Etiquetado de Roles Semánticos en el marco del corpus CoNLL 09

Vladímir Robles Bykbaev
vrobles@ups.edu.ec

Resumen

En este trabajo se aplican modelos de máxima entropía, a fin de etiquetar los roles semánticos que posee el corpus CoNLL 09. Se realizan dos aproximaciones: una primera basada en literales tácitos y una segunda que usa pesos para caracterizar los constituyentes de los predicados. Luego de analizar los resultados se sugieren mejoras en el proceso de entrenamiento, que permitirán obtener valores más bajos de error e incrementar el rendimiento general del sistema.

Palabras clave: Etiquetado de Roles Semánticos, modelos de máxima entropía, CoNLL 2009.

Abstract

In this work we applied the maximum entropy models to label semantic roles of the sentences in the CoNLL 09 corpus. We propose two approximations: the first one uses single literals and the second approximation introduces weights to obtain a better classifier of constituents in the sentences. After make several experiments we suggest improvements in the whole process to obtain a lower error rate.

Keywords: *Semantic Role Labeling, Maximum Entropy Models, CoNLL 2009.*

Forma sugerida de citar: Robles Bykbaev, Vladímir. 2011. "Etiquetado de Roles Semánticos en el marco del corpus CoNLL 09". Ingenius No. 5. Enero/Junio. Pp. 63-72.

1. Introducción

En los últimos años se han publicado una gran cantidad de estudios y artículos sobre las relaciones existentes entre la sintaxis del lenguaje y la semántica. Actualmente existe un cierto consenso en la existencia de tales relaciones [4] y cuáles son los elementos que participan de ellas. Sin embargo, debido a esta diversidad, existen varias líneas de investigación, así como herramientas desarrolladas para las tareas de análisis de roles semánticos [9]:

- **Enfoques basados en *corpus***. Aprendizaje de Máquina (Machine Learning), que busca obtener una función (clasificador) que asigne etiquetas que pertenezcan a una clase definida, verbigracia, agentes, pacientes, experimentador[1], etcétera, a nuevas muestras que se presenten al clasificador. Este enfoque se basa en dos tipos de aprendizaje:
  - *Supervisado*. Busca predecir el valor de la función que asigna la etiqueta de clase a nuevas entradas, luego de haber proporcionado un número determinado de muestras de entrenamiento.
  - *No supervisado*. No se proporcionan las clases *a priori*, por lo que la función debe ser capaz de determinar las clases con base en la agrupación de muestras que compartan las mismas características.
- **Enfoques basados en conocimiento**. Usan representación simbólica del conocimiento humano para resolver problemas dependientes del dominio y el idioma. Los formalismos más utilizados son las reglas, redes semánticas, marcos (frames), lógica de predicados, etcétera.

Cabe mencionar que de los dos enfoques citados, el primero es el más utilizado, debido a la gran cantidad de algoritmos que existen, a las posibilidades de generalización (en la mayoría de tareas), sencillez de implementación y otros factores. El enfoque instituido en conocimiento presenta ciertas dificultades, como el mantenimiento, que es muy costoso, ya que este tipo de solución es dependiente del dominio y el idioma en el que se trabaja.

En el presente trabajo nos fundamentaremos en un enfoque orientado al corpus, utilizando una metodología de tipo supervisada. La estructura que hemos seguido se detalla a continuación: en la sección 2 se describe de forma breve las características más importantes del método estadístico de máxima entropía, con una aproximación al procesamiento del lenguaje natural; en la sección 3 describimos el corpus empleado, así como las características inherentes al mismo. La sección 4 está divida en dos partes: la primera trata sobre la modelización y selección de constituyentes para la creación de los modelos de máxima entropía; la segunda parte versa sobre la metodología desarrollada para la extracción de información, así como del desarrollo de herramientas de soporte de preprocesado del corpus. En la sección 5 se pasa a revisar el estado del arte, esto es, técnicas empleadas y resultados obtenidos en las últimas competiciones. En la sección 6 indicamos los experimentos que se han realizado con base en la configuración de las funciones características extraídas. A continuación, en la sección 7, se detallan los resultados obtenidos. Finalmente la sección 8 contiene las conclusiones del trabajo y la sección 9 el trabajo futuro que se podría llevar a cabo para mejorar el rendimiento general del sistema.

