Resolución de Ambigüedades en el Etiquetado de un Texto mediante un Modelo de Regresión Logística[1]

DISAMBIGUATION IN TEXT LABELING THROUGH A LOGISTIC REGRESSION MODEL

Celina Beltrán
GRUPO INFOSUR UNR
Rosario, Argentina
beltranc@dat1.net.ar

Abstract

This paper seeks to solve, through statistical models, some of the ambiguities frequently observed during the text-labeling process. The statistical model presented is logistic regression. It is calculated on a training text manually labeled and supervised. The explanatory variables used to predict the right label and thus solve the ambiguity will be the label observed in the previous word and the label observed in the following word. Models to solve ambiguities are considered: determiner/clitic (DET/CL) and noun/verb (NOM/V). Models are then evaluated on a new text.

For the DET/CL ambiguity, the percentage of correct classification is 98.8%. The model assigns a higher probability to the DET label when the previous label is a verb or a preposition. With respect to the information of the following label, the model assigns a higher probability for a CL when followed by a verb than when followed by any other label.

In the resolution of the NOM/V ambiguity, the resulting model assigns a higher probability for NOM when the preceding label is a determiner or a preposition and when the following label is an adjective. The percentage of correct classification is 86.8%.

Keywords: Logistic regression model, dissambiguation

Resumen

En este trabajo se busca resolver mediante modelos estadísticos algunas de las ambigüedades observadas con frecuencia durante el proceso de etiquetado de un texto. El modelo estadístico planteado es el de regresión logística. Es estimado a partir de un texto de entrenamiento etiquetado y supervisado manualmente. Las variables explicativas utilizadas para predecir la etiqueta correcta, y así resolver la ambigüedad, serán la etiqueta observada en la palabra anterior y la etiqueta observada en la palabra siguiente. Se estiman modelos para resolver las ambigüedades: determinante/clítico (DET/CL) y nombre/verbo (NOM/V). Los modelos son luego evaluados en un nuevo texto.

Para la ambigüedad DET/CL, el porcentaje de clasificación correcta es 98.8%. El modelo otorga mayor probabilidad a la etiqueta DET cuando la etiqueta anterior es verbo o preposición. Con respecto a la información de la etiqueta posterior, el modelo asigna una probabilidad de CL mayor

[1] This paper is part of a Dissertation supervised by Dr. Gabriel G. Bès.

si le sigue un verbo que si le sigue otra etiqueta.

En la resolución de la ambigüedad NOM/V, el modelo resultante asigna una probabilidad mayor de NOM cuando la etiqueta anterior es determinante o preposición y si la etiqueta siguiente es un adjetivo. El porcentaje de clasificación correcta es 86.8%.

Palabras claves: Modelo de regresión logística, resolución de ambigüedades.

1. INTRODUCCION

Existen dos enfoques mediante los cuales se han abordado las tareas del procesamiento del lenguaje natural: uno basado en información lingüística y otro basado en técnicas estadísticas. La corriente estadística considera que es posible obtener / aprender la estructura de un lenguaje por la observación de grandes cantidades de textos a los que se aplica procedimientos estadísticos y métodos de generalización inductiva. La hipótesis de trabajo bajo este enfoque sería que cuanto más frecuente es un fenómeno lingüístico en un corpus, más se puede considerar que es relevante para el lenguaje al que pertenece el mismo y en sentido contrario, los fenómenos pocos frecuentes pueden indicar errores o inconsistencias. Se defiende esta posición basándose en la posibilidad de lograr una gran cobertura. Por otro lado, la corriente lingüística comprende una serie de etapas que comienza con la segmentación y lematización del texto (serie de expresiones en código ASCII), prosigue con el etiquetado y análisis morfológico para estar luego en condiciones de realizar un análisis sintáctico, que puede estar precedido por análisis pre sintáctico o de gramáticas locales. Su objetivo final es el análisis semántico que va a necesitar de las etapas anteriores. Dentro de esta corriente se aduce el logro de una mayor precisión.

Por lo tanto, existen al menos dos maneras de desambiguar las asignaciones categoriales: por técnicas estadísticas y por conocimiento de la lengua. Las primeras utilizan en general modelos probabilísticos los que van a observar las frecuencias de apariciones entre secuencias de categorías en un corpus de entrenamiento para deducir el sistema subyacente estadístico. El conocimiento de la lengua, en cambio, es utilizado cuando se ha logrado especificar un sistema sub-yacente expresado en reglas formales y explícitas, las cuales describen las representaciones que deben asignarse a los enunciados de una lengua, que hayan o no sido previamente observados en un corpus [1].

