Experiencia codificación automática INE Chile

Red de transmisión del conocimiento CEPAL

Codificación automatizada

Enero 2020

Contenidos

- Metodología para la codificación automática
- Algunos resultados

Énfasis en la metodología

Introducción

Algunos datos levantados en el marco de la producción estadística corresponden a texto "libre"

- Ocupación
- Actividad económica
- Producción
- Consumo

Estos datos deben ser codificados, para que sean útiles desde el punto de vista estadístico

Auxiliar de aseo. Limpió y realizó orden en instalaciones

¿Cómo se realiza la tarea de codificación?

Introducción

Contamos con clasificadores para esta tarea

- CIUO
- CAENES (CIIU para encuestas de hogares)
- CIIU
- CCIF (COICOP)

Permiten ordenar jerárquicamente distintos fenónemos

Auxiliar de aseo. Limpió y realizó orden en instalaciones

El clasificador CIUO nos dice que esto corresponde al Gran Grupo 9

Históricamente, esta tarea ha sido intensiva en trabajo manual

Alternativas

Las oficinas de estadísticas abordan la tarea de codificación mediante distintas estrategias

- Codificación manual
- Codificación automática
- Codificación manual asistida

Alternativas

Las oficinas de estadísticas abordan la tarea de codificación mediante distintas estrategias

- Codificación manual
- Codificación automática
- Codificación manual asistida

Codificación automática

- Por reglas
- Aprendizaje de máquinas (*machine learning*)

En el sistema basado en reglas el analista debe generar a priori las reglas

Codificación por reglas

Si están presentes *auxiliar de aseo* y *limpió*, uso el código **91**

Si está presente *obrero agrícola*, uso el código **92**

```
if (str_detect(texto, "auxiliar de aseo") &
    str_detect(texto, "limpió" )) {
  codigo <- 91
}
```

```
if (str_detect(texto, "obrero agrícola")) {
  codigo <- 92
}
```

✔ Implementa exactamente los criterios del clasificador

✘ Requiere de una gran cantidad de trabajo en la elaboración de las reglas

✘ Es muy costoso llegar a la totalidad de los casos

Codificación por *machine learning*

En lugar de que el analista genere las reglas, dejamos que un algoritmo las "aprenda" a partir de los datos

oficio	tareas	codigo
ingeniero de proyectos informaticos.	monitoreo de proyectos los revisa los avances.	25
chófer de taxi	transpotar pasajeros, revisar ruta a recorrer, realizar encomienda	83
gasfiter	realizo mantencion e instalacion de estanques de gas	71
ejecutiva de ventas	atender cliente, mostrar productos	52
obrero agricola	podar plantas de arándanos con tijera manual	92
repostera	mezclar ingredientes, preparar tortas, pasteles, rellenar	75

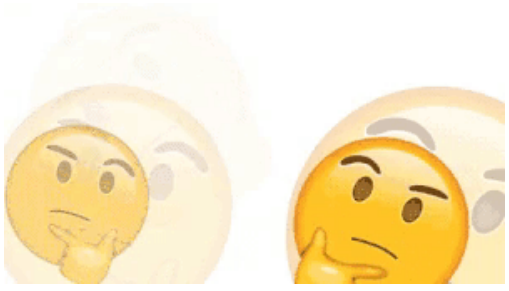
Si tenemos una cantidad razonable de ejemplos, es posible que nuestro algoritmo aprenda las reglas de clasificación

Codificación por *machine learning*

oficio	tareas	codigo
ingeniero de proyectos informaticos.	monitoreo de proyectos los revisa los avances.	¿?
chófer de taxi	transpotar pasajeros, revisar ruta a recorrer, realizar encomiendas	¿?
gasfiter	realizo mantencion e instalacion de estanques de gas	¿?
ejecutiva de ventas	atender cliente, mostrar productos	¿?
obrero agricola	podar plantas de arándanos con tijera manual	¿?
repostera	mezclar ingredientes, preparar tortas, pasteles, rellenar	¿?