1 Existen diversas formas de referirse a los elementos que participan dentro de una relación semántica. Una de ellas es la que considera a Agentes (los que inician la acción), Pacientes (entidad que padece la acción), Experimentadores (entidad consciente de la acción pero que no la controla) y Tema (entidad movida por alguna acción).

2. Materiales y métodos

2.1 Máxima entropía

Esta aproximación se fundamenta en el cálculo de la distribución de probabilidad que maximice la entropía o desconocimiento, lo que evitará alguna inducción de conocimiento que no se refleje en los datos. Las características de este algoritmo son las siguientes:

- Adecuado para la clasificación de multiclases.
- Coste computacional relativamente bajo.
- Muy robusto y preciso ante la escasez de datos.
- Se fundamenta en la aplicación de la siguiente fórmula basada en probabilidad [1]:

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{i=1}^{k} \lambda_i f_i(x,y)\right) \quad (1)$$

Donde:

fi = función característica.
λ = peso que se asigna a fi.
k = número de características a tomar en consideración.
Z(x) = factor de normalización.
P(y|x) = probabilidad de condicional de predecir una salida 'y' al haber visto una entrada 'x'.

Existe una fuerte relación entre la máxima entropía y las funciones binarias conocidas como *predicados contextuales* [1]:

$$f_{cp,y'}(x,y) = \begin{cases} 1 \text{ if } y=y' \text{ and } cp(x)=true \\ 0 \text{ en otro caso} \end{cases} \quad (2)$$

Donde cp es el predicado contextual que mapea el par elementos contexto (x) y salida (y) a dos posibles valores (verdadero y falso).

Por ejemplo, en el caso que nos atañe, si consideramos la oración "*Juan saludó a Pedro*", debemos tener en mente la siguiente premisa:

El verbo saludar tiene un solo sentido y requiere dos argumentos semánticos para conformar un sintagma verbal:

- Sujeto » Agente que inicia la acción.
- Complemento directo » Paciente que sufre la acción.

Con esto, si construimos nuestra función característica tendríamos lo siguiente:

$$f_{\text{verbo es saludar}}(x,y) = \begin{cases} 1 \text{ if } y=\text{saludar and} \\ \text{SujetoOración\_} \\ \text{Ubicado\_Antes } (x)=\text{trae} \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

2.2 Máxima entropía: una aproximación más cercana a la lingüística computacional

Uno de los ejemplos más claros e interesantes en el proceso de entendimiento de la máxima entropía lo proporciona [1]:

Vamos a suponer que se desea modelar las decisiones de un experto de traducción del francés al inglés. Este experto debe traducir la palabra **in**. Nuestro modelo deberá estimar la probabilidad de que el experto elija a una frase **f** como la posible traducción de la palabra **in**. Un aspecto importante será seleccionar las traducciones válidas de la palabra, esto es, *{dans, en, `a, au cours de, pendant}*. Con esta información podemos establecer la primera restricción a nuestro modelo:

$$p(dans) + p(en) + p(\grave{a}) + p(au \quad cours \quad de) + p(pedant) = 1$$

Como el lector se imaginará, existen infinitas formas de satisfacer esta ecuación. Si suponemos que el experto en traducción usa siempre estas frases, podemos proyectar que el valor de probabilidad de cada una de las palabras será de 1/5.

Sin embargo, el objetivo es que nos acerquemos a las decisiones del experto. Supongamos ahora que conocemos que el experto escoge un 30% de las ocasiones las palabras *dans* o *pen*. Nuestro modelo deberá ahora satisfacer las siguientes restricciones:

$$p(dans) + p(en) = \frac{3}{10}$$

$$p(dans) + p(en) + p(\grave{a}) + p(au \quad cours \quad de) + p(pedant) = 1$$

Si redistribuimos las probabilidades podremos observar que se cumplen los siguientes valores: *p(dans)=p(en)=3/20, p(`a)=p(au cours de)=p(pendant)=7/30*. Ahora vamos a imaginar que nos informan de un nuevo hecho, el experto escoge las palabras **dans** o **en**, en la mitad de los casos. Con esto podríamos agregar una nueva restricción a nuestro modelo *(p(dans)+p(en)=1/2)*. Ahora, al tratar de redistribuir la probabilidad, veremos que el resultado ya no es tan obvio como antes. Se nos plantean 3 interrogantes: 1. ¿Exactamente que es la distribución uniforme? 2. ¿Cómo medir la uniformidad de nuestro modelo? y 3. ¿Cómo encontrar el modelo que asigne la probabilidad más uniforme, considerando las constantes que se han definido?

