Glosador automático usando aprendizaje estructurado para el otomí de Toluca

Diego A. Barriga Martínez[1]
Victor Mijangos Crúz[1]
Ximena Gutierrez-Vasques[2]

5 de Noviembre 2020

Universidad Nacional Autónoma de México[1]

Universidad de Zúrich[2]

Etiquetadores automáticos

Etiquetadores automáticos

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

Etiquetadores automáticos

- 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

Etiquetadores automáticos

- 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*)

Glosado

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

El lenguaje natural

El lenguaje natural

- No obstante, el lenguaje natural es **complejo**
  - Presenta fenómenos que hacen de la construcción de etiquetadores automáticos una tarea difícil

El lenguaje natural

- No obstante, el lenguaje natural es **complejo**
  - Presenta fenómenos que hacen de la construcción de etiquetadores automáticos una tarea difícil
- Tradicionalmente el glosado se hace manualmente
  - Esto es lento y costoso ya que requiere de habilidades de un lingüista y un trabajo intimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller, 2018)

El lenguaje natural

- No obstante, el lenguaje natural es **complejo**
  - Presenta fenómenos que hacen de la construcción de etiquetadores automáticos una tarea difícil
- Tradicionalmente el glosado se hace manualmente
  - Esto es lento y costoso ya que requiere de habilidades de un lingüista y un trabajo intimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller, 2018)
- Adicionalmente, existen escenarios donde los métodos tradicionales no son efectivos como es el caso de los **bajos recursos digitales**

El lenguaje natural

- No obstante, el lenguaje natural es **complejo**
  - Presenta fenómenos que hacen de la construcción de etiquetadores automáticos una tarea difícil
- Tradicionalmente el glosado se hace manualmente
  - Esto es lento y costoso ya que requiere de habilidades de un lingüista y un trabajo intimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller, 2018)
- Adicionalmente, existen escenarios donde los métodos tradicionales no son efectivos como es el caso de los **bajos recursos digitales**
- La construcción de modelos automáticos que asistan este tipo de etiquetado surgen como **una tarea urgente**

El reto: Bajos recursos digitales

El reto: Bajos recursos digitales

- Los bajos recursos digitales son un escenario común en las lenguas mexicanas
  - A pesar de la gran diversidad lingüística, gran parte de las lenguas originarias no poseen contenido web ni publicaciones digitales
  - Esto propicia que carezcan de tecnologías del lenguaje

El reto: Bajos recursos digitales

- Los bajos recursos digitales son un escenario común en las lenguas mexicanas
  - A pesar de la gran diversidad lingüística, gran parte de las lenguas originarias no poseen contenido web ni publicaciones digitales
  - Esto propicia que carezcan de tecnologías del lenguaje
- Este entorno de experimentación supone importante **reto de invetigación**
  - Los métodos tradicionales de etiquetado automático requieren grandes cantidades de datos para funcionar correctamente
  - La escasez de los corpus y falta de normalización puede complicar la tarea

Objetivo

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 automático para el otomí de Toluca**

Objetivo

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

Objetivo

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

Corpus

Corpus

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

Corpus

- 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.

Corpus

- 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/`)

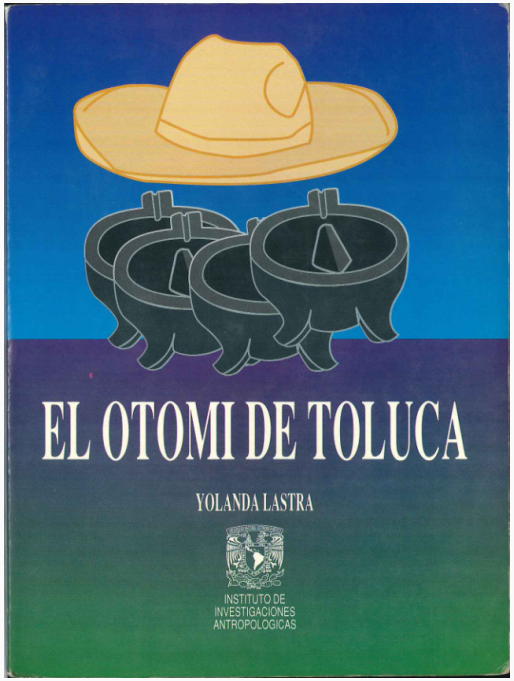

Figure 1: Portada de "El otomí de Toluca" de Lastra

Datos cualitativos del corpus

- La variante en particular es de la región de San Andrés Cuexcontitlan
- Incluye información morfosintáctica (etiquetas POS) y glosa.
- Se agregaron 81 casos poco usuales 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

