MEMORIA PRÁCTICA 4

COMPUTACIÓN INTELIGENTE Y LENGUAJE NATURAL

Melanie Torres Bisbal 146683
Roser Pruaño Milla 158688
Pedro Vílchez Blanco 148574

Resultatos del etiquetaje basado en unigramas del primer fichero.

Después de la realización del Modelo de Lenguaje (lexic.txt) a través del fichero corpus.txt, las comparaciones entre el primer test y nuestro primer etiquetado es el siguiente:

Salida del programa:

fichero1, aciertos: 0.88 %

El porcentaje de aciertos se encuentra en el intervalo [0,1] con lo cual vemos que nuestra tasa de aciertos es bastante elevada con nuestro Modelo de Lenguaje.
El problema del 0,12% viene dado a que hay ambigüedades que no podemos romper de manera correcta debido a la falta de información. Esta información creemos que se basa especialmente en el contexto, el cual no tenemos en cuenta en nuestro etiquetado.

Problemas hallados.

Veamos a continuación un extracto de las diferencias encontradas en ambos ficheros:

Nota aclaratoria: El primer [palabra-etiqueta] corresponde a nuestro etiquetado. El segundo corresponde al archivo con el que comparamos.

('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'primera', 'NC', '] vs [', 'primera', 'Adj', ']')
('difference: [', 'las', 'Pron', '] vs [', 'las', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'una', 'Pron', '] vs [', 'una', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'También', 'NP', '] vs [', 'También', 'Adv', ']')
('difference: [', 'ha', 'NP', '] vs [', 'ha', 'VAux', ']')
('difference: [', 'anunciada', 'V', '] vs [', 'anunciada', 'Adj', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'varios', 'Pron', '] vs [', 'varios', 'Det', ']')
('difference: [', 'latinoamericanos', 'NC', '] vs [', 'latinoamericanos', 'Adj', ']')
('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'Además', 'NP', '] vs [', 'Además', 'Adv', ']')
('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')
('difference: [', 'cubanos', 'NC', '] vs [', 'cubanos', 'Adj', ']')

('difference: [', 'ganadora', 'NC', '] vs [', 'ganadora', 'Adj', ']')

Lo que hemos podido observar es que muchos de los fallos al etiquetar son debidos a que no hacemos diferenciación de las categorías gramaticales de las palabras según su contexto.

Así pues, es fácil que el programa se equivoque al etiquetar palabras como “la, los”, ya que éstas pueden actuar como pronombre o determinante según la palabra a la cual acompañe. Incluso “la” podría hacer referencia a la nota musical, siendo entonces un nombre.

Este fallo corresponde a las dos primeras líneas remarcadas:

('difference: [', 'la', 'Pron', '] vs [', 'la', 'Det', ']')
('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')

Si nos basamos en el mismo tipo de fallo, encontramos que el programa nuevamente no hace diferencia entre palabras que pueden actuar como nombres comunes o adjetivos ya que no ve si viene precedido por un determinante o no o si acompaña a un nombre.

Esto provoca fallos tales como el encontrado en la quinta línea remarcada:

('difference: [', 'latinoamericanos', 'NC', '] vs [', 'latinoamericanos', 'Adj', ']')

Lo mismo sucede con algunas formas verbales dónde, al no ver si la palabra viene acompañada de una forma verbal auxiliar, ésta puede ser tratada de adjetivo o verbo según los datos proporcionados por el lexic.txt.

Esto conlleva que encontremos etiquetados como el que se ve en la cuarta línea remarcada:

('difference: [', 'anunciada', 'V', '] vs [', 'anunciada', 'Adj', ']')

Otro fallo al etiquetar bastante curioso es al tratar con palabras que comienzan por mayúsculas.
Aquí el programa lo que hace es etiquetarlas como nombres propios en vez de la categoría gramatical que le corresponde, cosa que se podría evitar si viéramos si hay otras palabras del contexto que las determinan como nombre o no.

Esto se corresponde con la tercera línea remarcada del extracto:

('difference: [', 'También', 'NP', '] vs [', 'También', 'Adv', ']')

Solución propuesta.

Para evitar este problema al etiquetar basándonos en el uso de la lengua, lo que se podría hacer es que, a medida que se hace el etiquetado, tener en cuenta las etiquetas anteriores y analizar la probabilidad más elevada de que se dé la siguiente etiqueta viendo todo lo acumulado.
Así conseguiríamos que muchos de los errores comentados anteriormente pudieran observar su contexto y añadir la etiqueta correcta.