Es aquí donde la metodología sustentada en máxima entropía entra en acción para responder a estas dos preguntas: dada una colección de hecho o restricciones, escoger el modelo que sea consistente con esas restricciones, pero en lo que no conoce asigne una distribución de probabilidad uniforme.

2.3 Corpus empleado

Para el desarrollo del presente trabajo se ha utilizado el *corpus* de la competición CoNLL 2009. Éste se encuentra anotado sintáctica y semánticamente y se divide en los siguientes ficheros:

- **Train**. Se conforma de 14.329 oraciones.
- **Development**. 1.655 oraciones.
- **Test**. Este *corpus* no está disponible.

Los *corpus* para el entrenamiento y el desarrollo están organizados en columnas, donde cada columna contiene la información semántica y los roles de los argumentos de cada sintagma. Las principales columnas son las que detallamos a continuación:

- **ID**. Esta columna contiene un valor numérico que identifica de forma única a cada constituyente de la oración.
- **Form**. Contiene el constituyente original.
- **Lemma**[2]. Raíz del constituyente.
- **POS**. *Parts Of Speech*. Contiene la etiqueta de cada constituyente. Este valor indica que puede ser un verbo (v), un adjetivo (a), un determinante (d), etcétera.
- **Feat**. Especifica la información complementaria sobre las características del consti-

2 Para cada columna se han realizado anotaciones manuales (lingüistas), para cada una de estas anotaciones se agrega una columna que señala una anotación propuesta por un etiquetador automático (*machine learning*). Las columnas que son anotadas automáticamente están precedidas por la letra 'p'.

tuyente, esto es, género (masculino o femenino), número (singular, plural o común), etc.

- **Head**. Formada por un valor numérico que indica el constituyente del que depende (es el ID del constituyente principal a este elemento).
- **PRED**. Esta columna tiene sentido para los verbos únicamente, ya que indica el sentido del verbo que está relacionado con los argumentos que poseen un rol semántico. Este sentido se corresponde con el *corpus* ANCORA.
- **APRED1, APRED2, ..., APREDn**. Estas columnas contienen el rol semántico de cada predicado, esto es, el APRED1 será el rol semántico del primer predicado.

En la siguiente tabla podemos apreciar un ejemplo de un fragmento del *corpus* CoNLL-Train:

ID	Form	Lemma	P
1	El	El	D
2	presidente	presidente	n
3	del	Del	s
4	órgano	órgano	n
5	regulador	regulador	a
6	de	De	s
7	las	El	d

Tabla 1. Fragmento del Corpus CoNLL – Train.

2.4 Diseño de la solución

En esta sección revisamos los elementos más importantes del diseño y desarrollo de la solución, verbigracia, la selección de características para entrenar al modelo de máxima entropía, el proceso de extracción de constituyentes del *corpus*, etcétera.

2.4.1 Selección de características para el Modelo

Uno de los aspectos más importantes a la hora de entrenar un modelo basado en máxima entropía, es seleccionar de forma correcta las propiedades (*features*) que permitan caracterizar de forma adecuada a cada predicado y sus argumentos. De acuerdo a [8] deben seleccionarse las palabras que contengan la mayor carga y por ende, información semántica. Por ello, hemos seleccionado los siguientes constituyentes y características, tomando en consideración lo que se plantea en [2], [5], [7] y [8]:

- **Verbos, nombres, adjetivos y adverbios**. Estos elementos contienen mucha información semántica, que puede ayudar a discriminar las características de cada predicado y rol – argumento.
- **Categoría gramatical**. Se especifican las categorías de cada elemento seleccionado, esto es, si son nombres, adjetivos, etcétera.
- **Posición respecto al verbo**. Valor numérico que indica si la palabra está antes o después del verbo (0=antes, 1=después).
- **Argumento - Rol semántico**. En cada característica se indica el constituyente y al final el rol semántico que tiene en el predicado.

