

Glosador automático usando aprendizaje estructurado para el otomí de Toluca

Diego A. Barriga Martínez¹

Victor Mijangos Cruz¹

Ximena Gutierrez-Vasques²

5 de Noviembre 2020

Universidad Nacional Autónoma de México¹

Universidad de Zúrich²

Etiquetadores automáticos

- Los etiquetadores automáticos son una tarea común del Procesamiento de Lenguaje Natural (PLN)

- Los etiquetadores automáticos son una tarea común del Procesamiento de Lenguaje Natural (PLN)
- Es usual que se utilicen métodos de *Machine Learning (ML)* para su construcción

- Los etiquetadores automáticos son una tarea común del Procesamiento de Lenguaje Natural (PLN)
- Es usual que se utilicen métodos de *Machine Learning (ML)* para su construcción
- En particular métodos basados en gráficas, por ejemplo Modelos Ocultos de Markov (*Hidden Markov Models, HMM*)

- El glosado es un tipo de etiquetado que asigna etiquetas a las unidades que conforman una palabra

- El glosado es un tipo de etiquetado que asigna etiquetas a las unidades que conforman una palabra
- Este etiquetado de textos es de suma importancia para el análisis y la documentación lingüística

- El glosado es un tipo de etiquetado que asigna etiquetas a las unidades que conforman una palabra
- Este etiquetado de textos es de suma importancia para el análisis y la documentación lingüística
- Tradicionalmente el glosado se hace manualmente
 - Esto es lento y costoso ya que requiere de habilidades de un lingüista y un trabajo íntimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller, 2018)

- El glosado es un tipo de etiquetado que asigna etiquetas a las unidades que conforman una palabra
- Este etiquetado de textos es de suma importancia para el análisis y la documentación lingüística
- Tradicionalmente el glosado se hace manualmente
 - Esto es lento y costoso ya que requiere de habilidades de un lingüista y un trabajo íntimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller, 2018)
- La construcción de modelos automáticos que asistan este tipo de etiquetado surgen como **una tarea urgente**

El lenguaje natural

- No obstante, el lenguaje natural es **complejo**

- No obstante, el lenguaje natural es **complejo**
- Adicionalmente, existen escenarios donde los métodos tradicionales no son efectivos como es el caso de los **bajos recursos digitales**

El reto: Bajos recursos digitales

El reto: Bajos recursos digitales



Figure 1: México

El reto: Bajos recursos digitales



Figure 1: México

- Los bajos recursos digitales son un escenario común en las lenguas mexicanas

El reto: Bajos recursos digitales



Figure 1: México

- Los bajos recursos digitales son un escenario común en las lenguas mexicanas
- Este entorno de experimentación supone importante **reto de invetigación**

Objetivo

- En este trabajo esta en el marco de los bajos recursos digitales.

Objetivo

- En este trabajo esta en el marco de los bajos recursos digitales.
- Nos enfocamos en la construcción de un **glosador para el otomí de Toluca**

- En este trabajo esta en el marco de los bajos recursos digitales.
- Nos enfocamos en la construcción de un **glosador para el otomí de Toluca**
- Diseño e implementación de un **etiquetador morfológico** basado en métodos de aprendizaje estructurado débilmente supervisado.

- En este trabajo esta en el marco de los bajos recursos digitales.
- Nos enfocamos en la construcción de un **glosador para el otomí de Toluca**
- Diseño e implementación de un **etiquetador morfológico** basado en métodos de aprendizaje estructurado débilmente supervisado.
- Específicamente, *Conditional Random Fields (CRFs)* (Lafferty et al., 2001)

Corpus

- Para este trabajo se utilizó un corpus en otomí basado en el trabajo de Lastra (1992).

- Para este trabajo se utilizó un corpus en otomí basado en el trabajo de Lastra (1992).
- El corpus fue etiquetado y glosado manualmente por el lingüista Víctor Germán Mijangos Cruz.

