Práctico 3: TA estadística

- En este práctico vamos a construir un sistema de estadístico con herramientas que implementan los modelos IBM 3 y 4 vistos en clase
- Por lo tanto vamos a
  1. Definir en corpus paralelo de entrenamiento
  2. Estimar P(f|e) o modelo de traducción
  3. Estimar P(e) o modelo de lenguaje
  4. Buscar la mejor traducción dados los alineamientos generados anteriormente

Práctico 3: qué recursos usar?

1. Textos paralelos: porción de Europarl para entrenamiento
   - Elegir un par de idiomas entre {de, en, es, fr, nl} del conjunto dev2006.* o un par entre {cz, de, en, es, fr} del conjunto nc-dev2007.*
     - http://www.famaf.unc.edu.ar/~pestrella/practico-3/corpus.zip
   - Elegir el archivo a traducir (por ej si hago Cz-Es elijo test.cz)
     - http://www.famaf.unc.edu.ar/~pestrella/practico-3/test-set.zip
2. Alinear textos con GIZA
   - Hay que compilarlo con gcc 2.8.2 o 2.95.1; versiones mayores pueden no andar bien
   - http://www.famaf.unc.edu.ar/~pestrella/practico-3/GIZA.zip
     - Si tuvieran muchos problemas con GIZA traten usando una nueva versión GIZA++ pero ver compatibilidad con ISI-decoder
     - http://www.famaf.unc.edu.ar/~pestrella/practico-3/GIZA++.tar.gz
3. Entrenar modelo de lenguaje con CMU-toolkit
   - http://www.famaf.unc.edu.ar/~pestrella/practico-3/CMU-Cam_Toolkit_v2.tar.gz
4. Decodificar con ISI-decoder
   - http://www.famaf.unc.edu.ar/~pestrella/practico-3/isi-rewrite-decoder-r0.7.0b.i686-linux.tgz

Práctico 3: qué hay que hacer?

- Usando los mismos archivos de entrenamiento y test:

1. Crear un sistema donde los alineamientos y el decoder usen el modelo IBM 3 (sys1)
   - Traducir y guardar la salida para uso posterior
2. Crear otro sistema que use el modelo IBM 4 (sys2)
3. Crear un sistema que use IBM 4 y pre-procese el texto
   - Antes de entrenar, pasar todo el texto a minúsculas y tokenizar
     - Scripts útiles http://www.statmt.org/wmt08/scripts.tgz
   - Luego de traducir, recomponer los tokens y las mayúsculas
4. Comparar la salida de los tres sistemas
   - Hay alguna diferencia en la salida de sys1, sys2 y sys3? A qué creen que se deben las diferencias?
   - Es alguno mejor que otro? Qué parámetros usan para determinar si una traducción es mejor que otra?

Práctico 3: TA estadística

- Para hacer este práctico van a tener que leer la documentación de cada herramienta, instalar librerías necesarias, etc
- Por favor documentar todo tipo de problemas, soluciones, conclusiones, ideas que surjan como posibles trabajos finales de la materia
- La próxima clase discutimos cómo les fue