A continuación, en la tabla 2, podemos observar un ejemplo de una oración y la función característica que se ha seleccionado para la misma:

Oración	Función Característica – Rol argumento
El presidente del órgano regulador de las telecomunicaciones se mostró partidario de completar esta liberalización de las telecomunicaciones con otras medidas que incentiven la competencia como puede ser abrir el acceso a la información de los clientes de Telefónica a otros operadores.	presidente=suj=n Telecomunicaciones=sn=n=0 regulador=s.a=a=0 órgano=sn=n=0 presidente=suj=n=0 mostrar=v=mostrar. c2 partidario=cpred=a=1 información=sn=n=1 operador=sn=n=1 presidente=suj=n=arg1-tem
	partidario=cpred=a Telecomunicaciones=sn=n=0 regulador=s.a=a=0 órgano=sn=n=0 presidente=suj=n=0 mostrar=v=mostrar. c2 partidario=cpred=a=1 información=sn=n=1 operador=sn=n=1 partidario=cpred=a=arg2-atr

Tabla 2. Ejemplo de función característica para una oración

Como podemos observar, para una oración se debe realizar el siguiente análisis:

a. Dividir la oración en predicados, a través de un análisis POS de tipo B. I. O. (*Begin - Input - Output*), sin embargo esto no ha sido posible, debido a que no existe un *corpus* etiquetado con estos tags. En lugar de ello se ha implementado una solución de análisis recursivo (siguiente apartado).

b. Se debe especificar el predicado y sus características para cada argumento - rol que tenga la oración.

c. Determinar las dependencias semánticas que tiene cada constituyente (respecto al verbo principal).

Una vez que se realiza este análisis, pasamos a entrenar el modelo de máxima entropía.

2.4.2 Metodologías de extracción y codificación de información

Uno de los primeros problemas que se presentó fue que no se contaba un etiquetador de tipo B. I. O. Para solucionar este inconveniente se tuvo que realizar un análisis basado en dependencias recursivas. En la Figura 1 podemos ver esquematizado el proceso llevado a cabo para extraer la información (*features*) del *corpus* de entrenamiento y de desarrollo.

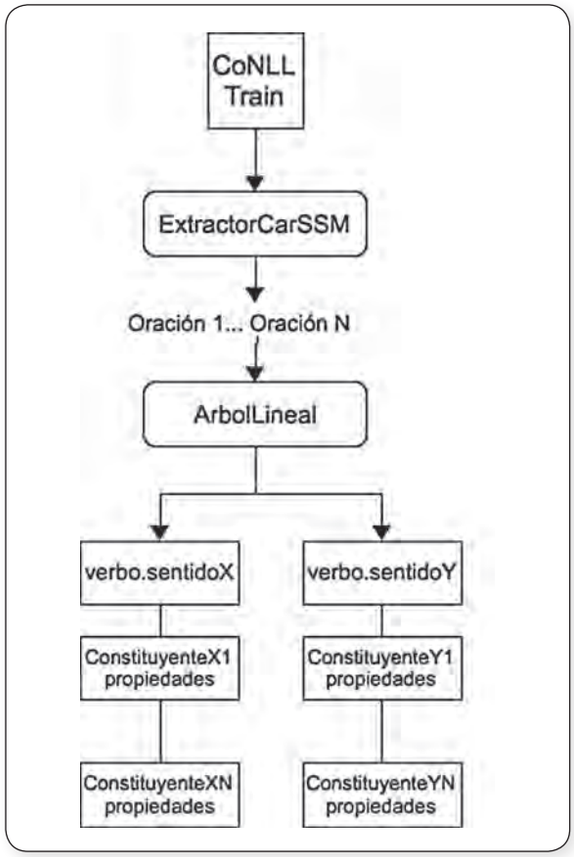

Figura 1. Etapa 1 del proceso etiquetado de roles semánticos

En esta primera etapa se realizan las siguientes tareas:

- **Exploración de dependencias**. Este proceso realiza una búsqueda recursiva de los

elementos que dependen directa o indirectamente del verbo. El objetivo es separar en predicados que estén caracterizados por [rol_1] verbo [rol_2]. Para poder ejecutar esta búsqueda nos basamos en la columna Head que posee el *corpus* y trabajamos con una clase de lista ligada lineal que abstrae esta información.

- **Extracción de características**. A medida que se realiza el análisis del **corpus** vamos extrayendo la información contenida en el mismo y la almacenamos en el fichero de características correspondientes.

