Glosador automático usando aprendizaje estructurado para el otomí de Toluca

Diego A. Barriga Martínez♦
Víctor Mijangos de la Cruz♦
Ximena Gutierrez-Vasques♣

5 de Noviembre 2020

Universidad Nacional Autónoma de México♦

Universidad de Zúrich♣

Etiquetadores automáticos

Etiquetadores automáticos

- El etiquetado automático es una tarea que asigna etiquetas a partes de un texto. Estas etiquetas agregan información lingüística a estos elementos
    - Esta tarea es un paso importante para el descubrimiento de las estructuras lingüísticas de un corpus

Etiquetadores automáticos

- El etiquetado automático es una tarea que asigna etiquetas a partes de un texto. Estas etiquetas agregan información lingüística a estos elementos
    - Esta tarea es un paso importante para el descubrimiento de las estructuras lingüísticas de un corpus
- Es usual que se utilicen métodos de *Machine Learning (ML)* para la construcción de etiquetadores automáticos

Etiquetadores automáticos

- El etiquetado automático es una tarea que asigna etiquetas a partes de un texto. Estas etiquetas agregan información lingüística a estos elementos
    - Esta tarea es un paso importante para el descubrimiento de las estructuras lingüísticas de un corpus
- Es usual que se utilicen métodos de *Machine Learning (ML)* para la construcción de etiquetadores automáticos
- 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 brinda información acerca del significado y propiedades gramaticales de las palabras

Glosado

- El glosado es un tipo de etiquetado que brinda información acerca del significado y propiedades gramaticales de las palabras
- Este etiquetado de textos es de suma importancia para el análisis y la documentación lingüística

Glosado

- El glosado es un tipo de etiquetado que brinda información acerca del significado y propiedades gramaticales de las palabras
- 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 intimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller and Hulden 2018)

Glosado

- El glosado es un tipo de etiquetado que brinda información acerca del significado y propiedades gramaticales de las palabras
- 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 intimo con hablantes nativos, los cuales, requieren capacitación en lingüística básica y de software (Moeller and Hulden 2018)
- La construcción de modelos automáticos que asistan este tipo de etiquetado surgen como **una tarea importante**

El lenguaje natural

El lenguaje natural

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

El lenguaje natural

- 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 investigación**

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 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 para el otomí de Toluca**
- Diseño e implementación de un **etiquetador morfológico** 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 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, McCallum, and Pereira 2001)

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 Mijangos de la 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 Mijangos de la Cruz.
- El corpus es un subconjunto del corpus paralelo español-otomí que se encuentra en la plataforma web Tsunkua (`https://tsunkua.elotl.mx/`)

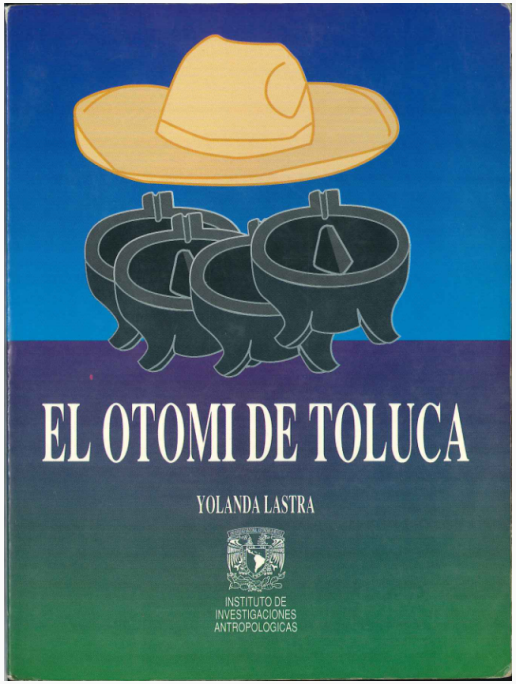

Figure 2: Portada de “El otomí de Toluca” de Lastra

Convenciones

IPA	ɨ	ɛ	ɔ	ʌ	ɵ
Ortografía práctica	u̲	e̲	a̲	i̲	o̲
Convención para este trabajo	μ	ϵ	α	ι	

Figure 3: Representación de cada vocal en IPA (alfabeto fonético internacional)

