# Language-Independent Named Entity Recognition

Aplicaciones de la Lingüística Computacional

Javier Pérez Afonso
Rodrigo Aldecoa García

14 de junio de 2011

## Introducción

El problema que nos ocupa consiste en reconocer los nombres propios (*Named Entities*, *NE*) que aparecen en un determinado texto. Las *NEs* pueden ser, por ejemplo, nombres de personas, organizaciones o lugares (países, ciudades, territorios...). La tarea que hemos tratado de resolver fue propuesta en 2002 en un conocido congreso de aprendizaje de lenguaje natural llamado *CoNLL* (*Conference on Computational Natural Language Learning*). Dicha tarea pretendía mejorar el reconocimiento automático de NEs independientemente del lenguaje en el que estuviese escrito un texto. Las NEs podían pertenecer a cuatro grupos: personas, lugares, organizaciones y otros nombres que no pudiesen ser incluidos en ninguno de estos tres primeros grupos. Para ello, la organización presentaba a los participantes corpus de entrenamiento y de test que debían utilizar para entrenar sus modelos mediante una aproximación de *machine learning*. En un principio, los corpus dados estaban escritos en dos idiomas, español y holandés, con la intención de que los participantes no utilizasen recursos idioma-específicos y que los modelos obtenidos se pudiesen aplicar posteriormente a textos escritos en cualquier idioma sin perder precisión.

### Datos

En nuestro trabajo hemos podido disponer de los corpus escritos en español y en holandés. Cada idioma consta de un corpus de entrenamiento y dos de test llamados *testa* y *testb*. El fichero *testa* debía ser utilizado para fijar unos buenos parámetros para el modelo de aprendizaje y el segundo era con el que se hacía la evaluación final. Los tres corpus están completamente anotados, para cada palabra se especifica su etiqueta que puede ser: *B-PER* en caso de *NEs* de personas, *B-LOC* para lugares, *B-ORG* para organizaciones y *B-MISC* para *NEs* que no se pueden clasificar en el resto de tipos. Las palabras que no son *NEs* tiene la etiqueta *O*.

Además también nos fue posible obtener el script en Perl que se utilizó en la prueba para determinar la calidad de cada uno de los sistemas de aprendizaje implementados por los distintos grupos. Su utilización es muy sencilla, se llama al script pasándole una solución de nuestro sistema para el conjunto de test (la asignación de etiquetas a cada palabra) y nos devuelve los siguientes parámetros: *precision*, *recall* y *F-score*. *Precision* en este caso representa el porcentaje de *NEs* que son correctamente clasificados del total de *NEs* que se han reconocido. El parámetro *recall*, por otro lado, es el porcentaje de *NEs correctamente* recuperados del total de *NEs* que existen en el texto. Para mostrar la calidad global de la clasificación obtenida partiendo de los valores obtenidos de *precision* y *recall*, podemos utilizar el valor de F-score, el cual viene dado por la siguiente fórmula: $F = 2 \times \frac{precision \times recall}{precision + recall}$

Podemos observar los resultados que obtuvieron en esta tarea los grupos que participaron en el *CoNLL-2002:*

| Spanish | precision | recall | F |
|---|---|---|---|
| CMP02 | 81.38 % | 81.40 % | 81.39 |
| Flo02 | 78.70 % | 79.40 % | 79.05 |
| CY02 | 78.19 % | 76.14 % | 77.15 |
| WNC02 | 75.85 % | 77.38 % | 76.61 |
| BHM02 | 74.19 % | 77.44 % | 75.78 |
| Tjo02 | 76.00 % | 75.55 % | 75.78 |
| PWM02 | 74.32 % | 73.52 % | 73.92 |
| Jan02 | 74.03 % | 73.76 % | 73.89 |
| Mal02 | 73.93 % | 73.39 % | 73.66 |
| Tsu02 | 69.04 % | 74.12 % | 71.49 |
| BV02 | 60.53 % | 67.29 % | 63.73 |
| MM02 | 56.28 % | 66.51 % | 60.97 |
| baseline | 26.27 % | 56.48 % | 35.86 |