- Para este trabajo se utilizó un corpus en otomí basado en el trabajo de Lastra (1992).
- El corpus fue etiquetado y glosado manualmente por el lingüista Víctor Germán Mijangos Cruz.
- El corpus es un subconjunto del corpus paralelo español-otomí que se encuentra en la plataforma web Tsunkua (<https://tsunkua.elotl.mx/>)

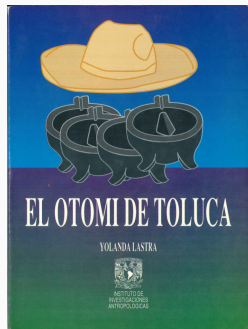


Figure 2: Portada de “El otomí de Toluca” de Lastra

- La variante con la que trabajamos es de la región de San Andrés Cuexcontitlan

- La variante con la que trabajamos es de la región de San Andrés Cuexcontitlan
- Se incluyó información morfosintáctica (*Part Of Speech, POS*) y glosa

- La variante con la que trabajamos es de la región de San Andrés Cuexcontitlan
- Se incluyó información morfosintáctica (*Part Of Speech, POS*) y glosa
- Se agregaron 81 casos con fenómenos poco frecuentes y, por tanto, particularmente difíciles de predecir.

Datos cuantitativos del corpus

Categoría	Cuenta
Tokens (POS)	8578
Tipos (POS)	44
Tokens (Glosa)	14477
Tipos (Glosa)	112
Total de oraciones etiquetadas	1786

Table 1: Tamaño del corpus

Textos	Número
Narrativos	32
Dialogados	4
Total de textos	36

Table 2: Textos del corpus

Arquitectura

1. Obtención del corpus en otomí

1. Obtención del corpus en otomí
2. Codificación

1. Obtención del corpus en otomí
2. Codificación
3. Preprocesamiento

1. Obtención del corpus en otomí
2. Codificación
3. Preprocesamiento
4. Fase de entrenamiento

1. Obtención del corpus en otomí
2. Codificación
3. Preprocesamiento
4. Fase de entrenamiento
5. Fase de evaluación

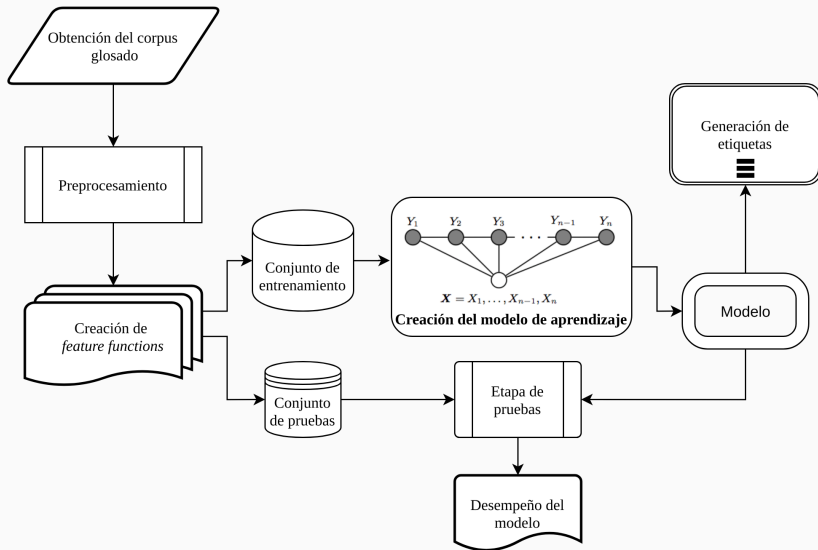


Figure 3: Arquitectura de aprendizaje

Feature Functions

“hi tótsogí” (*No lo he dejado*)

```
[  
  'bias',  
  'letterLowercase=ó',  
  'EOS',  
  'prevpostag=neg',  
  'letterposition=-5',  
  'prevletter=t>',  
  'nxtletter=<t',  
  'nxt2letters=<ts',  
  'nxt3letters=<tso',  
  'nxt4letters=<tsog'  
]
```

- *feature functions* de la letra ó

Evaluación

Se propusieron tres entornos de evaluación:

1. **Baseline:** Las *feature functions* fueron reducidas al mínimo con lo que se **simuló** un *HMM*
2. **POSLess:** Las *feature functions* fueron construidas ignorando la información de las etiquetas POS
3. **LinearCRF:** Toda la información lingüística del etiquetado manual es utilizada para la construcción de las *feature functions*

Para validar el desempeño utilizamos la técnica de *K-folds cross-validation* con $K = 10$ para cada modelo generado.

