Trabajo LC

Alejandro Granados Bañuls <algrabau@inf.upv.es>

Introducción

En este proyecto se han realizado varias ejecuciones del software SRLIM para la generación de un modelo de lenguaje del corpus DIHANA probando diferentes métodos para la genración de dicho modelo de lenguaje. Acto seguido se discutirán los resultados de los diferentes métodos para evaluar cuál es el más apropiado para el problema.

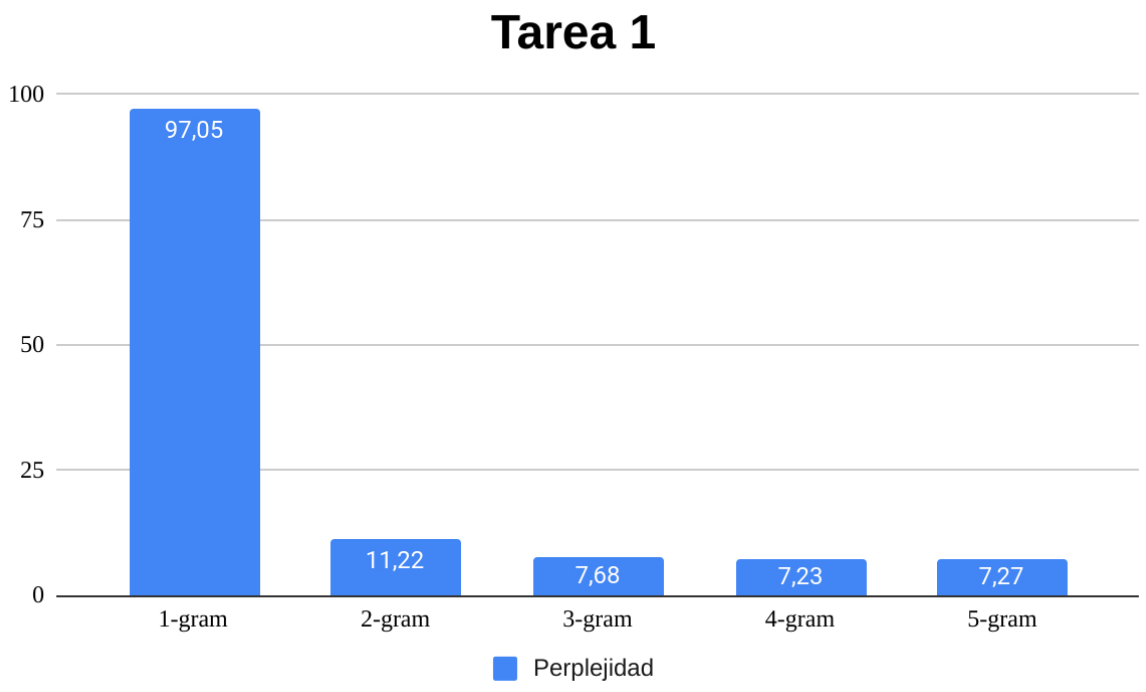
También se generará un modelo de lenguaje para el corpus Europarl para estimar el impacto de eliminar palabras del vocabulario según su frecuencia de aparición.

A continuación se va a mostrar los resaltudos de las ejecuciones y se van a comentar los resultados por tarea. Finalmente se comentara la conclusión que se obtiene a partir de la experimentación.

Tarea 1

	1-gram	2-gram	3-gram	4-gram	5-gram
Perplejidad	97,05	11,22	7,68	7,23	7,27

Tabla 1: Resultados tarea 1



Gráfica 1: Resultados tarea 1

A la vista de los resultados de a *Gráfica 1*, se puede observar que la perplejidad se reduce según se aumenta el parámetro de los n-gramas. Aunque esta tendencia parece reducirse a partir del parámetro 5. Esto puede deberse a una sobregeneralización a la hora de generar

el modelo de lenguaje, así que se descartará por el momento está solución, ya que tiene valores muy similares a los trigramas (3-gram) y a los 4-grams.

