Etiquetadores

- Esta tarea es común dentro del Procesameinto del Lenguaje Natural (PLN)
- Por ejemplo, asignar etiquetas POS (*Part Of Speech*) a las palabras puede ayudar a desambiguar su significado

Glosa

Lenguaje Natural

- Presenta fenómenos que hacen de la construcción de etiquetadores automáticos una tarea difícil

Bajos recursos digitales

- A pesar de la gran diversidad lingüística, gran parte de las lenguas originarias no poseen contenido web ni publicaciones digitales
  - 68 lenguas
  - 364 variantes dialectales
  - 11 familias linguisticas
- Los bajos recursos digitales propician que estas lenguas carezcan de tecnologías del lenguaje
  - Herramientas de preprocesamiento como analizadores morfologicos, datasets anotados

Retos de investigacion

- Los métodos tradicionales de etiquetado automático requieren grandes cantidades de datos para funcionar correctamente
- Dificultad para encontrar contenido digital
- Gran variacion dialectal y ortografica en textos
- Falta de normalización puede complicar la tarea
- El desarrollo e invetigación en esta area puede tener un impacto social positivo

Corpus

- El otomi como muchas lenguas mexicanas presenta una vasta variedad lo que implica diferentes otografias. En particular el otomi es tonal y distingue entre vocales orales y nasales. El inventario de vocales orales no se limita a las 5 del español. La representacion digital es un reto por la codificacion. Fue necesario modificar algunas vocales del otomi como se muestra en la tabla

Arquitectura

- Proponemos una arquitectura de aprendizaje estructurado utilizando CRFs para la generación de modelos capaces de glosar automáticamente textos del otomí de Toluca

Obtención del corpus

Codificación

- El corpus se obtuvo de un archivo de texto plano
- Cada renglon era una oración en otomí con glosa y una etiqueta POS por cada palabra
- Las frases estan estructuradas en listas de python validas
- En esta parte se sustituyeron las vocales del otomi por las que mencionamos anteriormente

Preprocesamiento

- Adecuación de la estructura del corpus para la creación de las *feature functions*
- Se construyeron las feature functions

CRFs

Los Conditional Random Fields (CRFs) son un framework para la crea- ción de modelos probabilístico utilizado en técnicas de aprendizaje estructu- rado. Tienen las ventajas de los Maximum Entropy Markov Models y, en principio, solucionan el label bias problem.

Bias problem

- Las transiciones que salen de un estado competen solo entre ellos, en lugar de entre todas las otras transiciones del modelo

Feature functions

- La extracción de estas características es importante porque capturar fenómenos lingüísticos necesarios para que la estructura de la lengua se pueda plasmar en el modelo de aprendizaje. Estas características están capturando, entres otras cosas, el contexto de la palabra y es importante para predecir la morfología.
- Se obtuvieron feature functions para cada letra dentro de las palabras que componen las frases del corpus

Evaluacion

- Para cada entorno se crearon diferentes modelos variando los hiperparametros del algoritmo de optimizacion L-BFGS para los CRFs
-

Ejemplos

- En el ejemplo 1 Vemos la aparición de etiquetas sistemáticas como stem y 3.icp, por mencionar algunas, por lo que la frase es correctamente etiquetada.
- En el ejemplo 2a se ve como el modelo hace una mala segmentación. Esto puede deberse por un lado a que la etiqueta 3.icp es muy frecuente en el corpus y que dado que la etiqueta previa es psd el modelo se inclina por estas etiquetas frecuentes.
- El patrón psd-3.icp es muy frecuente por lo que el modelo tiende a utilizarlo causando un mal etiquetado. Por otra parte, la combinación correcta de etiquetas (2.pot-stem) es poco frecuente. La versión del etiquetado esperado se puede ver en el ejemplo 2b