Memoria Trabajo Lingüística Computacional

Miguel This is a Placeholder Roberto Labadie Tamayo

Octubre 2022

Tarea I

Para evaluar la tarea de etiquetado morfosintágtico empleando un HMM entrenado sobre el corpus **cess-esp** se estudian dos variantes, una sobre una versión reducida del problema con un conjunto de categorías mas abarcadoras y la versión original usando las etiquetas tal cual aparecen en el corpus.
A través de una validación cruzada con 10 particiones del dataset se obtienen los resultados como se muestran en la Figura 1.

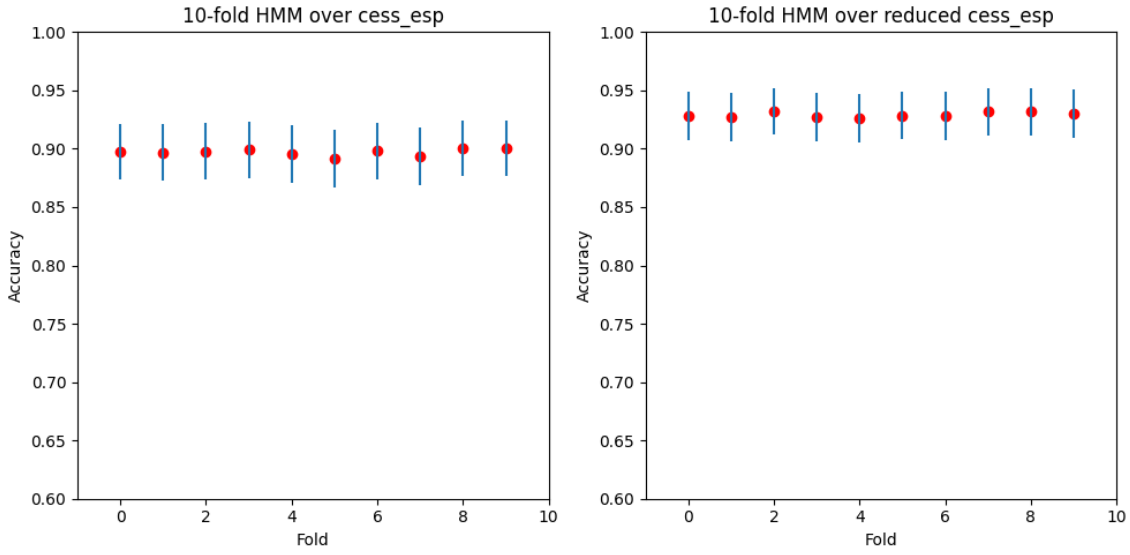

Figura 1: Hidden Markov Model sobre cess-esp (izquierda) y cess-esp reducido (derecha)

Como se puede apreciar el desempeño del modelo es mejor para el dataset de la tarea reducida gracias a que la cantidad de estados ocultos del modelo es menor por tanto la posibilidad de escoger una transición errada disminuye, esta observación se puede realizar incluyendo además los intervalos de un un 95 % de confianza estimados. Nótese ademas que en ambos modelos la $i - esima$ partición es exactamente la misma, i.e., el mismo conjunto de elementos ordenados en ambos modelos.
En la curva anterior se muestran los resultados para modelos de 1-gramas hasta 4-gramas sobre el test set, como es intuitivo modelos con construcciones más simples de elementos atómicos (gramas) tienden a presentar una distribución más suavizada y uniforme entre los elementos del vocabulario con respecto a construcciones más complejas (4-gramas y 5-gramas). De los modelos explorados el más robusto resulto el de 4-gramas, con una preplejidad de ~ 7.22.

cess-esp	**reduced cess-esp**
0.897±0.02	0.928±0.02
0.897±0.02	0.927±0.02
0.898±0.02	0.932±0.02
0.899±0.02	0.928±0.02
0.896±0.02	0.927±0.02
0.892±0.02	0.928±0.02
0.898±0.02	0.928±0.02
0.893±0.02	0.932±0.02
0.900±0.02	0.932±0.02
0.900±0.02	0.930±0.02

Cuadro 1: Perplejidad sobre modelos de 3-gramas y 4-gramas

Tarea II

De los resultados anteriores, resulta congruente pensar que empleando modelos de 3-gramas o 4-gramas podrían arrojar los mejores resultados, explorando distintos métodos de descuento como se muestra en la Figura 2 curva.
Nuevamente, para todos los casos los mejores resultados se obtienen para los modelos más complejos de 4-gramas, sin embargo el método no modificado de Kneser-Ney mejora los valores de perplejidad con respecto a las otras variantes de descuento hasta el valor de 7.04 en 4-gramas y 7.65 para 3-gramas.

