YUNIOR RAMÍREZ CRUZ*
HENRY ANAYA SÁNCHEZ*
REYNALDO J. GIL GARCÍA*
YAMILA COBOS CASTILLO**
Centro de Estudios de Reconocimiento de Patrones y Minería de Datos*
Santiago de Cuba, Cuba
Diseño de Aplicaciones, Tecnologías y Sistemas**
Ciudad de La Habana, Cuba
{yuniorrc, henry, gil}@csd.uo.edu.cu
yamila.cobos@datys.co.cu

Un enfoque híbrido al Reconocimiento de Nombres de Entidades para el español

Resumen

En este trabajo se presenta una aproximación preliminar al Reconocimiento de Nombres de Entidades para el idioma español que se basa en la combinación de formalismos del aprendizaje automático con un conjunto de heurísticas basadas en expresiones regulares. El aprendizaje es llevado a cabo sobre un corpus etiquetado manualmente. Los resultados obtenidos permiten corroborar que el desempeño de los formalismos de aprendizaje automático mejora con la aplicación de las heurísticas.

1 Introducción

El *Reconocimiento de Nombres de Entidades* (RNE) es una tarea del procesamiento del lenguaje natural. Ésta permite, dado un texto, identificar y clasificar las expresiones que aparecen en el mismo que se refieren a personas, lugares, organizaciones, fechas, horas, cantidades, etc. A pesar de la naturaleza intermedia del RNE, su importancia es crucial en las aplicaciones del procesamiento de textos tales como la Traducción Automática, la Extracción de Información, la Construcción Automática de Resúmenes, etc.

Los dos enfoques principales que se han dado a la resolución del problema de RNE se diferencian en la forma en la que se obtiene el conocimiento que utilizan para realizar el reconocimiento y la clasificación de los nombres de entidades. El primero de los enfoques está dado por aquellos sistemas que se valen del conocimiento aportado por especialistas, el cual generalmente se representa mediante un conjunto de reglas y heurísticas que guían el proceso de etiquetado de los textos. El segundo tipo de sistemas lo componen aquellos que utilizan formalismos del aprendizaje automático para construir un modelo que codifique las propiedades de las frases que constituyen nombres de entidades. Estos enfoques no son mutuamente excluyentes, ya que existen sistemas híbridos en los cuales se combina el empleo de los métodos de aprendizaje con el conocimiento previo codificado de forma manual.

Los sistemas que utilizan algoritmos de aprendizaje pueden ser supervisados o no supervisados. Los supervisados emplean un corpus etiquetado manualmente para efectuar el entrenamiento, mientras que los no supervisados buscan regularidades en los datos sin emplear ejemplos etiquetados. En la práctica, los sistemas basados en aprendizaje supervisado han obtenido mejores resultados que aquellos basados en aprendizaje no supervisado (Tjong Kim Sang, 2002; Tjong Kim Sang 2003).

Actualmente, la tendencia creciente en el RNE es utilizar formalismos del aprendizaje automático, debido a que éstos pueden adaptarse con mayor facilidad a diferentes dominios de aplicación. No obstante, resulta conveniente la integración de conocimiento lingüístico a estos sistemas debido a la simplicidad y generalidad que puede tener el mismo. Por tal motivo, en este trabajo combinamos clasificadores secuenciales probabilísticos con un conjunto de expresiones regulares, de modo que se pueda aprovechar la flexibilidad y la capacidad de manipular información contextual de los primeros con el conocimiento lingüístico representado por los segundos.

El resto del artículo se estructura como sigue. En la sección 2 se describe el sistema propuesto para el RNE. En la sección 3 se describen los experimentos realizados y por último, en la sección 4, se exponen las conclusiones.

2 Descripción del modelo utilizado para el Reconocimiento de Nombres de Entidades

Para llevar a cabo el RNE, combinamos técnicas de aprendizaje automático con un conjunto de heurísticas. El reconocedor recibe como entrada una secuencia de palabras etiquetadas con su información morfosintáctica, es decir, su parte de la oración y sus características morfológicas, y asigna una etiqueta a cada una. El conjunto de etiquetas se construye según la notación IOB (Ramshaw & Marcus, 1995), según la cual la primera palabra de un nombre de entidad de tipo *NE* se etiqueta como *B-NE* (*Beginning of NE*), las restantes palabras de dicho nombre de entidad como *I-NE* (*Inside NE*) y las palabras que no pertenecen a ningún nombre de entidad se etiquetan como *O* (*Outside*).