Una vez que se han extraído las características más importantes, pasamos a la segunda etapa (Figura 2):

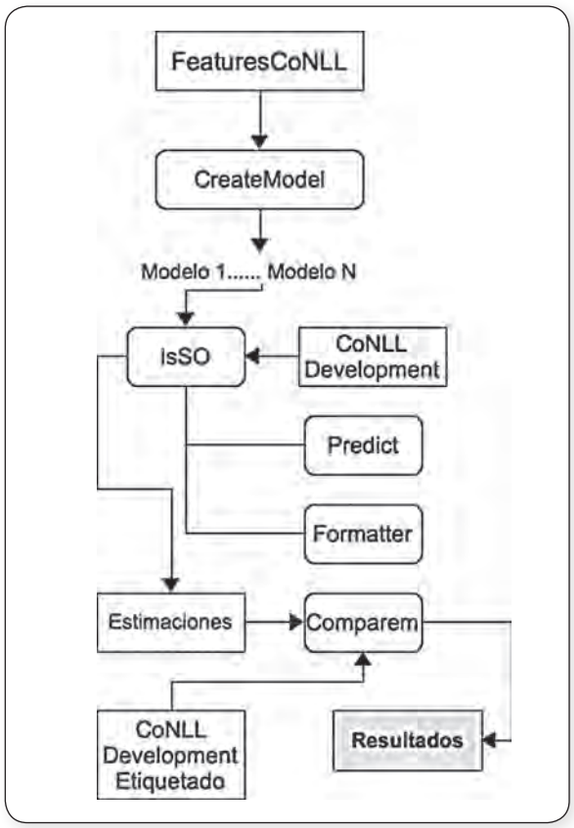

Figura 2. Etapa 2 del proceso etiquetado de roles semánticos

- **Creación del Modelo de Máxima Entropía**. En este paso se crea y entrena el modelo de máxima entropía en base a las características extraídas. Debido al alto número de elementos se generan N modelos, donde el tamaño dependerá de los bloques de datos a manejar [6].
- **Predicción de Valores de Test**. Con soporte en el corpus de desarrollo (que en nuestro caso lo usamos como test), se extraen las características de este corpus y se solicita a los modelos creados que estimen los roles de los argumentos dados.
- **Verificación de resultados**. Una vez que se han obtenido las estimaciones realizamos una comparación con el *corpus* de desarrollo que tienen las etiquetas reales.

2.5 Estado del Arte - BaseLine actual

De acuerdo a la página oficial de SemEval, los últimos resultados en tareas de etiquetado de roles semánticos tienen una precisión de alrededor del 84,30%. En la tabla 3 nos hemos permitido incluir los valores para esta tarea.

Test	Baseline			Best system		
	Prec.	Recall	Fl	Prec.	Recall	Fl
ca.*	83,28	76,88	79,95	84,72	82,12	83,4
es.*	81,61	76,05	78,73	84,3	83,98	84,14
*.in	82,07	80,4	81,38	84,71	84,12	84,41
*.out	82,88	71,48	76,76	84,26	81,84	83,03
.	82,42	76,46	79,32	84,5	83,07	83,78

Tabla 3. BaseLine de la última competencia de SemEval.
Fuente: http://www.lsi.upc.edu/~nlp/semeval/msacs_systems.html

Existen muchas aproximaciones para realizar la tarea de etiquetado de roles semánticos, entre las más importantes tenemos (excluyendo a MaxEnt):

- **Máquinas de soporte vectorial**. Esta alternativa se basa en redes neuronales para crear clasificadores. El objetivo que tienen, es mapear los datos de un espacio original hacia otro espacio euclidiano de mayor dimensionalidad [3].
- **Clasificadores o separadores lineales**. Estos enfoques han demostrado buenos resultados en esta tarea [9].

2.6 Experimentos realizados

Básicamente hemos realizado dos tipos de experimentos; el primero manejando valores literales tácitos, y el segundo, manejando pesos para los argumentos.

En la Tabla 4 podemos observar las funciones características para cada uno de los casos antes mencionados.

