UNIVERSITAT POLITÈCNICA DE VALÈNCIA
Departamento de Sistemas Informáticos y Computación

Lingüística Computacional
Modelos de lenguaje: SRILM

Luis Cardoza Bird.

Octubre de 2023

I. Introducción

La herramienta **SRILM** es utilizada para modelar y evaluar el lenguaje, por ende este trabajo busca demostrar la implementación de dicha herramienta utilizando como demostración dos corpus: “**Corpus Dihana**” y “**Corpus Europarl**”.

Se aplicarán técnicas de suavizado y comparativas para poder evaluar dichos modelos.

II. OBJETIVOS

A. Conocer las herramientas de modelización del lenguaje proporcionadas por el toolkit SRILM.
B. Construir modelos de lenguaje para un corpus pequeño (dihana).
C. Comparar las prestaciones de distintos modelos.
D. Construir modelos de lenguaje para un corpus grande (Europarl).

III. TAREAS

A. Tarea 1

La tarea central en este trabajo implica comparar modelos de lenguaje según la N en los N-gramas. Usando el **corpus Dihana**, con el descuento **Good-Turing** y suavizado por **backoff**, se varió el valor de N.

La tabla muestra los valores de perplejidad obtenidos:

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

La perplejidad disminuye hasta N=4. Después incrementa por el extenso rango de la ventana de N-gramas.

B. Tarea 2

En esta tarea apreciamos una diferencia en donde se contrasta la eficacia de diferentes métodos de descuento.

Usando el **corpus Dihana** y valores de N en 3 y 4, probamos los descuentos **Good-Turing**, **Witten-Bell**, **modified Kneser-Ney** y **unmodified Kneser-Ney**. Esto resulta en la estimación de 8 modelos de lenguaje distintos. La tabla que sigue muestra los valores de perplejidad para cada uno.

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	Modified Kneser-Ney	Backoff	15.061
Dihana	4	Modified Kneser-Ney	Backoff	15.312
Dihana	3	Unmodified Kneser-Ney	Backoff	14.508
Dihana	4	Unmodified Kneser-Ney	Backoff	14.387

C. Tarea 3

En la siguiente tarea realizamos una comparación respecto a los métodos de suavizado **backoff** e **interpolación**.

Aplicando los descuentos de **Witten-Bell** y **modified Kneser-Ney** al **corpus Dihana** y con valores N de 3 y 4, se generaron diversos modelos. Los valores de perplejidad se listan a continuación:

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	Modified Kneser-Ney	Backoff	15.061
Dihana	4	Modified Kneser-Ney	Backoff	15.312
Dihana	3	Modified Kneser-Ney	Interpolación	14.253
Dihana	4	Modified Kneser-Ney	Interpolación	14.024

Se han logrado obtener mejores valores de perplejidad.

D. Tarea 4

En la siguiente tarea se utilizará el corpus Europarl, que contiene actas del Parlamento Europeo en 11 idiomas, incluyendo el castellano e inglés.

El enfoque fue observar diferencias en modelos basados en distintos umbrales de vocabulario.

Se desarrollaron tres modelos considerando palabras con distintas frecuencias mínimas:

- Eliminando del vocabulario las palabras de frecuencia 1
- Eliminando del vocabulario las palabras de frecuencia <=5
- Eliminando del vocabulario las palabras de frecuencia <=9

Corpus	N	Descuento	Suavizado	Frecuencia	Perplejidad
Europarl	3	Good-Turing	Backoff	1	99.423
Europarl	4	Good-Turing	Backoff	1	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

Los resultados muestran que la perplejidad mejora al excluir palabras menos frecuentes, pero al hacerlo se puede observar que el modelo no reconoce esas palabras, lo que plantea la cuestión de cuán beneficioso es reducir el vocabulario.

E. Conclusiones

Durante este trabajo, aplicamos conceptos clave de la asignatura y usamos la perplejidad como indicador de la calidad del modelo.

Para el corpus Dihana, los mejores modelos se lograron con N=4, descuento Kneser-Ney modificado y suavizado por interpolación, aunque otros modelos con N=4 y descuento Kneser-Ney con backoff también rindieron bien. El cambio de suavizado de backoff a interpolación fue determinante.

En cuanto al corpus Europarl, los modelos más eficientes surgieron al excluir palabras que aparecen menos de 10 veces, pero no necesariamente son superiores a los que eliminan palabras con menos de 6 apariciones.