| Dutch | precision | recall | F |
|---|---|---|---|
| CMP02 | 77.83 % | 76.29 % | 77.05 |
| Flo02 | 76.95 % | 73.83 % | 75.36 |
| CY02 | 75.10 % | 74.89 % | 74.99 |
| WNC02 | 72.69 % | 72.45 % | 72.57 |
| BHM02 | 73.03 % | 71.62 % | 72.31 |
| Tjo02 | 74.01 % | 68.90 % | 71.36 |
| PWM02 | 72.56 % | 68.88 % | 70.67 |
| Jan02 | 70.11 % | 69.26 % | 69.68 |
| Mal02 | 70.88 % | 65.50 % | 68.08 |
| Tsu02 | 57.33 % | 65.02 % | 60.93 |
| BV02 | 56.22 % | 63.24 % | 59.52 |
| MM02 | 51.89 % | 47.78 % | 49.75 |
| baseline | 81.29 % | 45.42 % | 58.28 |

El baseline fue producido por un sistema el cual sólo anota como Named Entities aquellas que aparecen completas (pueden ser varias palabras) en el corpus de entrenamiento.

# Tarea

Nuestro objetivo es reconocer automáticamente, a partir de un corpus de entrenamiento, el mayor número de NEs posibles de un corpus de test. Como en la asignatura hemos aprendido cómo se puede utilizar Python para el análisis de textos, decidimos generar un script en Python que etiquete automáticamente las distintas palabras teniendo en cuenta el corpus de entrenamiento y el nuestro conocimiento previo sobre las características que debe tener una *Named Entity*. Para ello utilizaremos una potente herramienta de la que dispone Python (y otros lenguajes de scripting) para el reconocimiento de patrones en texto escrito: las expresiones regulares.

En principio, aunque disponemos de él, no vamos a utilizar el corpus testb para crear las expresiones regulares. Plantearemos una primera aproximación para crear un modelo base, comprobaremos su eficiencia a la hora de reconocer *NEs* del corpus de test y a partir de ahí iremos añadiendo expresiones regulares para intentar mejorar los valores de *precision* y *recall* y, por tanto, los del parámetro *F*.

## Primera aproximación

Nuestro primer intento en el reconocimiento de *Named Entities* es similar al que denominan *baseline* en el congreso anteriormente citado (CoNLL-2002). En primer lugar analizamos el texto de entrenamiento y guardamos en una tabla hash (llamada diccionario en Python) para cada palabra, las categorías que se le asignan y el número de veces que se le asigna cada una de ellas. Llegados a este punto disponemos de un conjunto de palabras y las frecuencias con las que son asignadas a cada una de las categorías (esta frecuencia se calcula simplemente dividiendo el número de veces que aparece en una categoría entre el total de veces que aparece la categoría).

Para clasificar las palabras del conjunto de test, el criterio es el siguiente:

- Si la palabra está en nuestro diccionario, es decir, aparece en el conjunto de entrenamiento, la etiquetamos con la categoría de mayor frecuencia. En el caso de que varias categorías tengan la misma frecuencia, le asignamos una de ellas al azar.
- Si la palabra no ha aparecido todavía consideramos que no es una *Named Entity*, etiquetándola por tanto como *O*.

En esta primera aproximación, podemos utilizar el conjunto de test conocido (*testa*) para generar el diccionario, ya que existirán muchas palabras que no aparezcan en el conjunto de entrenamiento y puedan ser útiles para clasificar el corpus de test desconocido (*testb*). A continuación podemos ver los resultados con y sin el conjunto conocido de entrenamiento y la posición del ranking que ocuparían respecto a los métodos utilizados en el CoNLL-2002.