Aún así es posible que algunas palabras siguieran mal etiquetadas ya que el uso de la lengua castellana es bastante libre, con lo cual el sentido de las palabras puede determinarse tanto por las palabras que la preceden con las que la proceden. Esto nos llevaría a especializar más el programa en cuanto al análisis del contexto, haciendo que tuviera en cuenta un marco de palabras a la izquierda y a la derecha de la palabra a etiquetar en cuestión. Esto se podría llevar a cabo en una doble ejecución del programa, la primera siendo un etiquetado viendo las etiquetas anteriores; y una segunda viendo el contenido general de la frase y analizando el marco alrededor de susodicha palabra.

A pesar de que el etiquetado mejoraría considerablemente, nos surgiría un nuevo problema. Este problema que sería que, al implementar todo esto, la cantidad de cálculos e información que debemos ir guardando a medida que avanzamos en el fichero con la ejecución del programa crecería de manera muy rápida y la ejecución sería muchísimo más lenta y costosa.

Resultatos del etiquetaje basado en unigramas del segundo fichero.

Después de la realización del Modelo de Lenguaje (lexic.txt) a través del fichero corpus.txt, las comparaciones entre el segundo test y nuestro segundo etiquetado es el siguiente:

Salida del programa:

fichero2, aciertos: 0.83 %

El porcentaje de aciertos se encuentra en el intervalo [0,1] con lo cual vemos que nuestra tasa de aciertos es bastante elevada con nuestro Modelo de Lenguaje.

Esta vez vemos que nuestro etiquetado se equivoca en un 0,17%, valor superior al anterior test.
Teniendo en cuenta que ambos tests son diferentes es altamente probable que este segundo contenga más palabras ambiguas que el anterior. Esto conllevaría a que el fallo en el etiquetado aumentara respecto al otro.

Así pues, nuevamente el problema del 0,17% viene dado a que hay ambigüedades que no podemos romper de manera correcta debido a la falta de información. Esta información creemos que se basa especialmente en el contexto, el cual no tenemos en cuenta en nuestro etiquetado.

Problemas hallados.

Veamos a continuación un extracto de las diferencias encontradas en ambos ficheros:

Nota aclaratoria: El primer [palabra-etiqueta] corresponde a nuestro etiquetado.
El segundo corresponde al archivo con el que comparamos.

[La, Pron] vs [La, Det]
[catalán, NC] vs [catalán, Adj]
[exterior, NC] vs [exterior, Adj]
[los, Pron] vs [los, Det]
[una, Pron] vs [una, Det]
[imcombustible, Desc] vs [imcombustible, NC]
[final, NC] vs [final, Adj]
[El, NP] vs [El, Det]
[los, Pron] vs [los, Det]
[once, NC] vs [once, Det]
[partido, V] vs [partido, NC]
[todo, Pron] vs [todo, Det]
[mucho, Det] vs [mucho, Adv]
[La, Pron] vs [La, Det]
[La, Pron] vs [La, Det]
[personal, NC] vs [personal, Adj]
[la, Pron] vs [la, Det]
[una, Pron] vs [una, Det]
[equilibraron, Desc] vs [equilibraron, V]

[los, Pron] vs [los, Det]
[once, NC] vs [once, Det]
[la, Pron] vs [la, Det]
[23-12, Desc] vs [23-12, Num]
[24-16, Desc] vs [24-16, Num]
[Massenburg_por_Roberts, Desc] vs [Massenburg_por_Roberts, NP]
[El, NP] vs [El, Det]
[local, NC] vs [local, Adj]
[36-22, Desc] vs [36-22, Num]
[los, Pron] vs [los, Det]
[Ricardo_Pittis, Desc] vs [Ricardo_Pittis, NP]
[los, Pron] vs [los, Det]
[local, NC] vs [local, Adj]
[la, Pron] vs [la, Det]
[diez, NC] vs [diez, Det]
[40-30, Desc] vs [40-30, Num]
[El, NP] vs [El, Det]
[cuatro, NC] vs [cuatro, Det]
[bien, NC] vs [bien, Adv]
[la, Pron] vs [la, Det]
[la, Pron] vs [la, Det]
[la, Pron] vs [la, Det]

Lo que hemos podido observar es que muchos de los fallos al etiquetar son debidos a que no hacemos diferenciación de las categorías gramaticales de las palabras según su contexto.

Nuevamente, es fácil que el programa se equivoque al etiquetar palabras como “la, los,una”, ya que éstas pueden actuar como pronombre o determinante según la palabra a la cual acompañe. Incluso “la” podría hacer referencia a la nota musical, siendo entonces un nombre.