Datos cualitativos del corpus

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

Datos cualitativos del corpus

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

Datos cualitativos del corpus

- 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

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

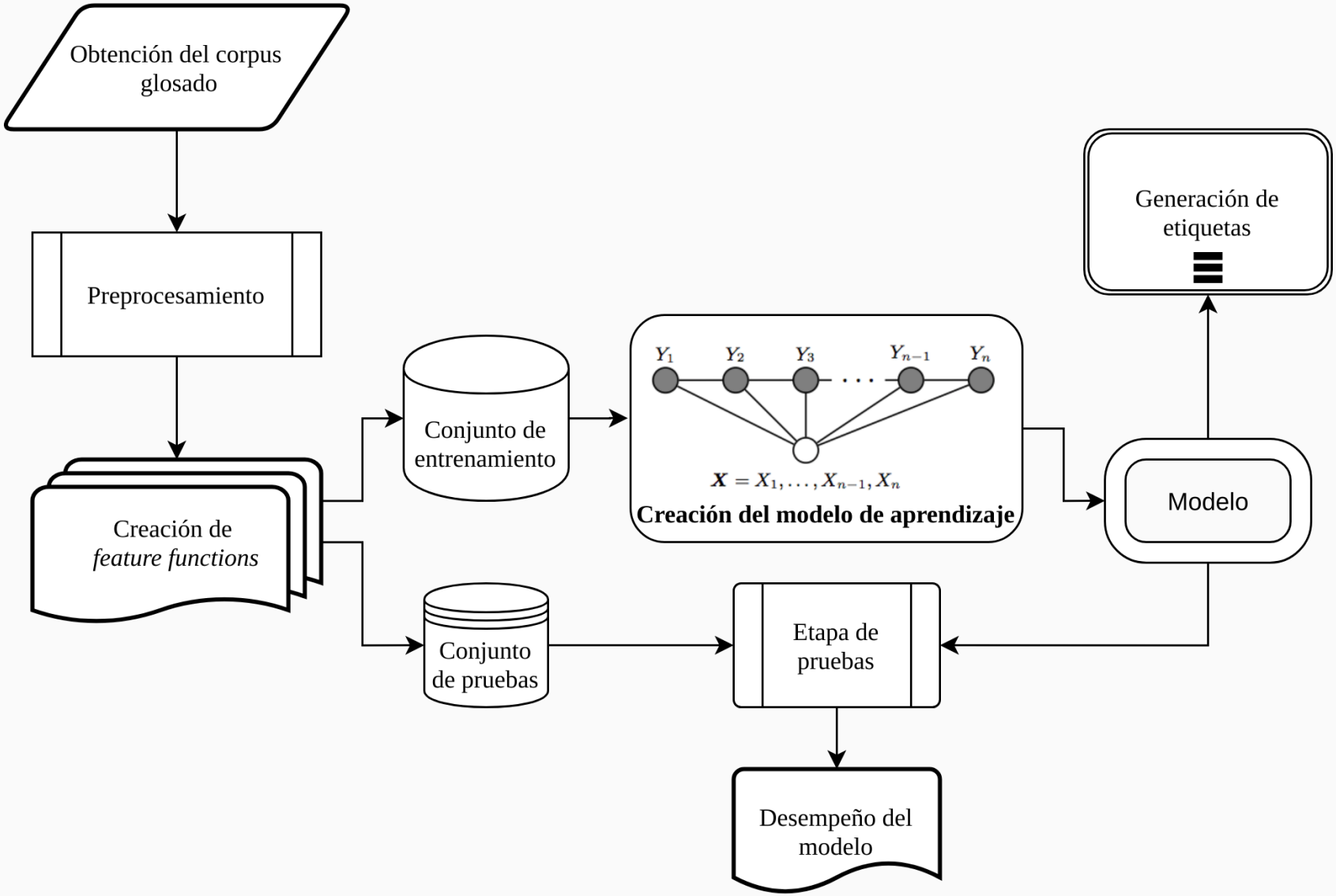

Figure 4: Arquitectura de aprendizaje

Feature Functions

“hi t**ó**tsogí” (*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**

```
[
  [["hi", "stem"], "neg"],
  [
    ["tó", "1.prf"],
    ["tsogí", "stem"],
    "v"
  ]
]
```

