# Laboratorio de Modelos de Lenguaje

David Barbas Rebollo

Lingüística Computacional
Octubre, 2020

## 1 Introducción

Este trabajo presenta la evaluación de distintos de modelos de n-gramas con el uso de diferente modelos de suavizado. Para ello se usan dos corpus, el corpus Dihana y el corpus Europarl. Además se utiliza la herramienta SRILM para la estimación y evaluación de los modelos.

Para la evalución de los modelos se calcula la perplejidad de los mismos. Esta métrica indica la capacidad de un modelo de predecir una nueva muestra. A continuación se expone la experimentación realizada en conjunto con los resultados obtenidos para cada tarea.

## 2 Tarea 1

La primera tarea del trabajo compara los distintos modelos en función de la N de los N-gramas. Para ello, se utiliza el corpus Dihana, el descuento Good-Turing y el suavizado Backoff.

En la siguiente tabla se puede observar como la perplejidad disminuye hasta N=4, a partir de ese punto parece que la perplejidad aumenta un poco debido al gran tamaño de la ventana de los n-gramas. En consecuencia, en los siguientes experimentos se utiliza sólo trigramas y cuatrigramas.

Table 1: Comparación de la perplejidad, variando N para el corpus Dihana, con descuento Good-Turing y suavizado Backoff.

| Corpus | N | Descuento | Suavizado | Perplejidad |
|---|---|---|---|---|
| Dihana | 1 | Good-Turing | Backoff | 97.050 |
| Dihana | 2 | Good-Turing | Backoff | 11.216 |
| Dihana | 3 | Good-Turing | Backoff | 7.678 |
| Dihana | 4 | Good-Turing | Backoff | 7.227 |
| Dihana | 5 | Good-Turing | Backoff | 7.271 |

# 3 Tarea 2

El objetivo de la segunda tarea es comparar la calidad de distintos métodos de descuentos disponibles. En específico, Good-Turing, Witten-Bell, Modified Kneser-Ney y Kneser-Ney. Para ello, se utiliza el corpus Dihana, el suavizado Backoff, trigramas y cuatrigramas.

Table 2: Comparación de la perplejidad, variando el método de descuento para el corpus Dihana, con suavizado Backoff, con trigramas y cuatrigramas.

| Corpus | N | Descuento | Suavizado | Perplejidad |
|---|---|---|---|---|
| Dihana | 3 | Good-Turing | Backoff | 7.678 |
| Dihana | 4 | Good-Turing | Backoff | 7.227 |
| Dihana | 3 | Witten-Bell | Backoff | 7.831 |
| Dihana | 4 | Witten-Bell | Backoff | 7.152 |
| Dihana | 3 | Mod. Kneser-Ney | Backoff | 8.324 |
| Dihana | 4 | Mod. Kneser-Ney | Backoff | 7.957 |
| Dihana | 3 | Kneser-Ney | Backoff | 7.653 |
| Dihana | 4 | Kneser-Ney | Backoff | 7.039 |

Observamos como utilizando un descuento de Kneser-Ney en vez de Good-Turing mejora ligeramente los resultados. Los resultados son similares excepto para el descuento Modified Kneser-Ney, ya que observamos un empeoramiento de los resultados.

# 4 Tarea 3

En la tercera tarea se comparan dos métodos de suavizado estudiados en clase, Backoff e interpolación, con los descuentos Witten-Bell y Modified Kneser-Ney. Para ello, se utiliza el corpus Dihana con trigramas y cuatrigramas.

Table 3: Comparación de la perplejidad, variando el método de descuento para el corpus Dihana, con suavizado Backoff e interpolación con trigramas y cuatrigramas.

| Corpus | N | Descuento | Suavizado | Perplejidad |
|---|---|---|---|---|
| Dihana | 3 | Witten-Bell | Backoff | 7.831 |
| Dihana | 4 | Witten-Bell | Backoff | 7.152 |
| Dihana | 3 | Witten-Bell | Interpolation | 7.365 |
| Dihana | 4 | Witten-Bell | Interpolation | 6.577 |
| Dihana | 3 | Mod. Kneser-Ney | Backoff | 8.324 |
| Dihana | 4 | Mod. Kneser-Ney | Backoff | 7.957 |
| Dihana | 3 | Mod. Kneser-Ney | Interpolation | 7.705 |
| Dihana | 4 | Mod. Kneser-Ney | Interpolation | 7.015 |

Al finalizar esta tarea, se observa como para este caso en concreto, el uso de suavizado por interpolación consigue mejorar los valores de perplejidad obtenidos previamente en evaluación.

# 5 Tarea 4

En la cuarta, y última, tarea, se ha utilizado el corpus Europarl. El objetivo de esta tarea es comparar la perplejidad en función del vocabulario usado para estimar los modelos. Se hacen 3 reducciones del vocabulario distintas. Donde se eleminan las palabras con frecuencia igual a 1, inferior o igual a 5 e inferior o igual a 9. Para ello, se usa el descuento Good-Turing y el suavizado Backoff.

Table 4: Comparación de la perplejidad variando el vocabulario utilizado con descuento Good-Turing, suavizado Backoff, trigramas y cuatrigramas.

| Corpus | N | Descuento | Suavizado | Frecuencia | Perplejidad |
|---|---|---|---|---|---|
| Europarl | 3 | Good-Turing | Backoff | $> 1$ | 83.181 |
| Europarl | 4 | Good-Turing | Backoff | $> 1$ | 75.524 |
| Europarl | 3 | Good-Turing | Backoff | $> 5$ | 80.581 |
| Europarl | 4 | Good-Turing | Backoff | $> 5$ | 73.145 |
| Europarl | 3 | Good-Turing | Backoff | $> 9$ | 78.995 |
| Europarl | 4 | Good-Turing | Backoff | $> 9$ | 71.701 |

Tras obtener los resultados de evaluación, se puede observar como la perplejidad mejora según se eliminan las muestras con menor frecuencia. Por otra parte, hay que tener en cuenta que al hacer estas reducciones de vocabulario, los modelos estimados no serán capaces de reconocer las palabras eliminadas.

# 6 Conclusiones

Al finalizar toda la experimentación comentada previamente, se puede ver que con este trabajo se pone en práctica varios de los conceptos estudiados en la asignatura.

Tomando el valor de la perplejidad como métrica de calidad de los modelos, vemos como para el corpus Dihana los mejores resultados obtenidos han sido con los parámetros: N=4, descuento Witten-Bell y suavizado Interpolation. Así que se observa que la exploración de la técnica de suavizado a ayudado a encontrar el mejor modelo.

Para el corpus Europarl, los mejores resultados se han conseguido con lo siguientes parámetros: N=4, descuento Good-Turing, suavizado Backoff y eliminando las palabras con frecuencia menor o igual a 9. Sin embargo, habría que estudiar el efecto de esta reducción de vocabulario para verificar esta conclusión.