En este trabajo se busca resolver ambigüedades mediante modelos estadísticos. El modelo estadístico que se utiliza es el de Regresión Logística. El mismo será estimado con un corpus de entrenamiento etiquetado y supervisado.

Se estima un modelo para cada una de estas ambigüedades presentadas:

- ✓ Determinante / Clítico
- ✓ Nombre / Verbo

2. EL MODELO DE REGRESION LOGISTICA

2.1. Formulación del modelo

El modelo de regresión logística múltiple [2] viene expresado por la ecuación

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

donde

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \ln\left(\frac{P(Y=1/\mathbf{x})}{P(Y=0/\mathbf{x})}\right)$$

y por lo tanto

$$P(Y=1) = \pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}.$$

El vector X es el vector de variables explicativas y éstas pueden estar medidas en distintas escalas. La interpretación de los coeficientes que las acompañan dependerá de la escala de las variables y la forma que se ingresen al modelo.

2.2. Interpretación de los coeficientes

En este caso las variables explicativas utilizadas para predecir la etiqueta correcta y así resolver la ambigüedad en estudio serán:

- *Etiqueta observada en la palabra anterior*
- *Etiqueta observada en la palabra siguiente*

Puesto que cada una de estas variables es categórica se ingresarán al modelo mediante un conjunto de variables dummies con una de sus categorías como la categoría de referencia.

El odds o cociente de las probabilidades

$$\left(\frac{P(Y=1/\mathbf{x})}{P(Y=0/\mathbf{x})}\right) = e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}$$

compara la probabilidad del evento Y=1 con la probabilidad del evento Y=0. El coeficiente β_j es una medida de cambio en la razón de probabilidades. Por ejemplo, si para una variable categórica el coeficiente de la categoría j es positivo, entonces el cociente de probabilidades aumentará, indicando que la probabilidad de Y=1 es mayor para la categoría j comparándola con la categoría de referencia.

La razón de odds es el cociente entre los odds calculados a partir de dos valores diferentes de variables explicativas. Considérese un ejemplo de regresión logística con una variable explicativa categórica con tres categorías, C_1, C_2, y C_3, por lo tanto el modelo incorpora dos variables dummies

$$g(\mathbf{x}) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

donde D_1 y D_2 son la variables que toman valores 0 y 1, por ejemplo:

	D_1	D_2
C_1	1	0
C_2	0	1
C_3	0	0

esto es

- cuando en la variable original asume la categoría C_1, le corresponde un valor de 1 para la variable D_1 y 0 para D_2
- cuando en la variable original asume la categoría C_2, le corresponde 0 en D_1 y 1 en D_2
- cuando la categoría asumida en la variable original es C_3 (categoría de referencia), le corresponde en ambas variables diseño el valor 0.

La razón de odds que compara el odds para la categoría C_1 y la categoría C_3, que es la de referencia, tiene la expresión

$$\frac{\left(\frac{P(Y=1/\mathbf{x}=C1)}{P(Y=0/\mathbf{x}=C1)}\right)}{\left(\frac{P(Y=1/\mathbf{x}=C3)}{P(Y=0/\mathbf{x}=C3)}\right)}=\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}=e^{\beta_1}$$

y por lo tanto e^{β_1} representa el cambio en la chance de P(Y=1) cuando X=C_1 que cuando X=C_3. Por ejemplo, si $\beta_1 > 0$ entonces la chance de P(Y=1) será mayor cuando la categoría de x es C_1 que cuando es C_3.

De la misma manera se compara el odds para la categoría C_2 y la categoría C_3

$$\frac{\left(\frac{P(Y=1/\mathbf{x}=C2)}{P(Y=0/\mathbf{x}=C2)}\right)}{\left(\frac{P(Y=1/\mathbf{x}=C3)}{P(Y=0/\mathbf{x}=C3)}\right)}=\frac{e^{\beta_0+\beta_2}}{e^{\beta_0}}=e^{\beta_2}$$

e^{β_2} representa el cambio en la chance de P(Y=1) cuando X=C_2 que cuando X=C_3. Si $\beta_2 > 0$ entonces la chance de P(Y=1) será mayor cuando la categoría de x es C_2 que cuando es C_3.

En esta aplicación el evento Y=1 corresponderá a la asignación de una de las categorías involucradas en la ambigüedad.