Tarea III

Al sustituir el suavizado mediante back-off por un esquema de interpolación sobre los descuentos de Witten-Bell y Kneser-Ney modificado, los resultados obtenidos en términos de perplejidad se muestran en el Cuadro 2

Descuento	3-gram		4-gram	
	Back-off	*Interpol.*	*Back-off*	*Interpol.*
Witten-Bell	7.83	**7.36**	7.15	**6.58**
Kneser-Ney	8.32	**7.70**	7.96	**7.02**

Cuadro 2: Perplejidad sobre modelos de 3-gramas y 4-gramas

Donde como se puede observar en cada caso los valores son mejorados tanto para los modelos de 3-gramas como para 4-gramas, donde se obtienen los mejores resultados.

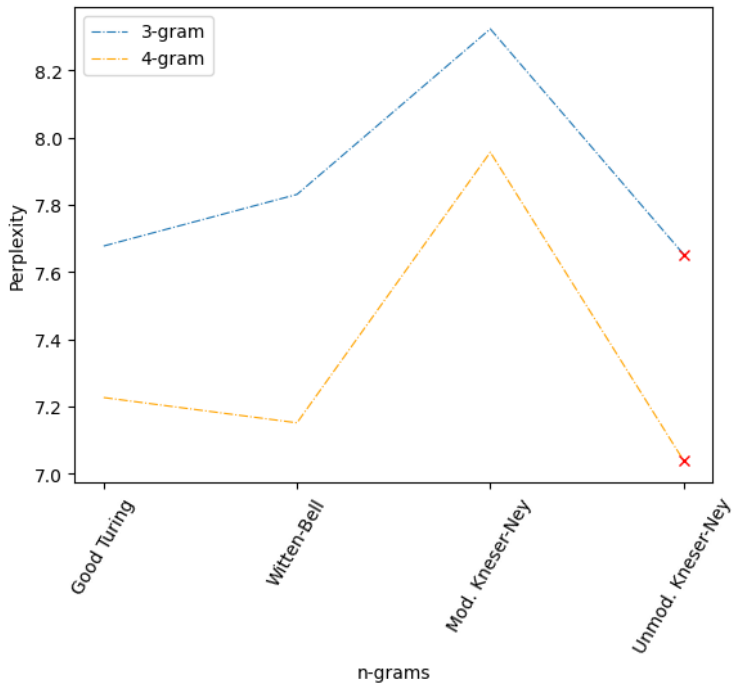

Figura 2: Resultados de Perplejidad sobre Dihana test set para modelos de 1-gramas a 4-gramas y distintos métodos de descuento .

Tarea IV

Para estudiar como impacta el uso de conjuntos léxicos con distintas cargas semánticas, se analiza el desempeño de modelos de 3-gramas y 4-gramas bajo las mismas condiciones de la tarea I, i.e., descuento de Good-Turing y suavizado de back-off empleando el conjunto de datos Europarl. Este estudio consiste simplemente en remover del vocabulario términos de forma escalonada que aparecen 1, 5 o menos y 9 o menos veces.
Como se puede observar, modelos de 4-gramas poseen un mejor comportamiento en todas las variaciones. Para el valor de 0 (no eliminando palabra alguna del vocabulario) se obtienen los resultados más altos de perplejidad lo cual es coherente con el hecho de que elementos más uniformemente distribuidos poseen menor entropía y por tanto perplejidad. Esta condición de uniformidad se aproximará mientras nos acercamos al valor del 9 ya que se eliminan términos extremos. Por otro lado, términos menos frecuentes dentro de un corpus son en gran parte aquellos que caracterizan el lenguaje de manera específica y por tanto lo complejizan de ahí gran parte de este decrecimiento.

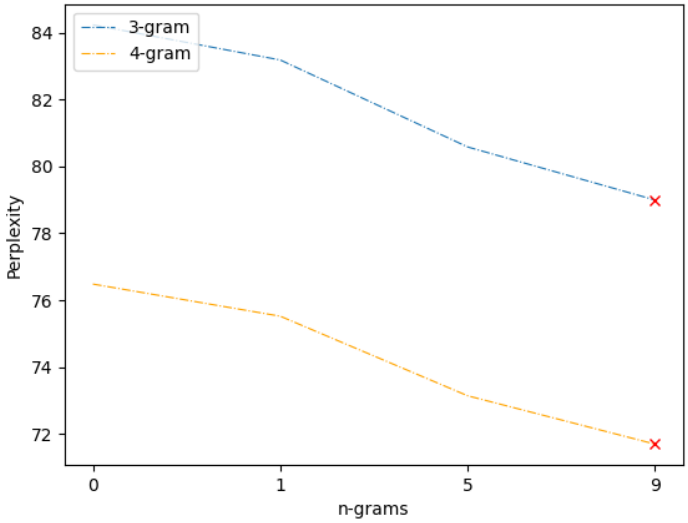

Figura 3: Resultados de Perplejidad sobre reducciones de vocabulario para dataset Europal en modelos de 3-gramas y 4-gramas.