| Spanish | precision | recall | F |
|---|---|---|---|
| CMP02 | 81.38 % | 81.40 % | 81.39 |
| Flo02 | 78.70 % | 79.40 % | 79.05 |
| CY02 | 78.19 % | 76.14 % | 77.15 |
| WNC02 | 75.85 % | 77.38 % | 76.61 |
| BHM02 | 74.19 % | 77.44 % | 75.78 |
| Tjo02 | 76.00 % | 75.55 % | 75.78 |
| PWM02 | 74.32 % | 73.52 % | 73.92 |
| Jan02 | 74.03 % | 73.76 % | 73.89 |
| Mal02 | 73.93 % | 73.39 % | 73.66 |
| Tsu02 | 69.04 % | 74.12 % | 71.49 |
| BV02 | 60.53 % | 67.29 % | 63.73 |
| MM02 | 56.28 % | 66.51 % | 60.97 |
| sinTest | 56.05 % | 66.58 % | 60.86 |
| conTest | 55.66 % | 66.18 % | 60.46 |
| baseline | 26.27 % | 56.48 % | 35.86 |

| Dutch | precision | recall | F |
|---|---|---|---|
| CMP02 | 77.83 % | 76.29 % | 77.05 |
| Flo02 | 76.95 % | 73.83 % | 75.36 |
| CY02 | 75.10 % | 74.89 % | 74.99 |
| conTest | 72.97 % | 78.92 % | 74.74 |
| sinTest | 70.59 % | 79.29 % | 74.69 |
| WNC02 | 72.69 % | 72.45 % | 72.57 |
| BHM02 | 73.03 % | 71.62 % | 72.31 |
| Tjo02 | 74.01 % | 68.90 % | 71.36 |
| PWM02 | 72.56 % | 68.88 % | 70.67 |
| Jan02 | 70.11 % | 69.26 % | 69.68 |
| Mal02 | 70.88 % | 65.50 % | 68.08 |
| Tsu02 | 57.33 % | 65.02 % | 60.93 |
| BV02 | 56.22 % | 63.24 % | 59.52 |
| MM02 | 51.89 % | 47.78 % | 49.75 |
| baseline | 81.29 % | 45.42 % | 58.28 |

## NEs compuestas

Tras estudiar brevemente la composición del corpus de entrenamiento, vemos como muchas de las NEs que aparecen en el texto constan de varias palabras, por ejemplo: *Abogado General del Estado*, *Antena-3 Televisión* o *Naciones Unidas*. También hemos observado cómo muchas de esas palabras no se pueden reconocer como NEs mediante nuestra aproximación anterior: *Televisión* es etiquetado como O, es decir, se considera que no es ningún nombre propio.

Para resolver este tipo de problemas, hemos recorrido el corpus de test y, cada palabra que comienza en mayúscula cuya palabra anterior esté anotada como una NE, se etiqueta también como NE con esa misma categoría. En las palabras anteriores, si se detecta que *Televisión* no está etiquetada como NE pero está en mayúscula y su palabra anterior está clasificada como *B-ORG* (nombre propio de organización), se reclasifica como NE de esa clase *B-ORG*. Como se puede apreciar en las siguientes tablas, el método da muy buen resultado y mejora cualitativamente sobre las soluciones que obteníamos anteriormente:

| Spanish | precision | recall | F |
|---|---|---|---|
| CMP02 | 81.38 % | 81.40 % | 81.39 |
| Flo02 | 78.70 % | 79.40 % | 79.05 |
| CY02 | 78.19 % | 76.14 % | 77.15 |
| WNC02 | 75.85 % | 77.38 % | 76.61 |
| BHM02 | 74.19 % | 77.44 % | 75.78 |
| Tjo02 | 76.00 % | 75.55 % | 75.78 |
| PWM02 | 74.32 % | 73.52 % | 73.92 |
| Jan02 | 74.03 % | 73.76 % | 73.89 |
| Mal02 | 73.93 % | 73.39 % | 73.66 |
| Tsu02 | 69.04 % | 74.12 % | 71.49 |
| conTest | 64.74 % | 70.62 % | 67.56 |
| sinTest | 64.47 % | 70.77 % | 67.47 |
| BV02 | 60.53 % | 67.29 % | 63.73 |
| MM02 | 56.28 % | 66.51 % | 60.97 |
| baseline | 26.27 % | 56.48 % | 35.86 |