2.3. Clasificación. Análisis discriminante predictivo con regresión logística

En el caso de regresión logística con variable respuesta dicotómica se tienen dos grupos para predecir, Y=1 e Y=0. La asignación de cada unidad a cada grupo2 se realiza utilizando el valor estimado de la probabilidad de Y=1 e Y=0 obtenidas a partir del modelo ajustado. Estableciendo un punto de corte que en general es 0.5, si la probabilidad de Y=1 para la unidad es superior a 0.5 entonces a dicha unidad se la clasifica en el grupo Y=1, caso contrario se la clasifica en el grupo Y=0.

Utilizando este procedimiento es posible construir una tabla de clasificación en un nuevo corpus utilizado como evaluación (Tabla 1).

[2] En esta aplicación cada unidad a clasificar es una palabra y cada grupo corresponde a cada una de las etiquetas involucradas en la ambigüedad.

Tabla 1: Tabla de clasificación.

Grupo predicho / *Grupo observado*	*Y=0*	***Y=1***
Y=0	n_{00}	n_{01}
Y=1	n_{10}	n_{11}

El porcentaje de clasificación correcta viene dado por

$$\frac{n_{00} + n_{11}}{n}.100\%$$

donde

$$n = n_{00} + n_{10} + n_{01} + n_{11}$$

es el total de unidades en el corpus de evaluación.

3. ESTIMACION DEL MODELO DE REGRESION LOGISTICA

3.1. Corpus de entrenamiento

El corpus de entrenamiento para estimar el modelo presentado se conformó con textos provenientes de las páginas web de periódicos argentinos. El mismo está constituido por 8541 palabras y 412 oraciones.

El corpus de entrenamiento se etiquetó utilizando el software Smorph y luego se supervizó manualmente para controlar los resultados. Este conjunto de textos analizado morfológicamente constituyó la muestra sobre la cual se extrajo la información acerca de las ambigüedades en estudio.

Se conformaron dos bases de datos. Una base de datos corresponde a los casos observados de ambigüedad determinante/clítico (DET/CL) y la otra base corresponde a los casos observados de la ambigüedad nombre/verbo (NOM/V).

3.2. Selección del modelo

Para seleccionar el modelo adecuado en cada caso se utiliza la comparación de modelos mediante el test de razón de verosimilitudes. Este procedimiento compara los logaritmos de las verosimilitudes de los dos modelos: el modelo completo (con todas las variables explicativas) y el modelo reducido (con ausencia de alguna variable) [3]. El contraste chi-cuadrado para la reducción en la verosimilitud proporciona una medida de la "mejora" en el modelo al incluir la variable en cuestión. Si esta reducción no es estadísticamente significativa entonces la variable en estudio no es útil para "explicar" la etiqueta de una palabra.

4. RESULTADOS

4.1. Ambigüedad DET/CL

El primer modelo de regresión logística estimado es para resolver la ambigüedad DETERMINANTE/CLITICO (DET/CL). La estimación de dicho modelo se realiza sobre una base de datos que contiene la información de un corpus de entrenamiento etiquetado y luego supervisado. De esta manera, para cada palabra, se tiene la etiqueta correcta y las etiquetas asignadas por el analizador en el caso que la palabra admitiera más de una etiqueta y por lo tanto surge la ambigüedad. Por ejemplo, si en el texto de entrenamiento se lee:

LA CRISIS EN LA PROVINCIA

en la base de datos se encontrará la información y estructura de la tabla 2.

Tabla 2: Base de datos.

nro secuencia	palabra observada	etiqueta 1	etiqueta 2	etiqueta correcta	Ambigüedad
1	**LA**	**det**	**cl**	**det**	**si**
2	CRISIS	nom	S/E	nom	no
3	EN	prep	S/E	prep	no
4	**LA**	**det**	**cl**	**det**	**si**
5	PROVINCIA	nom	S/E	nom	no

Así se construye una base de datos con aquellos casos en los cuales existe la ambigüedad DET/CL (n=641).

La variable binaria respuesta se define de modo tal que toma el valor 1 cuando la etiqueta correcta es CL y toma el valor 0 cuando la etiqueta correcta es DET.

Para este caso las variables explicativas utilizadas son:

- *Etiqueta observada en la palabra anterior*
  - ✓ Etiqueta Verbo
  - ✓ Etiqueta Preposición
  - ✓ Otra Etiqueta
- *Etiqueta observada en la palabra siguiente*
  - ✓ Etiqueta Verbo
  - ✓ Etiqueta Nombre
  - ✓ Etiqueta Adjetivo
  - ✓ Otra Etiqueta

Para la categorización de estas dos variables se tuvo en cuenta las frecuencias observadas en el texto de entrenamiento.