Resultados

Reportamos el *accuracy* promedio por cada modelo generado en los diferentes entornos de experimentación

Modelo	Accuracy
linearCRF_l2_zero	0.9516
POS_Less	0.9499
baseline_HMMLike_zero	0.8762

Table 3: Comparación de modelos de diferentes entornos con mejor *accuracy*

Función de pérdida

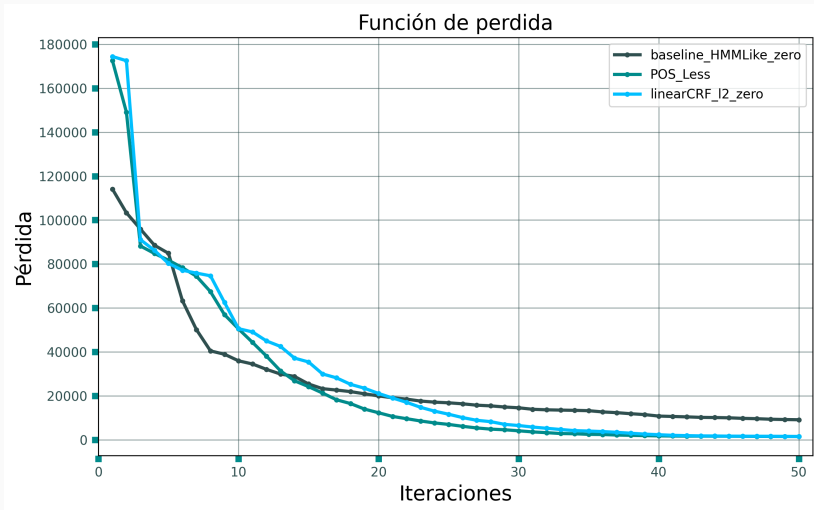


Figure 4: Función de pérdida de los modelos con mejor desempeño

$b\mu$ m-bi-' μ n-gí ya dó-ráhi- ι -'wi
STEM MED-3.ICP-STEM-1.OBJ STEM 1.CPL-STEM
'Cuando me pegaban pues me quitaba'

Figure 5: Modelo linearCRF_l2_zero

- a. má-ndé bi-ni
CTR-F-STEM 3.CPL-STEM
‘*La tarde lo tuve’
- b. mánde bi-ni
STEM 3.CPL-STEM
‘Ayer lo tuve’

Figure 6: Modelo baseline_HMMLike_zero

Conclusiones

- La información lingüística codificada en las *feature functions* es muy importante para **mejorar el desempeño** del etiquetador.

- La información lingüística codificada en las *feature functions* es muy importante para **mejorar el desempeño** del etiquetador.
- Notamos que las etiquetas *POS* parecen **no ser restrictivas** lo cual es bueno para lenguas de bajos recursos digitales

Conclusiones

- La información lingüística codificada en las *feature functions* es muy importante para **mejorar el desempeño** del etiquetador.
- Notamos que las etiquetas *POS* parecen **no ser restrictivas** lo cual es bueno para lenguas de bajos recursos digitales
- Cuando quitamos información en la construcción de las *feature functions* la **frecuencia de las instancias** tiene mayor peso

Conclusiones

- La información lingüística codificada en las *feature functions* es muy importante para **mejorar el desempeño** del etiquetador.
- Notamos que las etiquetas *POS* parecen **no ser restrictivas** lo cual es bueno para lenguas de bajos recursos digitales
- Cuando quitamos información en la construcción de las *feature functions* la **frecuencia de las instancias** tiene mayor peso
- Concluimos que para entornos de bajos recursos digitales, donde la frecuencia de las instancias es menor, es necesario **brindar un contexto amplio** y agregar **información lingüística detallada**

Gracias | Jamädi

¿Dudas?