| Dutch | precision | recall | F |
|---|---|---|---|
| sinTest | 80.25 % | 82.42 % | 81.32 |
| conTest | 80.69 % | 81.87 % | 81.27 |
| CMP02 | 77.83 % | 76.29 % | 77.05 |
| Flo02 | 76.95 % | 73.83 % | 75.36 |
| CY02 | 75.10 % | 74.89 % | 74.99 |
| WNC02 | 72.69 % | 72.45 % | 72.57 |
| BHM02 | 73.03 % | 71.62 % | 72.31 |
| Tjo02 | 74.01 % | 68.90 % | 71.36 |
| PWM02 | 72.56 % | 68.88 % | 70.67 |
| Jan02 | 70.11 % | 69.26 % | 69.68 |
| Mal02 | 70.88 % | 65.50 % | 68.08 |
| Tsu02 | 57.33 % | 65.02 % | 60.93 |
| BV02 | 56.22 % | 63.24 % | 59.52 |
| MM02 | 51.89 % | 47.78 % | 49.75 |
| baseline | 81.29 % | 45.42 % | 58.28 |

## Otras reglas

Con el objetivo de ir mejorando poco a poco los tres parámetros que miden la calidad de la clasificación (*precision*, *recall* y *F-score*), hemos incluído también una serie de refinamientos de la clasificación obtenida hasta el momento:

- Todas las palabras que comienzan por minúscula se reclasifican como *O*. Es decir, ninguna de ella será considerada como una NE.
- En muchos de los lenguajes conocidos las palabras muy habituales están formadas por pocas letras. Dado que teníamos problemas con algunas de estas palabras, por ejemplo *del* en *Abogado General del Estado* (hay que considerarla como NE de persona), decidimos que: a las palabras menores de 5 letras que se encuentren entre dos NEs del mismo tipo, se les asigna ese tipo.

La primera de las reglas fue la mejor a la hora de asignar correctamente un mayor número de NEs. La segunda sin embargo no lo fue tanto, pero sí que conseguimos mejorar unas décimas. El resultado de la aplicación de estas dos reglas fue el siguiente:

| Spanish | precision | recall | F |
|---|---|---|---|
| CMP02 | 81.38 % | 81.40 % | 81.39 |
| Flo02 | 78.70 % | 79.40 % | 79.05 |
| CY02 | 78.19 % | 76.14 % | 77.15 |
| WNC02 | 75.85 % | 77.38 % | 76.61 |
| BHM02 | 74.19 % | 77.44 % | 75.78 |
| Tjo02 | 76.00 % | 75.55 % | 75.78 |
| PWM02 | 74.32 % | 73.52 % | 73.92 |
| Jan02 | 74.03 % | 73.76 % | 73.89 |
| Mal02 | 73.93 % | 73.39 % | 73.66 |
| Tsu02 | 69.04 % | 74.12 % | 71.49 |
| conTest | 69.48 % | 70.61 % | 70.04 |
| sinTest | 69.39 % | 70.67 % | 70.02 |
| BV02 | 60.53 % | 67.29 % | 63.73 |
| MM02 | 56.28 % | 66.51 % | 60.97 |
| baseline | 26.27 % | 56.48 % | 35.86 |