Los tipos de nombres de entidades que se reconocen son nombres de personas, organizaciones, lugares y misceláneos (por ejemplo: obras de arte). Además, por su utilidad, se reconocen cantidades y entidades temporales. Dentro de las cantidades se distinguen los porcentajes y las cantidades monetarias, mientras que las entidades temporales se dividen en fechas y horas.

El reconocedor de nombres de entidades está compuesto por dos unidades: un clasificador secuencial y un conjunto de heurísticas. El clasificador secuencial se entrena a partir de un corpus y produce una primera secuencia de etiquetas. Cada heurística es definida mediante una expresión regular y se encarga de corregir las posibles omisiones de nombres de entidades. Sin embargo, no corrigen errores relacionados con el tipo de los nombres de entidades reconocidos ya que la información contextual, que es la que más útil resulta para este propósito, se maneja de forma más natural por el clasificador secuencial. La omisión de nombres de entidades en el reconocimiento puede deberse a la presencia de términos nunca vistos durante el proceso de entrenamiento.

2.1 Modelos de clasificadores secuenciales utilizados

Como clasificadores secuenciales utilizamos dos tipos de modelos markovianos, los modelos ocultos de Markov (HMM, del inglés *hidden Markov model*) (Rabiner, 1989) y los modelos de Markov por proyecciones (PMM, del inglés *projection Markov model*) (Punyakanok & Roth, 2002). Los modelos markovianos tienen la capacidad de modelar de forma natural el carácter secuencial de la clasificación, así como manipular fácilmente la información de contexto.

En los HMM, si se consideran como observaciones a las palabras tal y como aparecen en el texto, surgen los siguientes problemas. Primeramente, en un corpus cualquiera, las palabras individuales aparecen pocas veces, lo que hace que sus probabilidades de emisión sean muy bajas. En segundo lugar, en un corpus no aparecen todas las palabras del idioma, de modo que ante una palabra desconocida se haría necesario estimar su probabilidad de emisión de manera *ad hoc*.

Para enfrentar tales problemas, en esta propuesta se consideran como observaciones vectores de rasgos binarios definidos sobre palabras en lugar de considerar palabras individuales. De esta forma, se obtiene un conjunto de observaciones finito, el cual se restringe a los posibles vectores binarios que se pueden formar. Con esto se pretende que el modelo no memorice *las palabras* sino que más bien se centre en las "situaciones". Por ejemplo, si el nombre propio *Ludovico* no aparece en el corpus de entrenamiento, éste se trataría de la misma forma que *José* ya que podrían estar representados por el mismo vector de rasgos.

Aunque la cantidad de vectores binarios que se pueden formar es en teoría 2^m, donde m es la cantidad de rasgos con los cuales se describen las palabras, en realidad la cantidad de vectores que se forman es menor, debido a las dependencias entre los rasgos. Por ejemplo, si tenemos el rasgo f_i que consiste en que la palabra comience con letra mayúscula y el rasgo f_j que consiste en que sea un nombre propio, nunca se presentará la situación de que f_j sea verdadero y f_i sea falso.

Aunque la aparición de observaciones desconocidas durante el proceso de inferencia se reduce notablemente, aún existe la posibilidad de que éstas aparezcan. Para manipularlas se aplican técnicas de suavizado (Molina, 2004) a la estimación de las probabilidades de emisión. A las probabilidades de transición, por su parte, no se les aplican técnicas de suavizado, de manera que las restricciones concernientes a las secuencias de etiquetas legales queden codificadas en la matriz de probabilidades de transición.

Adicionalmente, se estima la probabilidad de clasificación de un vector de rasgos $<f_1,\ldots, f_m>$ con una etiqueta c_i, o sea, $P(c = c_i \mid f_1 = v_1,\ldots, f_m = v_m)$ y se utilizan las probabilidades obtenidas de esta forma en un combinador (Punyakanok & Roth, 2002) basado en modelos de Markov por proyecciones.

2.2 Tipos de rasgos utilizados

Se utilizan rasgos que se refieren a la morfología de las palabras (por ejemplo, si la palabra comienza o no con mayúscula), a su información morfosintáctica (parte de la oración, etc.), a su pertenencia a algún diccionario, así como rasgos léxicos.