Este fallo corresponde a las tres primeras líneas remarcadas:

('difference: [', 'La', 'Pron', '] vs [', 'La', 'Det', ']')
('difference: [', 'los', 'Pron', '] vs [', 'los', 'Det', ']')
('difference: [', 'una', 'Pron', '] vs [', 'una', 'Det', ']')

Si nos basamos en el mismo tipo de fallo, encontramos que el programa nuevamente no hace diferencia entre palabras que pueden actuar como nombres comunes o adjetivos ya que no ve si viene precedido por un determinante o no o si acompaña a un nombre. También hay error de etiquetaje cuando se trata de nombres comunes y determinantes.

Esto provoca fallos tales como el encontrado en la cuarta y sexta línea remarcada:

('difference: [', 'final', 'NC', '] vs [', 'final', 'Adj', ']')
('difference: [', 'once', 'NC', '] vs [', 'once', 'Det', ']')

Lo mismo sucede con algunas formas verbales dónde, al no ver si la palabra viene acompañada de una forma verbal auxiliar, ésta puede ser tratada de nombre común o

verbo según los datos proporcionados por el lexic.txt. Como hemos visto en el caso del primer test, esto también podría suceder entre verbos y adjetivos.

Esto conlleva que encontremos etiquetados como el que se ve en la séptima línea remarcada:

('difference: [', 'partido', 'V', '] vs [', 'partido', 'NC', ']')

De nuevo el etiquetar palabras que comienzan por mayúsculas suponen un problema.
Aquí el programa lo que hace es etiquetarlas como nombres propios en vez de la categoría gramatical que le corresponde, cosa que se podría evitar si viéramos si hay otras palabras del contexto que las determinan como nombre o no.

Esto se corresponde con la quinta línea remarcada del extracto:

('difference: [', 'El', 'NP', '] vs [', 'El', 'Det', ']')

Otro problema hallado, que no hemos podido observar en el extracto del test anterior, es el hecho de que contenga palabras que no existen en el corpus. Esto conlleva a que dichas palabras no hayan sido etiquetadas y, por lo tanto, cuando aparecen en un texto a etiquetar se les ponga una etiqueta errónea ya que no la encuentra.

Vemos este caso en las tres últimas líneas remarcadas:

('difference: [', '23-12', 'Desc', '] vs [', '23-12', 'Num', ']')
('difference: [', '24-16', 'Desc', '] vs [', '24-16', 'Num', ']')
('difference: [', 'Massenburg_por_Roberts', 'Desc', '] vs [',
'Massenburg_por_Roberts', 'NP', ']')

Solución propuesta.

Como en el caso del fichero anterior, para evitar este problema podríamos tener en cuenta las etiquetas anteriores y analizar la probabilidad más elevada de que se dé la siguiente etiqueta viendo todo lo acumulado.
Así conseguiríamos que muchos de los errores comentados anteriormente pudieran observar su contexto y añadir la etiqueta correcta.

Como hemos comentado anteriormente, debido al uso de la lengua castellana, es posible que aún resten palabras mal etiquetadas debido a que no se tienen en cuenta las palabras que preceden la palabra a etiquetar ni las que la proceden. Esto nos llevaría a especializar más el programa en cuanto al análisis del contexto, haciendo que tuviera en cuenta un marco de palabras a la izquierda y a la derecha de la palabra a etiquetar en cuestión.
Como hemos dicho, esto se podría llevar a cabo en una doble ejecución del programa, la primera siendo un etiquetado viendo las etiquetas anteriores; y una segunda viendo el contenido general de la frase y analizando el marco alrededor de susodicha palabra.

Además, hemos visto que deberíamos mejorar el programa haciendo que tuviera en cuenta otros rasgos de la lengua castellana como el uso de caracteres especiales (“/”, “-”, etc) y así poder asignar apropiadamente las etiquetas a las palabras que las usan o contienen y, además, saber tratar las palabras que no existen en el corpus de manera general. En referencia a esto último, nosotros hemos decidido utilizar la etiqueta 'Desc' para mostrar aquellas palabras que no han podido ser etiquetadas, en vez de buscar la etiqueta adecuada.

Al añadir esta nueva mejora, esto hace que el programa realizaría más cálculos y movería más información que en el test anterior, con lo cual seguiríamos teniendo el mismo problema que antes y puede que agravado. pesar de que el etiquetado mejoraría considerablemente, nos surgiría un nuevo problema. Este problema que sería que, al implementar todo esto, la cantidad de cálculos e información que debemos ir guardando a medida que avanzamos en el fichero con la ejecución del programa crecería de manera muy rápida y la ejecución sería muchísimo más lenta y costosa.