| Dutch | precision | recall | F |
|---|---|---|---|
| sinTest | 82.97 % | 81.43 % | 82.19 |
| conTest | 83.08 % | 80.89 % | 81.97 |
| CMP02 | 77.83 % | 76.29 % | 77.05 |
| Flo02 | 76.95 % | 73.83 % | 75.36 |
| CY02 | 75.10 % | 74.89 % | 74.99 |
| WNC02 | 72.69 % | 72.45 % | 72.57 |
| BHM02 | 73.03 % | 71.62 % | 72.31 |
| Tjo02 | 74.01 % | 68.90 % | 71.36 |
| PWM02 | 72.56 % | 68.88 % | 70.67 |
| Jan02 | 70.11 % | 69.26 % | 69.68 |
| Mal02 | 70.88 % | 65.50 % | 68.08 |
| Tsu02 | 57.33 % | 65.02 % | 60.93 |
| BV02 | 56.22 % | 63.24 % | 59.52 |
| MM02 | 51.89 % | 47.78 % | 49.75 |
| baseline | 81.29 % | 45.42 % | 58.28 |

### Reglas testadas que no funcionaron

- Asignar la categoría *B-ORG* a aquellas palabras que fuesen todo mayúsculas. La mayoría de veces podrían corresponder a siglas y lo más probable es que fuese una organización. Es posible que este método ante otro texto hubiese mejorado la clasificación, sin embargo aquí empeoraba: algunas frases de los textos eran todo mayúsculas, etiquetaba todo ello como NEs de organizaciones y por tanto el error crecía considerablemente.

- Asignar categorías de NEs aleatorias a palabras que comiencen en mayúscula y no estén a principio de línea. Al aplicar este método el error aumentaba dramáticamente ya que no sólo es importante reconocer si son NEs o no, sino también clasificarlas según el tipo al que corresponden.

## Conclusiones

Nuestra aproximación parece funcionar razonablemente bien si la comparamos con las de los grupos que se presentaron en el *CoNLL2002*. Por supuesto son dos tipos de clasificaciones totalmente distintas y nosotros partíamos con la clara ventaja del diccionario creado por los conjuntos de entrenamiento. Las frecuencias de las categorías de cada palabra han sido muy útiles a la hora de realizar una correcta clasificación.

Una pregunta obvia que le surgirá al lector es la enorme diferencia entre la bondad de las clasificaciones del español y el holandés. Nuestra hipótesis era que esa diferencia era causada porque las palabras en español eran más ambiguas, es decir, podían pertenecer, de media, a más categorías que las del holandés. Tras un simple análisis pudimos comprobar que íbamos por el camino correcto. Mientras que en el español una palabra pertenecía a 1.17 categorías,

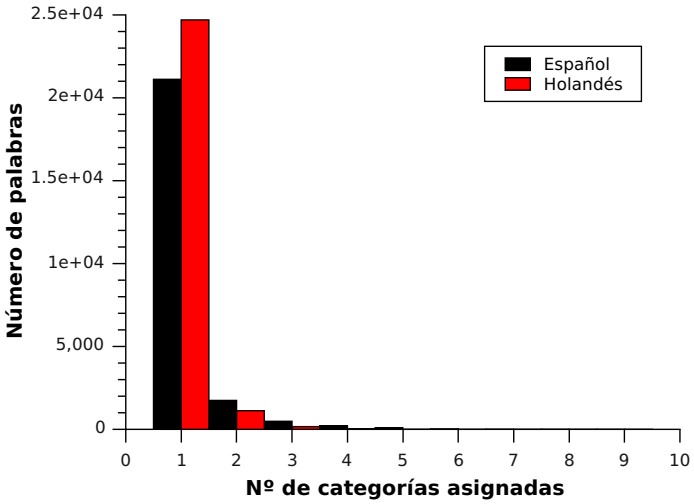


Figura 1: Número de palabras por número de categorías

en el holandés únicamente a 1.06. Lo que pueden parecer unos pocos decimales, puede distorsionar cualitativamente una correcta clasificación. En la figura 1 podemos ver el número de palabras que pertenecen a cada número de categorías.

Por tanto, aunque pueda haber alguna otra razón subyacente para que nuestros métodos sean muy superiores al clasificar palabras holandesas, podemos asumir que la mayor ambigüedad del español hace que la clasificación mediante frecuencias de categorías obtenidas gracias a un conjunto de entrenamiento sea la causante de dicho efecto.