Índice general

Etiquetador automático de la morfología del otomí usando predicción estructurada

Diego Alberto Barriga Martínez

22 de enero de 2020

Capítulo 1

Etiquetador morfológico para el otomí

En este capítulo se mostrará el pipeline y arquitectura utilizada para la realización de esta tesis. Adicionalmente, se explicará el diseño, implementación y las feature functions elegidas para el correcto glosado de frases en otomí.

1.1. CRF

1. Que son 2. Como funcionan

1.2. Corpus

1. De donde viene 2. Tipo de otomi, quien lo recolecta y quien lo glosa 3. Tamaño de tipos y tokens. 4. Numero de etiquetas diferentes 5. Tipos de POS y cuantos de cada uno

Descripción general del corpus. Orientar que la arquitectura esta enfocada en resolver el problema de low resources

1.3. Arquitectura

Para esta tesis proponemos una arquitectura de aprendizaje estructurado supervisado utilizando un método gráfico, Conditional Random Field (CRF),

que permitirá la predicción de secuencias que describen las unidades morfológicas (glosa) dentro de una palabra en otomí

Se utilizaron CRFs para predecir secuencias de glosa, que será la salida Y dadas las observaciones X que son el texto previamente glosado. Puntualmente, se utiliza el modelo gráfico 1st-order Markov CRF with dyad features. Adicionalmente, es utilizado el algoritmo de aprendizaje de Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) como se menciono en TODO.

Con base en el trabajo previo para del idioma Lezgi (Moeller, S. & Hulden, M, 2018) se plantea como hipótesis que dado el tamaño del corpus y la glosa que contiene se obtendrá texto correctamente glosado con una precisión de al menos 80 %. El objetivo de esta arquitectura es obtener almenos un bla % de preciosion Ya que el resultado esperado es la generación de etiquetas que, en principio, dependen unas de otras un método basado en grafos como los CRF puede ser adecuado. Se definierón de aprender un conjunto de feature functions que describen TODO el contexto y brindan información útil para la fase de entrenamiento.

En este trabajo se utilizará un corpus del otomí que, además, cumple la característica de estar glosado. Se tomará el corpus Tsunkua (Elotl, 2019) etiquetado por el lingüista Víctor Germán Mijangos de la Cruz y que está basado en el trabajo El otomí de Toluca (Lastra, 1992).

El modelo de aprendizaje semi-supervisado, para la generación de glosa para el otomí se describe a continuación:

HACER ESTA PARTE POR PUNTOS MAS BREVES Y CADA PUNTO HACERLO SECCION Y PROFUNDIZAR

- Obtener el corpus en otomí previamente glosado y obtenerlo en un formato que especifique la información de las oraciones a nivel de letra especificando su Bio Label.

- Los CRF toman como entrada los datos X que corresponden al corpus en otomí introducido en las feature functions asociados de forma biyectiva con la etiqueta Bio Label que le corresponde. Con base en esto se entrenará un modelo que busque maximizar el logaritmo de verosimilitud con el método de aprendizaje L-BFGS

- Posterior se obtendrá un modelo entrenado con el que se generarán etiquetas de glosa para el otomí. Por lo tanto, el modelo recibirá párrafos de texto en otomí y retornará el texto glosado.

- Se considera exitosa la predicción si se logra maximizar la correcta clasificación de las secuencias de salida. Para determinar si la predicción fue exitosa se utilizaron técnicas típicas de ML como K-folds que consiste en tomar K fragmentos de los datos de entrada para utilizarlos para probar el modelo y asi obtener una precisión, recall y F-score.

MENCIONAR python, version, paquetes y donde corrio. En promedio cuanto tarda en correr en la maquina. Mencionar el original en c++

Capítulo 2

Experimentación y Resultados

Cualitativos y cuantitativos
Hablar del Base line y como mejoró con mas features

2.1. Corpus de evaluación

Aqui se habla del K fold y de como se introdujo el corpus retador ?

2.2. Análisis de resultados

Capítulo 3

Conclusiones