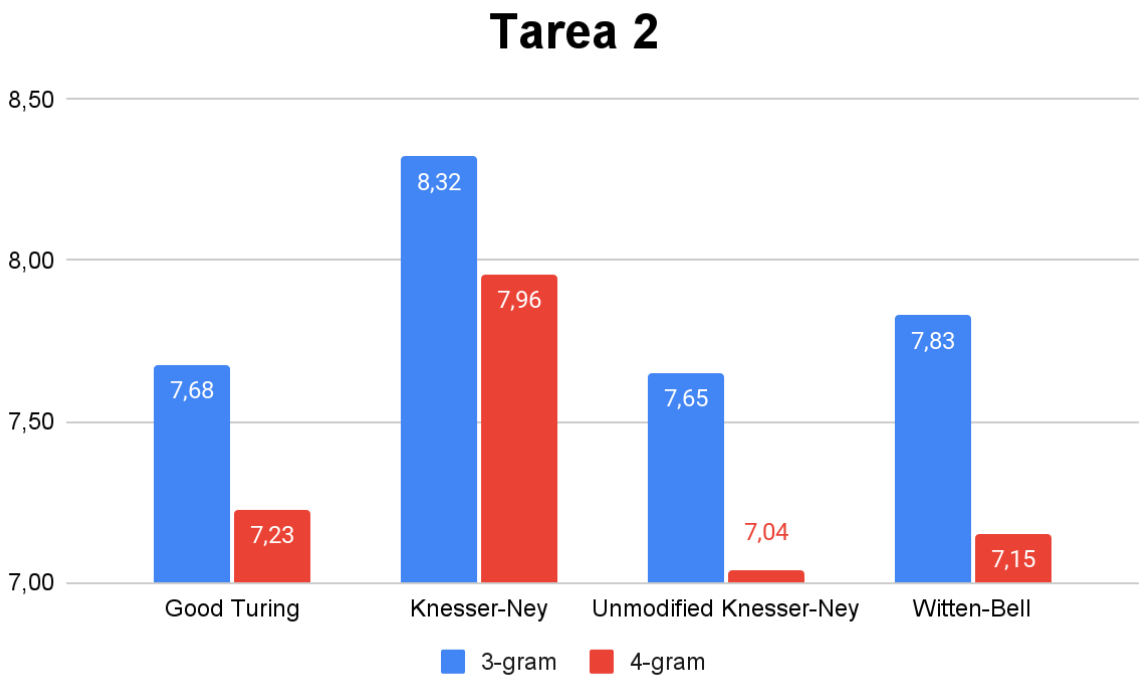
Respecto a la perplejidad, se puede afirmar que la capacidad de generalziación del modelo de lenguaje basado en unigramas (1-gram) no está dando buenos resultados. Cuando se pasa a un modelo de bigramas (2-gram) la perplejidad baja drástcamente, con lo que se puede concluir que la generalización con bigramas es mejor con unigramas.

En el caso de trigramas y 4-gram, da valores muy parecidos, así que tal vez sería preferible usar un modelo basado en trigramas que en 4-grams por el hecho de que en el caso de los 4-grams la complejidad sería mucho mayor que en el caso de los trigramas.

Tarea 2

	Good Turing	Knesser-Ney	Unmodified Knesser-Ney	Witten-Bell
3-gram	7,68	8,32	7,65	7,83
4-gram	7,23	7,96	7,04	7,15

Tabla 2: Resultados tarea 2



Gráfica 2: Resultados de la tarea 2

En este caso se han comparado varios métodos de descuento a la hora de generar el modelo de lenguaje con trigramas y 4-grams (que son los parámetros que mejor resultado han dado en la tarea anterior). Es necesario recalcar que la ejecución de la tarea anterior tenía cómo descuento Good Turing, por eso los valores son los mismos.

