PRÁCTICA DE MODELOS DE LENGUAJE

1.- Objetivos de la práctica:

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

2.- Recursos y herramientas

- SRILM [1] es una herramienta que soporta la estimación y evaluación de modelos de lenguaje basados en N-gramas. Contiene los siguientes componentes:
    - Un conjunto de librerías C++ que implementan modelos de lenguaje, que soportan estructuras de datos y funciones de utilidad variada.
    - Un conjunto de programas ejecutables construidos sobre estas librerías que realizan las tareas fundamentales tales como el entrenamiento de modelos y la prueba de los mismos, etc.
    - Una colección de scripts variados que facilitan las operaciones.
- El corpus en castellano del proyecto DIHANA [2] que encontraremos en el directorio DIHANA en el apartado de recursos de PoliformaT. Sobre este corpus aplicaremos la herramienta para estimar diferentes modelos de lenguaje. El corpus consiste en 900 diálogos adquiridos con 225 usuarios a los que se les proporcionó varios escenarios y objetivos que debían lograr. Los usuarios interrogaban en habla espontánea a un sistema de diálogo que proporcionaba información sobre horarios, precios y servicios de trenes de largo recorrido.
- El corpus European Parliament (Europarl) [3] construido por Philipp Koehn, MT Summit 2005. Contiene actas del Parlamento Europeo en 11 lenguas, entre ellas castellano e inglés. La versión que manejamos es v3 que data de 28 de septiembre de 2007. El corpus en castellano contiene 1.476.106 oraciones y 41.408.300 palabras. Se suelen utilizar los datos de 200-10 a 2000-12 como conjunto de prueba. Los datos que vamos a utilizar de este corpus, ya preprocesado los encontraremos en el apartado de recursos de PoliformaT.

3.- Descripción de la práctica:

La práctica consiste en la evaluación de las prestaciones de los diferentes modelos de lenguaje de n-gramas y los diferentes métodos de suavizado para su uso como modelos de lenguaje en el corpus dihana. En esta experimentación se estudiará cómo afectan diversos parámetros a las prestaciones del modelo: el método de suavizado, los métodos de descuento, etc. Adicionalmente se trabajará también con un corpus de mayor tamaño, el Europarl.

Para ello se utilizará el paquete SRILM. Es un paquete de libre distribución para usos académicos bajo una licencia "open source community license" y puede conseguirse en la dirección http://www.speech.sri.com/projects/srilm/.

En la carpeta *Corpus Dihana* de recursos de PoliformaT hay un fichero de entrenamiento y uno de prueba para la experimentación con el corpus Dihana. En la carpeta Corpus Europarl se encuentran los ficheros de entrenamiento y prueba para la experimentación con el corpus Europarl, junto con un par de ficheros que contienen el vocabulario.

Formato de los ficheros de oraciones del corpus Dihana: cada oración en una línea del fichero ASCII. Cada oración es una secuencia de "palabras" separadas por blancos. Ej.

```
El pr'oximo s'abado d'ia trece antes de las doce
```

Formato de los ficheros de oraciones Europarl: cada oración en una línea del fichero con codificación UTF-8. Cada oración es una secuencia de "palabras" (hay algunos errores en las transcripciones y traducciones) separadas por blancos. Ej.

```
el parlamento rechaza la petición al presidente
```

TAREAS

El alumno deberá realizar las tareas que se detallan a continuación, y en este orden:

Tarea 0: Uso del paquete SRILM:

En la dirección web http://www.speech.sri.com/projects/srilm/ obtendréis información sobre el uso del paquete.

Tarea 1: Comparación de los diferentes modelos de lenguaje según el valor de N.

Utilizando el **corpus Dihana** y el suavizado por defecto de la herramienta (descuento **Good-Turing** y esquema de suavizado **backoff**) se pide obtener las perplejidades de los modelos para **N= 1,2,3,4 y 5**.

A continuación se detallan los pasos para realizar una ejecución para un valor determinado del parámetro *N*:

1.- Construcción del modelo a partir del conjunto de entrenamiento.

ngram-count –order *N* –lm modelo*N* –text entrenamiento

2.- Evaluación de cada modelo con su conjunto de prueba correspondiente, es decir, cálculo de la perplejidad.

ngram –order *N* –lm modelo*N* –ppl prueba

Tarea 2: Comparación de los diferentes modelos de lenguaje según el método de descuento utilizado.

Utilizando el **corpus Dihana** y el valor de **N= 3,4**, se pide estimar diferentes modelos de lenguaje variando los métodos de descuento. Los pasos a seguir son los mismos que en la tarea anterior pero en este caso, incluyendo en la estimación del modelo las opciones adecuadas para ir variando el método de descuento: Good-Turing, Witten-Bell, modified Kneser-Ney y unmodified Kneser-Ney. Comparar para cada método de descuento los diferentes resultados de perplejidad.

Tarea 3: Comparación del suavizado por backoff e interpolación.

Utilizando el **corpus Dihana** y el valor de **N =3,4**, se pide estimar diferentes modelos de lenguaje para los descuentos Witten-Bell y modified Kneser-Ney, pero esta vez bajo un esquema de **interpolación** en lugar de backoff. (hay una opción de "ngram-count" que permite hacerlo). Comparar para cada método de descuento los resultados de perplejidad cuando el modelo es backoff o interpolado.

Tarea 4: Modelos de lenguaje para el corpus Europarl con diferentes tallas de vocabulario (*Opcional*).

Fijando un valor de **N= 3,4** y el método de descuento y suavizado por defecto, se pide estimar un modelo de lenguaje para el corpus Europarl, tomando como **conjunto de entrenamiento** y **conjunto de prueba** los que se proporcionan en el PoliformaT.
En lugar de tomar como vocabulario para los modelos de lenguaje el que se estima a partir del conjunto de entrenamiento, vamos a entrenar proporcionando a la herramienta un fichero con el vocabulario. En el PoliformaT encontraréis dos ficheros de vocabulario, en los dos aparecen todas las palabras del corpus de entrenamiento ordenadas por frecuencia de aparición en el corpus. En uno tenemos las palabras con su frecuencia de aparición y en otro, siguiendo la misma ordenación, se proporcionan sólo las palabras.
Hay que realizar varios experimentos:
- Eliminando del vocabulario las palabras de frecuencia 1.
- Eliminando del vocabulario las palabras de frecuencia <=5
- Eliminando del vocabulario las palabras de frecuencia <=9

Para cada fichero de vocabulario hay que entrenar un modelo de lenguaje para **N= 3,4**.
Se pide evaluar la perplejidad de los diferentes modelos entrenados con el conjunto de prueba que se proporciona.

Tarea 5: Conclusiones.
A partir de los resultados obtenidos en las anteriores experimentaciones, se pide una pequeña discusión acerca de qué métodos han proporcionado los mejores modelos de lenguaje para el **corpus Dihana** (y el **corpus Europarl** si se resuelve la parte opcional).

REFERENCIAS

[1] A. Stolke. "SRILM - An Extensible Language Modeling Toolkit", in Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.
[2] J.M.Benedí, E.Lleida, A.Varona, M.J.Castro, I.Galiano, R.Justo, I.López and A.Miguel. "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA". In Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 1636–1639. European Language Resources Association, 2006.
[3] P. Koehn. "Europarl: A Parallel Corpus for Statistical Machine Translation", MT Summit 2005.