El modelo de regresión logística estimado con las dos variables explicativas, etiqueta de la palabra anterior y posterior, se presenta en la tabla 3 y la clasificación en la tabla 4.

Tabla 3: Modelo de regresión completo.

VARIABLE	***ESTIMADOR***	***Significación de la variable****
Etiqueta Anterior		
Anterior(1)	−2.88	0.000
Anterior (2)	−11.76	
Etiqueta posterior		
Posterior (1)	14.16	0.000
Posterior (2)	0.21	
Posterior (3)	0.15	
Intercepto	−12.89	0.92

*La significación de cada variable se realizó mediante el test del cociente de verosimilitud.

Tabla 4: Tabla de clasificación[3]

Grupo predicho / *Grupo observado*	**Determinante**	***Clítico***
Determinante	456	5
Clítico	1	25

El porcentaje de clasificación correcta para los determinantes es de 98.9% y para clítico es de 96.2% y el porcentaje de clasificación correcta global es 98.8%.

El coeficiente correspondiente a la primera variable indicadora de la etiqueta anterior se refiere a la comparación de la probabilidad de CL cuando la etiqueta anterior es V y cuando es otra etiqueta. Este coeficiente resulta negativo, por lo cual se puede interpretar que la probabilidad de CL es menor si lo antecede un verbo que si la etiqueta anterior es otra. Similarmente, si analizamos el segundo coeficiente, correspondiente a la segunda variable indicadora, compara la probabilidad de CL cuando la etiqueta anterior es preposición y cuando es otra etiqueta. Este valor resulta también negativo indicando que la probabilidad de CL disminuye si lo antecede una preposición. Dicho de otra manera, cuando la etiqueta anterior es verbo o preposición es más probable que la etiqueta correcta sea DET que CL.

Con respecto a la etiqueta posterior, el coeficiente correspondiente a la primera variable indicadora compara la probabilidad de CL cuando la etiqueta siguiente es V y cuando es otra etiqueta. Este coeficiente resulta positivo, por lo cual se interpreta que la probabilidad de CL es mayor si le sigue un verbo que si le sigue otra etiqueta.

[3] Evaluación del modelo sobre un nuevo corpus de características similares al de entrenamiento.

4.2. Ambigüedad NOM/V

El segundo modelo de regresión logística estimado es para resolver la ambigüedad NOMBRE/VERBO (NOM/V).

La estimación del modelo se realiza sobre la base de datos presentada en el punto 4.1 que contiene la información de un corpus de entrenamiento etiquetado y luego supervisado. Para esta aplicación, se retuvo la información de los casos en los cuales el analizador morfológico le asignó las dos etiquetas: Nombre y Verbo (n=326).

La variable binaria respuesta se define de modo tal que toma el valor 1 cuando la etiqueta correcta es V y toma el valor 0 cuando la etiqueta correcta es NOM.

Para este caso las variables explicativas utilizadas son:

- *Etiqueta observada en la palabra anterior*
  - ✓ Etiqueta Preposición
  - ✓ Etiqueta Determinante
  - ✓ Otra Etiqueta
- *Etiqueta observada en la palabra siguiente*
  - ✓ Etiqueta Preposición
  - ✓ Etiqueta Adjetivo
  - ✓ Otra Etiqueta

De la misma manera que en la modelización anterior, para la categorización de estas dos variables se tuvo en cuenta las frecuencias observadas en el texto de entrenamiento.

El modelo de regresión logística estimado con las dos variables explicativas, etiqueta de la palabra anterior y posterior, se presenta en la tabla 5 y la tabla de clasificación es la 6.

Tabla 5: Modelo de regresión completo.

VARIABLE	***ESTIMADOR***	***Significación de la variable****
Etiqueta Anterior		
Anterior(1)	−1.43	0.000
Anterior (2)	−3.93	
Etiqueta posterior		
Posterior (1)	−0.69	0.000
Posterior (2)	−0.72	
Intercepto	−0.69	

*La significación de cada variable se realizó mediante el test del cociente de verosimilitud.