Los diccionarios que se utilizan contienen títulos de personas (por ejemplo: *Sr.* o *Ing.*), nombres de personas, apellidos, cargos, tipos de organizaciones (por ejemplo: *Federación* o *Asociación*), tipos de lugares, nombres de ciudades, regiones, países y divisiones administrativas, nombres de monedas, los meses del año, etc. Por su parte, mediante los rasgos léxicos se chequea la palabra que se analiza, la palabra que le antecede o la que se encuentra dos posiciones por delante de ella. Mediante éstos se chequea la ocurrencia de palabras sintácticamente importantes, tales como conjunciones o preposiciones que indican pertenencia, origen, destino, etc.

Una combinación adecuada de estos rasgos permite que un modelo sea capaz de aprender patrones que funcionen de forma similar a como lo haría un conjunto de reglas. No obstante, esto es aprendido a partir del corpus; o sea, no es necesario fijar previamente ni el conjunto de reglas ni la forma que pueden tener éstas, sino que es posible aprender las relaciones teniendo en cuenta los datos que se observan durante el entrenamiento.

En la Tabla 2.1 se describen los tipos de rasgos utilizados.

Tipo de rasgo	Descripción de los rasgos que componen el conjunto
A	Rasgos sobre la morfología: uso de mayúsculas, formato de fecha y hora y símbolos de monedas.
B	Rasgos sobre la información morfosintáctica: sustantivos propios y comunes, adjetivos y participios en su función adjetiva, numerales, siglas e identificadores.
C	Pertenencia a alguno de los diccionarios definidos.
D	Rasgos léxicos referidos a la palabra que se analiza, preposiciones que indican destino u origen (*a*, *desde*, *hacia*, etc.), conjunciones, comas, etc.
E	Rasgos léxicos referidos a la palabra que precede a la que se analiza, preposiciones que indican destino u origen (*a*, *desde*, *hacia*, etc.), conjunciones, comas, etc.
F	Rasgos léxicos referidos a la palabra que se encuentra dos posiciones antes de la palabra que se analiza, preposiciones que indican destino u origen (*a*, *desde*, *hacia*, etc.), conjunciones, comas, etc.
G	Pertenencia a diccionarios escogidos: títulos de personas, nombres no personales, tipos de lugares y tipos de organizaciones.
H	Pertenencia a los diccionarios que forman parte de G, además de los diccionarios de nombres de ciudades, divisiones administrativas, países y regiones.

Tabla 2.1. Tipos de rasgos utilizados

2.3 Heurísticas para la corrección de omisiones

Las heurísticas utilizadas tienen la forma de expresiones regulares sobre el alfabeto formado por los rasgos binarios definidos sobre las palabras.

Estas heurísticas pueden lograr un alto grado de generalidad. Por ejemplo, una expresión regular de la forma *(Título + ξ) SP$^+$((ξ + 'del' + 'de la' + 'de las' + 'de los') SP$^+$)**, donde *SP* significa sustantivo propio, es capaz de reconocer nombres de personas tales como *Ing. Juan Carlos Pérez de la Cruz* y *María de las Mercedes Zamora Suárez del Villar*.

Debe notarse que, a pesar de la generalidad de las heurísticas, éstas no deben utilizarse como único recurso en la solución del problema del RNE ya que éstas no pueden resolver los casos de ambigüedad como la existente entre los nombres de personas y los nombres de lugares y entre éstos y los nombres de organizaciones. Por ejemplo, en la oración *"La actividad se desarrolló en el Guillermón Moncada"* la expresión regular anterior clasificaría *Guillermón Moncada* como un nombre de persona a pesar de que en este caso se trata de un nombre de lugar.

3 Resultados experimentales

Los experimentos se llevaron a cabo sobre un corpus etiquetado manualmente que incluye 239706 palabras, el cual, con el propósito de realizar una validación cruzada, fue dividido en tres partes de aproximadamente la misma cantidad de oraciones.