- Frase glosada en el corpus

Evaluación

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

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 perdida

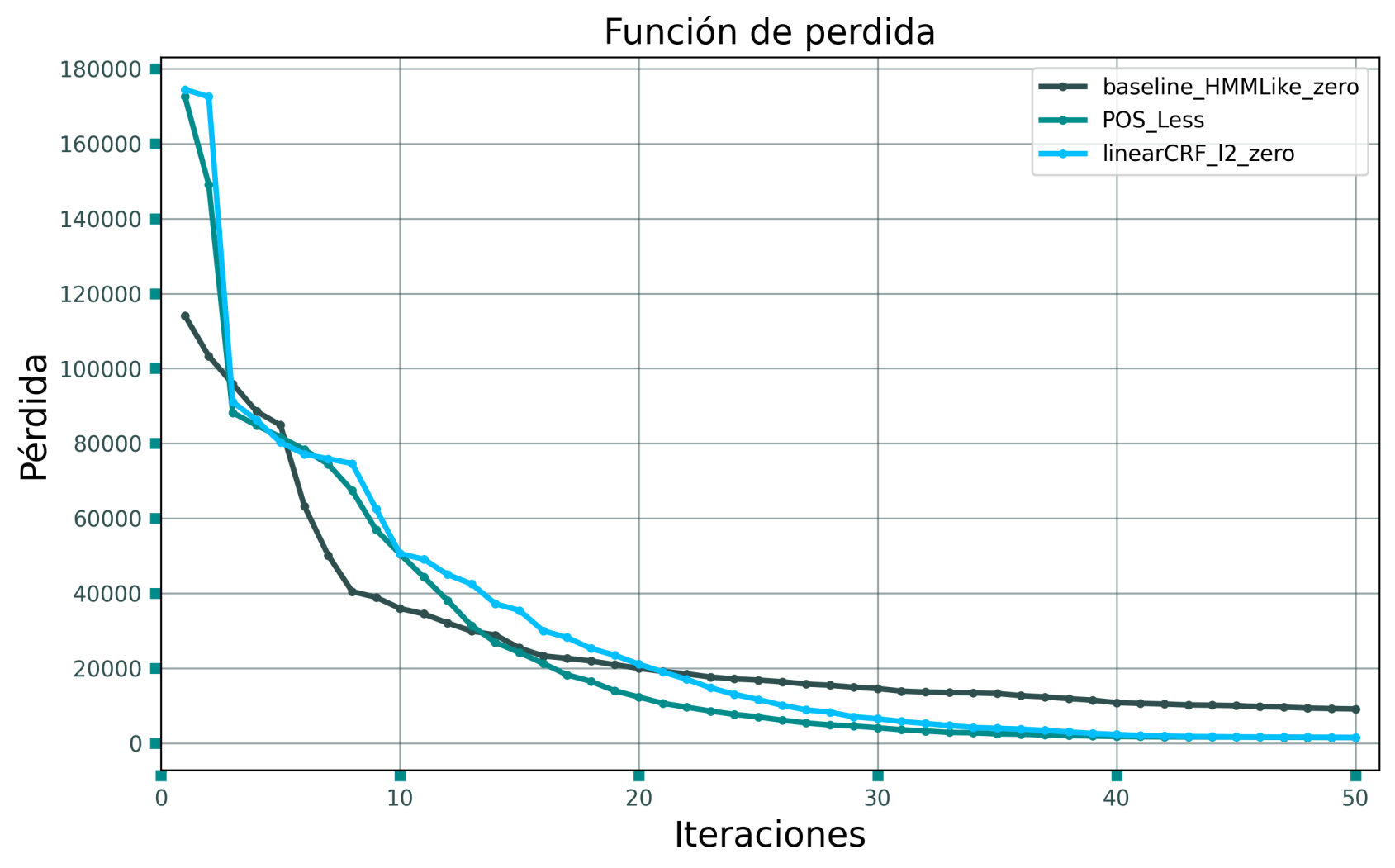

Figure 5: Función de perdida de los modelos con mejor desempeño

Ejemplos

bμ	m-bi-'μn-gí		ya	dó-ráhi-ι-'wi
STEM	MED-3.ICP-STEM-1.OBJ		STEM	1.CPL-STEM
'Cuando me pegaban pues me quitaba'				

