestra la FSE alcanzada usando la anotación de sintagmas verbales. En este caso la mejora es mayor que en el de los SNs. La razón puede estar en que los SVs son más difíciles de detectar de forma no supervisada, ya que pueden tener un número variable de complementos y adjuntos. Por ello la información del análisis superficial tendría un efecto mayor. Para el menor nivel de supervisión (21,81 %) se obtiene una mejora del 5,34 % sobre el resultado de referencia del IGNS.

4.3. El Efecto de los Sintagmas Preposicionales

Finalmente consideramos que los SPs de determinados niveles del árbol estén anotados previamente. La gráfica etiquetada SP en la Figura 8 muestra los resultados para este caso. El porcentaje de mejora en este caso es más bajo, así como el grado de supervisión de los distintos niveles, lo que indica que los SPs aparecen menos. Para el menor nivel de supervisión (5,51 %) se obtiene una mejora del 0,81 % sobre el resultado de referencia del IGNS.

4.4. Efecto combinado

Podemos considerar distintos tipos de sintagmas anotados simultáneamente. La gráfica etiquetada SV+SN+SP en la Figura 8 muestra la FSE alcanzada usando la anotación de los tres tipos de sintagmas por nivel de supervisión. En este caso los niveles de supervisión se corresponden exactamente con los niveles del árbol de análisis, ya que en todos los niveles hay algún tipo de sintagma de uno u otro tipo. El efecto combinado de los tres tipos de sintagmas, SN, SVs y SPs, mejora el efecto obtenido para cada uno de ellos por separado. Los resultados más relevantes son los que se obtienen para niveles bajos de supervisión, que pueden obtenerse con un sistema de análisis superficial. Podemos ver que anotando sólo los sintagmas de los tres tipos considerados en el primer nivel del árbol (26,40 % de supervisión) se obtiene una mejora del 7,06 %.

Nivel superv.	CSE.	PSE	FSE
1(26,40 %)	84.16	78.40	81,18(7,06 %)
2(47,93 %)	87.42	79.56	83,31(9,87 %)
3(58,51 %)	88.90	80.15	84,3(11,18 %)
4(62,70 %)	89.25	80.38	84,6(11,58 %)
5(63,76 %)	89.34	80.41	84,64(11,63 %)
6(63,91 %)	89.36	80.42	84,65(11,64 %)
7(63,93 %)	89.36	80.42	84,65(11,64 %)

Cuadro 2: Cobertura, precisión y medida-F sin etiquetas del sistema semisupervisado para los tres tipos de sintagmas.

El cuadro 2 muestra los resultados numéricos obtenidos por nivel de supervisión cuando se consideran conjuntamente los tres tipos de sintagmas. El número entre paréntesis al lado del nivel indica el grado de supervisión asociado al nivel. Podemos observar que los resultados de cobertura y precisión están bastante balanceados. El número entre paréntesis junto a la medida-F representa la mejora obtenida respecto al valor de referencia de 75.82 del sistema no supervisado.

5. Conclusiones

Hemos propuesto un método semisupervisado para la inducción gramatical. El modelo propuesto introduce un pequeño grado de supervisión en un modelo de IG no supervisada basado en patrones de etiquetas léxicas. La supervisión que se añade corresponde al análisis superficial de ciertos sintagmas, y se introduce de forma muy natural en el modelo. Los resultados muestran que incluso para los niveles más bajos de supervisión, es decir contando sólo con la anotación de los sintagmas del primer nivel del árbol, se obtiene una mejora relevante para todos los tipos de sintagmas, especialmente los verbales. Por lo tanto la propuesta permite mejorar el rendimiento del sistema no supervisado sin utilizar textos analizados sintácticamente, ya que los analizadores superficiales no requieren este tipo de anotaciones para su entrenamiento.

Bibliografía

Araujo, Lourdes y Jose Ignacio Serrano. 2008. Highly accurate error-driven method for noun phrase detection. *Pattern Recognition Letters*, 29(4):547–557.

Bod, Rens. 2006. Unsupervised parsing with u-dop. En *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*, páginas 85–92, Morristown, NJ, USA. Association for Computational Linguistics.

Bourigault, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. En *Proc. of the Int. Conf. on Computational Linguistics (COLING-92)*, páginas 977–981.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. En *Proc. of 1st Conference on Applied Natural Language Processing, ANLP*, páginas 136–143.

Druck, Gregory, Gideon Mann, y Andrew McCallum. 2009. Semi-supervised learning of dependency parsers using generalized expectation criteria. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, páginas 360–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Haghighi, Aria y Dan Klein. 2006. Prototype-driven grammar induction. En *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, páginas 881–888. Association for Computational Linguistics.

Klein, Dan y Christopher D. Manning. 2005. Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, 38(9):1407–1419.

Marcus, Mitchell P., Beatrice Santorini, y Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Petrov, Slav. 2010. Products of random latent variable grammars. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, páginas 19–27.

Pla, F., A. Molina, y N. Prieto. 2000. Tagging and chunking with bigrams. En *Proc. of the 17th conference on Computational linguistics*, páginas 614–620.

Ramshaw, Lance A. y Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040. informal publication.

Sang, E. F. T. K. 2002. Memory-Based Shallow Parsing. *ArXiv Computer Science e-prints*, Abril.

Santamaría, Jesús y Lourdes Araujo. 2010. Identifying patterns for unsupervised grammar induction. En *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, páginas 38–45, Uppsala, Sweden, July. Association for Computational Linguistics.

Sha, Fei y Fernando Pereira. 2003. Shallow parsing with conditional random fields. En *Proceedings of HLT-NAACL 2003*, páginas 213–220.

Voutilainen, A. 1993. Nptool, a detector of english noun phrases. En *Proc. of the Worshop on Very Large Corpora (ACL)*, páginas 48–57.