# Modelos de lenguaje:
SRILM

Ramon Ruiz Dolz

Octubre 2018

# 1 Introducción

Este trabajo presenta la evaluación y comparativa de distintos modelos de n-gramas así como métodos de suavizado. Para ello se ha hecho uso de dos corpus distintos, el corpus Dihana y el corpus Europarl. Además, se ha utilizado la herramienta SRILM para la estimación y la evaluación de los distintos modelos de lenguaje.

Para la evaluación de los modelos estimados se realiza el calculo de la perplejidad. Esta medida nos indica, un modelo determinado, cómo de bien es capaz de predecir una nueva muestra. A continuación se expone la experimentación y los resultados obtenidos para cada tarea realizada en este trabajo.

# 2 Tarea 1

La primera tarea realizada en este trabajo ha consistido en la comparación de distintos modelos de lenguaje en función de la N de N-gramas. Para ello, haciendo uso del corpus Dihana, el descuento Good-Turing y suavizado por backoff, se ha ido variando el valor de N. Se han obtenido los siguientes valores de perplejidad:

| **Corpus** | **N** | **Descuento** | **Suavizado** | **Perplejidad** |
|---|---|---|---|---|
| Dihana | 1 | Good-Turing | Backoff | 160.768 |
| Dihana | 2 | Good-Turing | Backoff | 18.539 |
| Dihana | 3 | Good-Turing | Backoff | 15.041 |
| Dihana | 4 | Good-Turing | Backoff | 14.896 |
| Dihana | 5 | Good-Turing | Backoff | 15.075 |

Como se puede observar, la perplejidad disminuye hasta N=4, a partir de este punto empieza a aumentar debido al gran tamaño de la ventana de los N-gramas. Es por esto que, para los siguientes experimentos se utilizaran los valore 3 y 4 asociado a la N de los N-gramas.

# 3 Tarea 2

Mediante la segunda tarea ha sido posible comparar la calidad de los distintos métodos de descuento disponibles. Para ello, se ha hecho uso del corpus Dihana, y asignando a N los valores 3 y 4 se han probado los métodos de descuento Good-Turing, Witten-Bell, modified Kneser-Ney y unmodified Kneser-Ney. Por lo tanto, en esta tarea se han estimado 8 modelos de lenguaje distintos. Los valores de perplejidad para cada uno de ellos se pueden observar a continuación:

| Corpus | N | Descuento | Suavizado | Perplejidad |
|---|---|---|---|---|
| Dihana | 3 | Good-Turing | Backoff | 15.041 |
| Dihana | 4 | Good-Turing | Backoff | 14.896 |
| Dihana | 3 | Witten-Bell | Backoff | 15.035 |
| Dihana | 4 | Witten-Bell | Backoff | 14.724 |
| Dihana | 3 | Modif. Kneser-Ney | Backoff | 15.061 |
| Dihana | 4 | Modif. Kneser-Ney | Backoff | 15.312 |
| Dihana | 3 | Kneser-Ney | Backoff | 14.508 |
| Dihana | 4 | Kneser-Ney | Backoff | 14.387 |

# 4 Tarea 3

Mediante la tercera tarea se ha realizado la comparativa entre dos de los métodos de suavizado estudiados. Concretamente se han comparado los métodos de backoff e interpolación.

Para ello se han aplicado los métodos de descuento de Witten-Bell y modified Kneser-Ney sobre el corpus Dihana. Además se han asignado los valores de 3 y 4 a N. La comparación de los valores de perplejidad entre el uso de backoff y de interpolación son los siguientes:

| Corpus | N | Descuento | Suavizado | Perplejidad |
|---|---|---|---|---|
| Dihana | 3 | Witten-Bell | Backoff | 15.041 |
| Dihana | 4 | Witten-Bell | Backoff | 14.896 |
| Dihana | 3 | Witten-Bell | Interpolación | 14.754 |
| Dihana | 4 | Witten-Bell | Interpolación | 14.708 |
| Dihana | 3 | Modif. Kneser-Ney | Backoff | 15.061 |
| Dihana | 4 | Modif. Kneser-Ney | Backoff | 15.312 |
| Dihana | 3 | Modif. Kneser-Ney | Interpolación | 14.253 |
| Dihana | 4 | Modif. Kneser-Ney | Interpolación | 14.024 |

Al finalizar esta tarea, se ha observado como para este caso concreto, la estimación de los modelos mediante el suavizado por interpolación nos ha permitido obtener unos mejores valore de perplejidad a la hora de realizar la evaluación.

# 5 Tarea 4

Finalmente, en la última tarea se ha utilizado un corpus diferente. Para la realización de esta tarea se ha hecho uso del corpus Europarl que contiene actas del Parlamento Europeo en 11 lenguas, entre ellas castellano e inglés.

En este experimento se ha tratado de observar las diferencias en los modelos estimados en función del vocabulario utilizado para estimar. Para ello se han estimado tres modelos diferentes, uno eliminando del vocabulario las palabras de frecuencia uno. Otro eliminando del vocabulario todas las palabras con una frecuencia menor o igual a cinco y, finalmente otro eliminando del vocabulario

todas las palabras con frecuencia menor o igual a nueve. Los valores de perplejidad asociados a cada modelo son los siguientes:

| **Corpus** | **N** | **Descuento** | **Suavizado** | **Frecuéncia** | **Perplejidad** |
|---|---|---|---|---|---|
| Europarl | 3 | Good-Turing | Backoff | >=2 | 99.423 |
| Europarl | 4 | Good-Turing | Backoff | >=2 | 89.919 |
| Europarl | 3 | Good-Turing | Backoff | >5 | 96.241 |
| Europarl | 4 | Good-Turing | Backoff | >5 | 87.018 |
| Europarl | 3 | Good-Turing | Backoff | >9 | 94.306 |
| Europarl | 4 | Good-Turing | Backoff | >9 | 85.262 |

Al finalizar los experimentos se puede observar como la perplejidad del modelo mejora eliminando las muestras con menor frecuencia. Esto, sin embargo no es sinónimo de mejora, puesto que hay que tener en cuenta que al eliminar esas palabras del vocabulario, el modelo estimado no será capaz de reconocerlas. Por lo tanto hay que evaluar hasta que punto puede ser interesante la eliminación de las palabras poco frecuentes del vocabulario.

# 6 Conclusiones

Con la realización de este trabajo ha sido posible poner en práctica gran parte de los conceptos estudiados en la asignatura. Tomando el valor de la perplejidad como referencia a la hora de determinar la calidad de un modelo, se puede afirmar que, para el corpus Dihana se han obtenido los mejores modelos de lenguaje con los parámetros N=4, descuento de Kneser-Ney modificado y suavizado por interpolación. Sin embargo también se han obtenido buenos modelos con N=4, descuento Kneser-Ney con backoff. De hecho, ha sido la variación del suavizado, de backoff a interpolación, la que ha permitido encontrar el mejor modelo del lenguaje.

Por otra parte, con el corpus Europarl, los mejores resultados se han obtenido eliminando las palabras con frecuencia menor a 10 veces del vocabulario utilizado para la estimación del modelo. Sin embargo, como ya se ha comentado, esto no convierte necesariamente a este modelo mejor que el estimado eliminando las palabras con frecuencia menor a 6 veces.