Figure 6: Modelo linearCRF_l2_zero

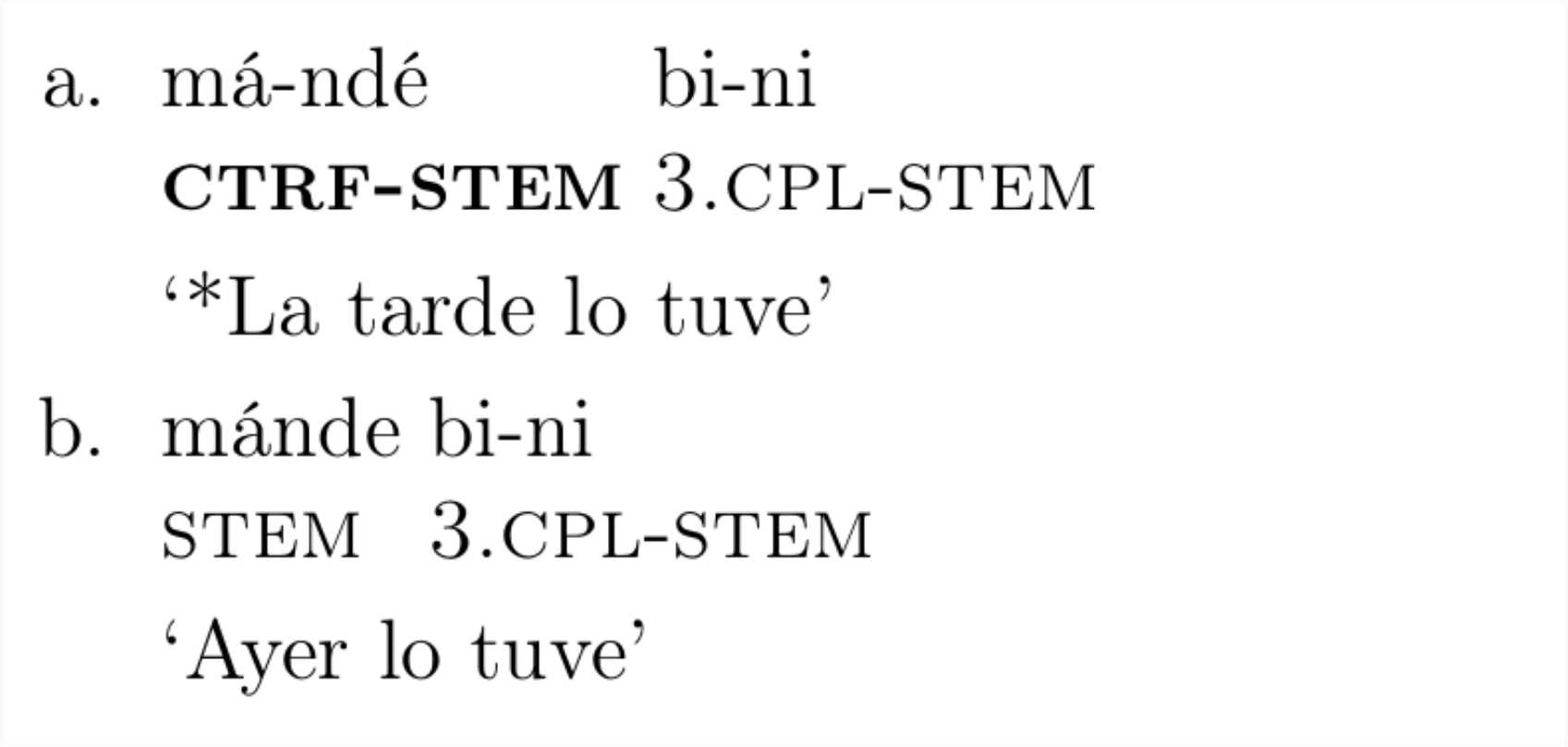

Figure 7: Modelo baseline_HMMLike_zero

Conclusiones

Conclusiones

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

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

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**

Gracias | Jamädi

¿Dudas?

Bibliografia

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data."

Lastra, Yolanda. 1992. *El Otomí de Toluca*. Instituto de Investigaciones Antropológicas, UNAM.

Moeller, Sarah, and Mans Hulden. 2018. "Automatic Glossing in a Low-Resource Setting for Language Documentation." In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 84–93.