Función Característica	Valor
Literal	presidente=suj=n Telecomunicaciones=sn=n=0 regulador=s.a=a=0 órgano=sn=n=0 presidente=suj=n=0 mostrar=v=mostrar. c2 partidario=cpred=a=1 información=sn=n=1 operador=sn=n=1 presidente=suj=n=arg1-tem
Pesos	presidente<=>suj<=>n Telecom unicaciones<=>sn<=>n<=>0=0.3 regulador<=>s.a<=>a<=>0=0.3 órgano<=>sn<=>n<=>0=0.3 presidente<=>suj<=>n<=>0=0.7 mostrar<=>v<=>mostrar.c2=1.0 partidario<=>cpred<=>a<=>1=0.7 información<=>sn<=>n<=>1=0.3 operador<=>sn<=>n<=>1=0.3 presidente<=>suj<=>n<=>arg1-tem

Tabla 4. Funciones características con valores literales y con pesos

Con estos valores característicos se han realizado varios experimentos de entrenamiento, trabajando con valores parciales del *corpus* (debido al factor tiempo). En el siguiente apartado podemos observar los resultados que se han obtenido de estas dos experimentaciones

3. Resultados obtenidos

Como podemos apreciar en las figuras 3 y 4, se han obtenido valores que aún están lejanos al baseline, sin embargo, si realizamos una reestimación de las funciones características y de los pesos, podremos seguir mejorando el porcentaje de error.

Otro punto muy importante a tener en cuenta, es que del total del corpus no se está utilizando todo el modelo, ya que se manejan únicamente 17.000 muestras de las más de 100.000 muestras disponibles.