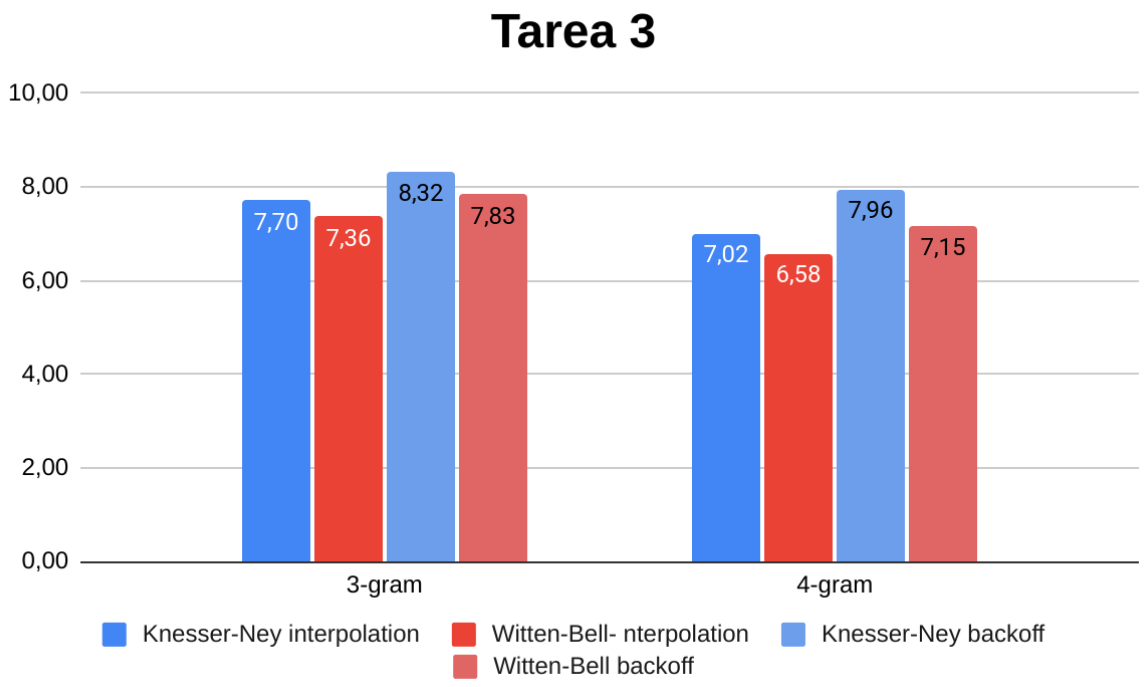
A partir del gráfico anterior se puede concluir que el mejor descuento para el actual problema es el *Unmodified Knesser-Ney*, ya que se concluye que tanto para trigramas como para 4-grams propoirciona mejores resultados.

A parte de haber elegido el descuento de *Unmodified Knesser-Ney*, se podría concluir que sería mejor usar 4-grams que trigramas (siempre que fuera posible y asequible el tratar con la complejidad de este caso), pues la diferencia con el caso de trigramas es relativamente más grande que en la tarea 1.

Tarea 3

	Knesser-Ney interpolation	Witten-Bell- nterpolation	Knesser-Ney backoff	Witten-Bell backoff
3-gram	7,70	7,36	8,32	7,83
4-gram	7,02	6,58	7,96	7,15