Arquitectura

1. Obtención del corpus en otomí

Arquitectura

1. Obtención del corpus en otomí
2. Codificación

Arquitectura

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

Arquitectura

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

Arquitectura

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

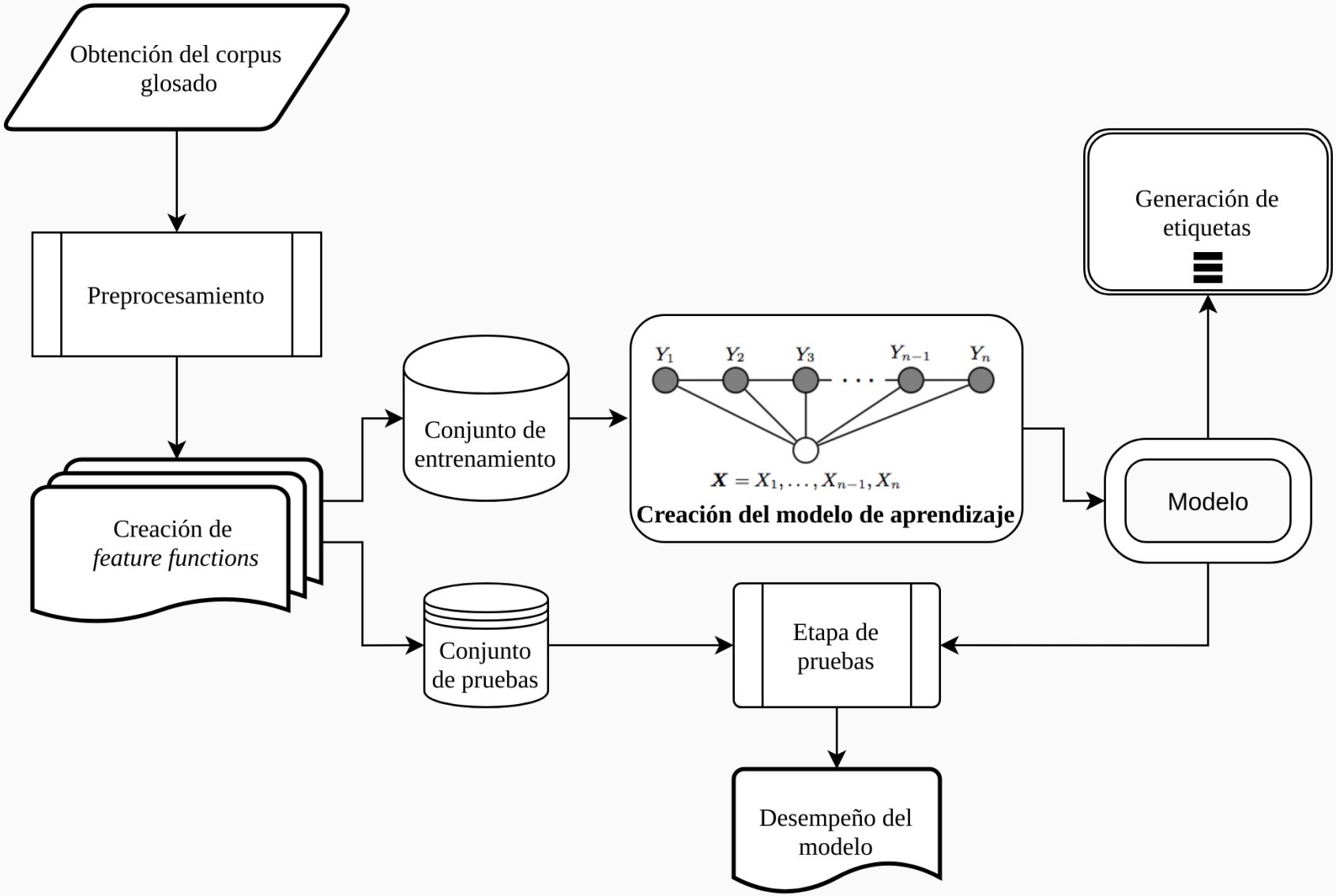

Figure 2: Arquitectura de aprendizaje

Feature Functions

Se obtuvieron *feature functions* por cada letra.

“hi tótsogí” (*No lo he dejado*)

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

Feature functions para la letra ó

Evaluación

Evaluación

Se propusieron tres entornos de evaluación:

1. **Baseline**: Las *feature functions* fueron reducidas al mínimo
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

Resultados

Reportamos el *accuracy* promedio de cada modelo generado