El coeficiente correspondiente a la primera variable indicadora de la etiqueta anterior se refiere a la comparación de la probabilidad de V cuando la etiqueta anterior es preposición y cuando es otra

etiqueta. Este coeficiente resulta negativo, por lo cual se puede interpretar que la probabilidad de V es menor si lo antecede una preposición que si la etiqueta anterior es otra. Similarmente, si analizamos el segundo coeficiente, correspondiente a la segunda variable indicadora, compara la probabilidad de V cuando la etiqueta anterior es determinante y cuando es otra etiqueta. Este valor resulta también negativo indicando que la probabilidad de V disminuye si lo antecede un determinante. Dicho de otra manera, cuando la etiqueta anterior es determinante o preposición es más probable que la etiqueta correcta sea NOM que V.

Con respecto a la etiqueta posterior, la probabilidad de que sea V disminuye cuando le sigue una preposición o un determinante, es decir que la probabilidad de NOM aumenta si la etiqueta siguiente es un adjetivo o una preposición.

Tabla 6: Tabla de clasificación[4]

Grupo predicho / *Grupo observado*	*Nombre*	***Verbo***
Nombre	226	0
Verbo	34	0

El porcentaje de clasificación correcta para los nombres es de 100% y para verbo es de 0% y el porcentaje de clasificación correcta global es 86.9%.

5. CONCLUSIONES

En este estudio se muestra que el modelo estimado para resolver la ambigüedad DET/CL otorga mayor probabilidad a la etiqueta DET cuando la etiqueta anterior es verbo o preposición. Con respecto a la información de la etiqueta posterior, el modelo asigna una probabilidad de CL mayor si le sigue un verbo que si le sigue otra etiqueta. La precisión resultante es del 98,8%. Esta información que surge del modelo estadístico estimado del corpus de entrenamiento coincide con reglas lingüísticas que se pueden definir con el mismo propósito [4]. Por ejemplo, para resolver esta ambigüedad en el análisis de SMORPH [5], se define en MPS [6] la siguiente regla:

CLITICO+VERBO

```
S1 [ L1, 'EMS','cl'] S2[ L2, 'EMS','v']  --> S1+S2 [ L1+L2, 'EMS','cl_v'].
```

```
"cuando se tenga la etiqueta CL, independientemente de que tenga otra etiqueta,
si le sigue una palabra con etiqueta V, entonces se está frente a un CL+V"
```

En la resolución de la ambigüedad NOM/V, el modelo resultante otorga una probabilidad mayor de NOM cuando la etiqueta anterior es determinante o preposición y si la etiqueta siguiente es un

[4] Evaluación del modelo sobre un nuevo corpus de características similares al de entrenamiento.

adjetivo. Se consigue trabajar con una precisión del 86,8%. En este caso el modelo asigna a todos los casos en los que la etiqueta es V, la etiqueta NOM, provocando una reducción de la precisión.

Este modelo resuelve la ambigüedad de la misma manera que actúa la siguiente regla lingüística definida en MPS para un análisis realizado en SMORPH:

DETERMINANTE+NOMBRE+ADJETIVO

```
S1 [ L1, 'EMS','det'] S2[ L2, 'EMS','nom'] S3[ L3, 'EMS','adj'] --> S1+S2+S3 [
L1+L2+L3, 'EMS','det_nom_adj'].
```

```
"cuando se tenga la etiqueta NOM, independientemente de que tenga otra etiqueta,
si la antecede una palabra con et DET y le sigue una palabra con etiqueta ADJ,
entonces se está frente a un DET+NOM+ADJ"
```

Referencias

[1] Bès, Gabriel G. ; Solana, Zulema; Beltrán, Celina "Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico". Publicación de Facultad de Filosofía y Letras de la Universidad Nacional de Cuyo. JALIMI 2005.

[2] Hosmer, David; Lemeshow, Stanley. "Applied Logistic Regression". Jhon Wiley & Sons. New York. 1989.

[3] Stokes, Maura E.; Davis, Charles S.; Koch, Gary G. “Categorical Data Analysis using the SAS® System”. SAS Institute Inc. 1999.

[4] Beltrán, Celina. Comparación y evaluación de dos etiquetadores. Revista INFOSUR. Nro. 2 Agosto 2008.

[5] Aït-Mokhtar, Salah; Rodrigo Mateos, José Lázaro 1995 Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. SEPLN Revista nº 17 pags 29-41.

[6] MPS ha sido especificado en el GRIL por Caroline Hagège, José Rodrigo, Gabriel G. Bès y Faiza Abacci, e implantado en C++ en un contexto Windows por Faiza Abacci.