Tabla 3: Resultados tarea 3



Gráfica 3: Resultados tarea 3

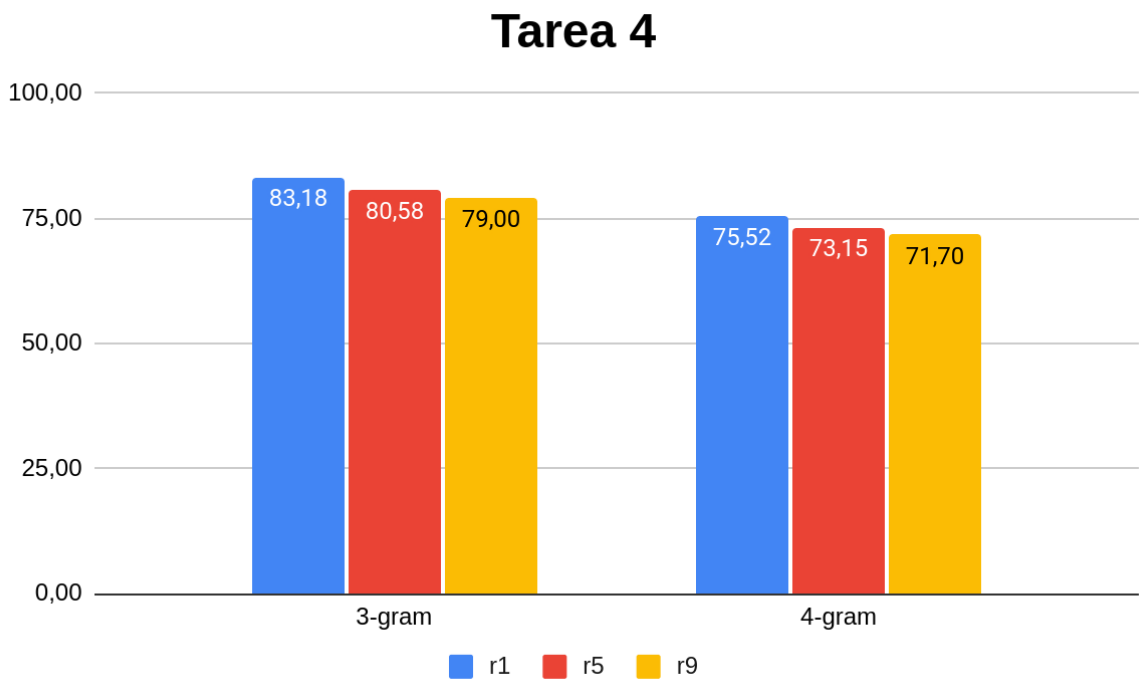
Para esta tarea se ha debía usar el método de interpolación en lugar del de back-off (el que se realiza por defecto) y comparar ambos métodos.

Cómo se puede observar, el método de interpolación proporciona mejroes resultados que el método de back-off, tanto para trigramas como para 4-grams. Así que el método a usar será el de interpolación en el problema a resolver.

Tarea 4

	r1	r5	r9
3-gram	83,18	80,58	79,00
4-gram	75,52	73,15	71,70

Tabla 4: Resultados de la tarea 4



Gráfica 4: Resultados de la tarea 4

En esta tarea se ha obtenido un modelo de lenguaje en el corpus Europarl eliminando ciertas unidades del vocabulario según la frecuencia de aparición. Se han generado modelos de lenguaje en trigramas y 4-grams cómo en las tareas 1, 2 y 3.

Se han hecho tres pruebas eliminando las palabras que sólo aparecían 1 vez, las que sólo aparecían menos de 6 veces y las que sólo aparecían menos de 10 veces. Con estas modificaciones se puede ver que la perplejidad disminuye según se eliminan palabras del vocabulario.

El hecho de que la perplejidad disminuya es algo que aporta beneficios al modelo de lenguaje, pero eliminar muchas palabras del vocabulario no es algo que sea bueno para el problema, pues aunque se reduzca la entropía (hay menos n-grams) también se pierde información.

Así que puede que eliminando las plabras que tengan una frecuencia menor a 2 o 3 sea beneficioso (pueden ser errores de escritura) porque reduce la perplejidad, pero podría ser

más pejudicial que beneficioso. Así que la eliminación de palabras según la frecuencia de aparición debería realizarse con frecuencias bajas.

Tarea 5

Con los datos obtenidos en las tareas anteriores y las conclusiones obtenidas en cada una, se puede afirmar que un buen modelo de lenguaje para el problema planteado debería estar basado en 4-grams con el método de descuento *Unmodified Knesser-Ney* y el método de interpolación, eliminando las palbras cuya frecuencia sea menor a 1.