

Machine translation

Sentence based statistical translation: MOSES

Adrián Vázquez Barrera
Polytechnic University of Valencia
January 2022

Introducción

En esta práctica se pretenden construir sistemas de traducción utilizando pares de frases bilingües. Para este fin, utilizaremos la herramienta Moses para entrenar modelos estadísticos. En la próxima sesión, se construirá un sistema similar basado en redes neuronales con el toolkit NMT-Keras.

Contexto

Para la realización de esta práctica, utilizaremos un conjunto de datos bilingüe de la tarea EuTrans para Español-Inglés, en el contexto de un turista frente al mostrador de un hotel. El corpus ya se encuentra tokenizado por lo que no será necesario someterlo a este proceso, no obstante, sí es conveniente limpiarlo.

También será necesario crear un modelo de n-gramas para el idioma de destino, en este caso se ha apostado por un modelo de trigramas y Kneser-Kney como método de descuento.

Con a penas 5 iteraciones y empleando los valores por defecto de los parámetros vistos en la práctica, obtenemos un BLEU del 92.02%.

Ejercicio 1

El ejercicio 1 radica en entrenar el modelo sin el posterior ajuste de pesos log-lineal. Como cabría esperar, el BLEU obtenido cae hasta los 88.42. Esto se debe a que MERT (o MIRA como veremos en ejercicios posteriores) ayuda al ajuste de los pesos dado un modelo de desarrollo, haciéndolo dependiente de este y añade coste computacional extra.

Ejercicio 2

Con el doble de iteraciones la mejoría es muy leve y con cincuenta o cien apreciamos el efecto del sobre-ajuste, perdiendo incluso calidad en la traducción.

MERT	
Iteraciones	BLEU
5	92.02%
10	92.52%
50	91.83%
100	91.93%

Ejercicio 3

Analizando los resultados obtenidos reflejados en la tabla, vemos como es imprescindible contar con un buen modelo de lenguaje para entrenar modelos de traducción estadísticos.

N-grama	BLEU
1	0%
2	88.79%
3	92.02%
4	89.05%

Con unigramas es imposible conseguir un modelo útil, porque es como no tener modelo de lenguaje. El uso de bigramas empeora el BLEU obtenido, esto se debe a la falta de historia pasada y contexto previo de las oraciones. Con cuatrigramas también se experimenta una caída en el BLEU, esto es debido a que, a mayor número de n-gramas obtenemos un menor número de combinaciones no vistas en el entrenamiento.

Ejercicio 4

Según el manual de MOSES, cuando se proporciona el flag `-batch-mira`, se reemplaza la llamada a MERT por una a `kbmira`. De esta forma, se activa el parámetro de `tunning k-best batch MIRA`, que lleva a cabo `online-training` utilizando listas `k-mejores` agregadas como una aproximación del espacio de búsqueda.

El BLEU obtenido al realizar esta modificación es de 90.95, que es inferior al obtenido con MERT. No obstante, debemos tener en cuenta que MIRA converge más rápido con una solución bastante buena considerando su menor coste computacional.

Ejercicio 5

El método de Kneser-Ney utilizado originalmente, ofrece mejores resultados que Good Turing y Unmodified Kneser-Ney. Esto es debido a que el primero emplea un contador de los elementos vistos una vez para imputar una probabilidad a los elementos no vistos. Por otro lado, Kneser-Ney se basa en el contexto para imputar dicha probabilidad. En este sentido, se comprende que la versión más refinada del método sea la que mejor funcione, aunque la diferencia sea poco significativa.

Método	BLEU
Good Turing	89.85%
Unmodified Kneser-Ney	91.80%

Ejercicio 6

Para aplicar moses monótono, es necesario establecer el parámetro -distortion-limit a cero, o lo que es lo mismo, no reordenar. Con esto se consigue una traducción literal, con un BLEU de 91.05.

Es evidente que este tipo de traducción puede ser más rápida, pero penaliza el BLEU obtenido. Sobre todo si traducimos del Español a Inglés donde el orden de los elementos gramaticales cambia a la hora de construir las frases.

Bibliografia

1. Philipp Koehn. MOSES: Statistical Machine Translation System. User Manual and Code Guide. University of Edinburgh. 2014.
<http://www.statmt.org/moses/manual/manual.pdf>
2. SRILM - The SRI Language Modeling Toolkit
<http://www.speech.sri.com/projects/srilm/>