Con el objetivo de mostrar el efecto de la combinación de los clasificadores secuenciales y las heurísticas, así como evaluar la importancia de diferentes tipos de rasgos, se construyó un conjunto de reconocedores usando distintos tipos de clasificadores secuenciales y combinaciones de rasgos. Para cada uno de estos se calcularon los promedios de los valores de precisión, relevancia y F_1 en cada una de las divisiones del corpus. Los resultados obtenidos se muestran en la Tabla 3.1. En la misma, la primera columna especifica el clasificador y las combinaciones de rasgos utilizados en cada caso, mientras que para cada una de las medidas, la columna de la izquierda muestra el valor obtenido por el clasificador sin utilizar las heurísticas y la columna de la derecha muestra el valor obtenido al combinar el clasificador secuencial con las heurísticas.

Clasificadores y Combinaciones de rasgos	Precisión		Relevancia		F_1	
	Clasif.	+ ER	Clasif.	+ ER	Clasif.	+ ER
HMM (ABC)	0,6933	0,6900	0,6867	0,7133	0,6933	0,7000
HMM (ABCD)	0,7100	0,7033	0,7000	0,7300	0,7067	0,7167
HMM (ABD)	0,6167	0,6200	0,6000	0,6300	0,6100	0,6267
PMM (ABCD)	0,7600	0,7300	0,7200	0,7900	0,7400	0,7667
PMM (ABCDE)	0,7700	0,7500	0,7433	0,7900	0,7533	0,7633
PMM (ABCE)	0,7700	0,7500	0,7667	0,8000	0,7667	0,7767
PMM (ABDEG)	0,7333	0,7200	0,7267	0,7700	0,7333	0,7467
PMM (ABCDEF)	0,7867	0,7567	0,7500	0,7900	0,7667	0,7733

Tabla 3.1. Resultados experimentales

Como se puede apreciar en la tabla, la utilización de las expresiones regulares provoca un descenso en el valor de la precisión. Esto se debe a la clasificación incorrecta de algunos nombres de entidades que habían sido omitidos. Sin embargo, el valor de la relevancia en todos los casos aumenta debido a la recuperación de una mayor cantidad de nombres mediante las expresiones regulares. Dado que el aumento de la relevancia es mayor que la disminución de la precisión, el valor de la medida F_1 aumenta en todos los casos.

Se puede observar que el clasificador basado en PMM que utiliza la combinación de rasgos ABCE obtiene el mejor resultado.

Conclusiones

En este artículo se presentó un enfoque híbrido al Reconocimiento de Nombres de Entidades para el idioma español que combina clasificadores secuenciales markovianos con un conjunto de heurísticas en forma de expresiones regulares.

Los resultados experimentales avalan nuestro criterio de que las medidas de calidad de un sistema para RNE en español basado en clasificadores secuenciales probabilísticos aumentan con la integración de conocimiento lingüístico mediante expresiones regulares. Además, estos resultados permiten corroborar que la elección de los rasgos que se utilizan para representar las palabras influye notablemente en la calidad del etiquetado.

Como trabajo futuro nos proponemos definir nuevos tipos de rasgos y utilizar modelos de máxima entropía (Borthwick, 1999) para la estimación de las probabilidades, así como campos aleatorios condicionales (Lafferty, McCallum & Pereira, 2001) como combinadores.

Referencias bibliográficas

Borthwick, A.: *A Maximum Entropy Approach to Named Entity Recognition*, Ph.D. Thesis, Computer Science Department, New York University, New York, USA, 1999.

Lafferty, J., McCallum A., Pereira F.: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 2001.

Molina, A.: *Desambiguación en procesamiento de lenguaje natural mediante técnicas de aprendizaje automático*, Tesis doctoral, Universidad Politécnica de Valencia, 2004.

Punyakanok, V., Roth, D.: *Inference with Classifiers: The Phrase Identification Problem*, Technical Report, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 2002.

Rabiner, L. R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, Vol. 77, No. 2, 1989.

Ramshaw, L. A., Marcus, M. P.: *Text Chunking using Transformation-based Learning*, Proceedings of the Third Annual Workshop on Very Large Corpora, Boston, MA, USA, 1995.

Tjong Kim Sang, E.: *Introduction to the CoNLL-2002 Shared Task: Language Independent Named Entity Recognition*, Proceedings of CoNNL-2002, Taipei, Taiwan, 2002.

Tjong Kim Sang, E.: *Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition*, Proceedings of CoNNL-2003, Edmonton, Canada, 2003.