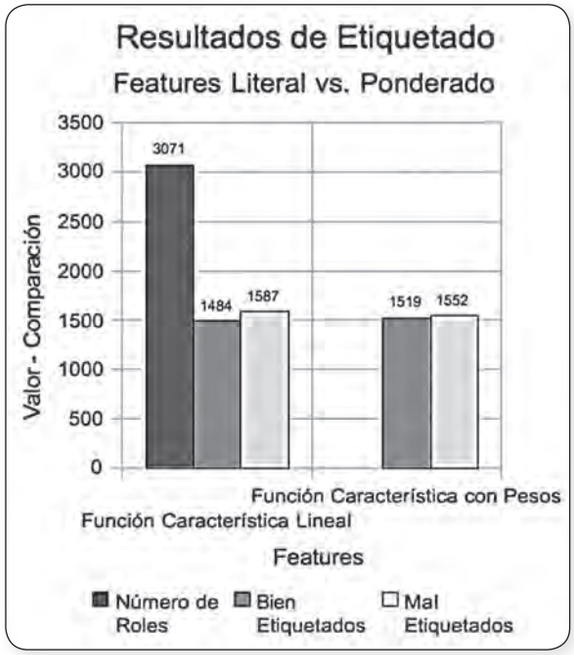

Figura 3. Resultados obtenidos, valores absolutos

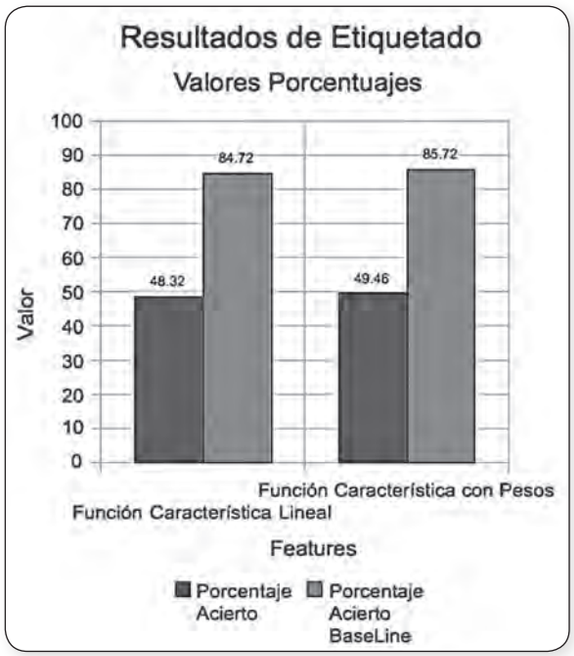

Figura 4. Resultados obtenidos, porcentajes

Figura 5. Resultados obtenidos distribución alfabética, valores totales

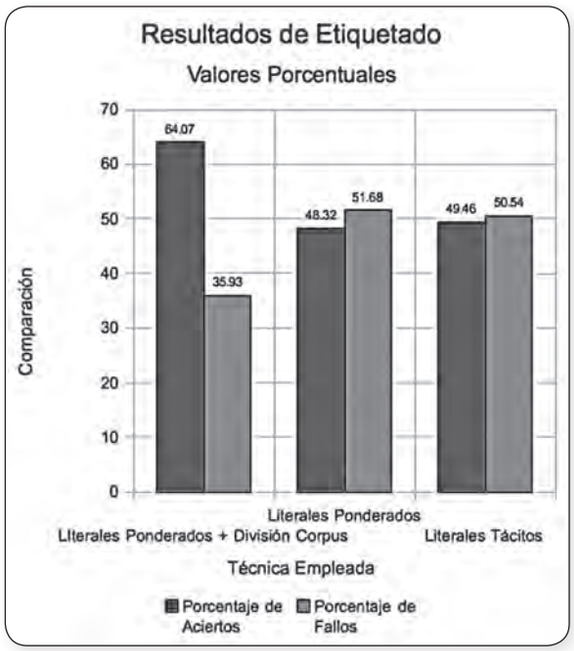

Figura 6. Resultados obtenidos distribución alfabética, valores totales

Como experimento final hemos decido crear un modelo de máxima entropía que trabaje por verbos ordenados alfabéticamente, esto es, un modelo para los verbos que inician con la letra 'a', otro para los que inician con la letra 'b', etcétera.

Con este esquema se han alcanzado los mejores resultados, llegando a obtener una tasa de aciertos del 64%. En la Figura 5 podemos observar los valores obtenidos.

4. Discusión

Luego de haber realizado este trabajo experimental, hemos llegado a las siguientes conclusiones:

- La elección de los elementos de la función característica es uno de los aspectos más importantes para la estructuración de un modelo de máxima entropía.
- Se ha logrado mejorar ligeramente el porcentaje de acierto al ponderar los constituyentes de los predicados.

- Al crear una partición de modelos basados en el tipo de verbo con el que se trabaja, la tasa de acierto sube en 16 puntos porcentuales, esto es debido a que el modelo se logra entrenar en su totalidad.
- La falta de un etiquetador de tipo B. I. O. impide que se logren extraer de mejor manera los elementos dependientes semánticamente del verbo.
- Se debería procurar manejar una mayor parte del corpus o su totalidad de ser posible, a fin de obtener una mejor tasa de aciertos.

4.1 Trabajo futuro

Como trabajo futuro podemos proponer los siguientes literales:

- Crear un modelo de entropía para cada verbo, siguiendo la estructura alfabética de ANCORA.
- Usar un etiquetador B. I. O. en la fase previa al entrenamiento, a fin de usar constituyentes de las cláusulas inmediatas y superiores.
- Utilizar predicados de mayor significación y carga semántica para entrenar los modelos.
- Revisar el uso de máquinas de soporte vectorial y comparar los resultados obtenidos con este tipo de herramientas.

Referencias bibliográficas

[1] Berger, A.L., Della Pietra, S. A., Della Pietra, V.J. (1996) *A Maximum Entropy Approach to Natural Language Processing*. IBM T. J. Watson Research Center.

[2] Choi, Y., Breck, E., Cardie, C. (2006) *Joint Extraction of Entities and Relations for Opinion Recognition*. Cornell University.

[3] C. Cortes, & Vapnik. 1995. *Support-Vector Networks*. Machine Learning.

[4] Cruz, F., (2005). *Etiquetado Estadístico de Roles Semánticos*. Universidad de Sevilla.

[5] Guildea, D., Jurafsky, D. (2002) *Automatic Labeling of Semantic Roles. In Computational Linguistics*.

[6] Zhang, L. (2004) *Maximum Entropy Modeling Toolkit for Python and C++*, December 2004.

[7] Malouf, R., (2002) *A comparison of algorithms for maximum entropy parameter estimation*. In Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002). 49-55.

[8] Navarro, B., Moreda, P., Fernández, B., Marcos, R., Palomar, M. *Anotación de roles semánticos en el corpus 3LB*. Universidades del País Vasco, de Barcelona, Politécnica de Cataluña, Politécnica de Valencia y de Alicante.

[9] Moreda, P. Mayo de 2008. *Los Roles Semánticos en la Tecnología del Lenguaje Humano: Anotación y Aplicación*. Tesis Doctoral, Universidad de Alicante.