Con nuestro algoritmo ya entrenado, podemos predecir la etiqueta de datos nuevos

¿Qué pasa si los datos nuevos no se parecen a los del entrenamiento?



Pre procesamiento

Edición de las glosas originales

- Remover signos de puntuación
- Conversión a minúscula
- Remover palabras con poco significado (*stopwords*)

glosa	glosadep
GOBERNACION MARITIMA (DIRECTEMAR)	gobernacion maritima directemar
DISTRIBUIDORA DE CECINAS (WINTER)	distribuidora cecinas winter
ELABORACION DE JUGOS NATURALES Y SANEWICHEROS	elaboracion jugos naturales sanewicheros

Necesitamos representar los textos de manera numérica

Bolsa de palabras (TF-IDF) y Word embeddings

Esquema TF-IDF

Una palabra es importante cuando:

- Se repite mucho dentro de una glosa
- Se repite poco dentro de todas las glosas

	V1	gobernacion	maritima	directemar	elaboracion	jugos	naturales	elaboraciond
1	gobernacion maritima directemar	3.42898161498785	2.57510965066609	4.78596301598098	0	0	0	0
2	elaboracion jugos naturales sanewicher...	0	0	0	1.43995252579	2.77418118543287	2.92413001832303	0
3	elaboraciond jugos naturales sandwiches	0	0	0	0	2.77418118543287	2.92413001832303	3.97304965933812
4	servicio mantencion soldadura calderia ...	0	0	0	0	0	0	0
5	universidad	0	0	0	0	0	0	0
6	rectificacion motores camion alto tonel...	0	0	0	0	0	0	0
7	cafeteria pasteleria	0	0	0	0	0	0	0
8	gobernacion provincial iquique	3.42898161498785	0	0	0	0	0	0
9	ministerio obras publicas	0	0	0	0	0	0	0
0	servicios psicologia	0	0	0	0	0	0	0

Esquema word embeddings

La representación TF-IDF no captura la semántica de las palabras

Pensemos en las siguientes glosas de actividad económica:

- *elaboración vienesas* (expresión chilena)
- *fabricación salchichas*

En el esquema TF-IDF ambos vectores son ortogonales

Para una persona es trivial determinar que ambos textos son muy similares

Word embeddings es una técnica que permite construir representaciones mucho más ricas semánticamente

Estamos usando el trabajo del profesor Jorge Perez

Esquema word embeddings

La representación de la palabra *fabricación* luce así:

```
array([ 0.21481  , -0.62736  , -0.4004   , -0.08525  ,  0.12641  ,
        0.13286  ,  0.066812 ,  0.19013  , -0.28346  , -0.54088  ,
        0.11488  ,  0.034526 ,  0.29506  ,  0.073115 ,  0.12948  ,
        0.35718  , -0.23242  , -0.20224  ,  0.12519  , -0.096514 ,
       -0.36164  ,  0.22901  , -0.11947  ,  0.053282 ,  0.35719  ,
        0.28415  , -0.30769  , -0.077862 , -0.16879  ,  0.45927  ,
        0.3204   , -0.18164  ,  0.023082 , -0.43119  ,  0.20989  ,
       -0.41418  ,  0.23208  ,  0.25281  ,  0.24712  ,  0.088254 ,
        0.081171 ,  0.24265  ,  0.066244 , -0.12523  ,  0.15953  ,
       -0.1294   ,  0.26078  , -0.25779  ,  0.24374  ,  0.67421  ,
       -0.11043  , -0.2476   ,  0.099461 ,  0.07903  , -0.24899  ,
        0.013924 ,  0.027634 , -0.33037  , -0.06468  , -0.16753  ,
       -0.083145 , -0.61762  , -0.001967 , -0.21073  , -0.11562  ,
       -0.20126  , -0.35138  ,  0.075967 , -0.11816  , -0.31851  ,
       -0.29583  ,  0.39431  , -0.55547  , -0.017207 , -0.53197  ,
       -0.095595 ,  0.12578  ,  0.45695  ,  0.10897  , -0.15482  ,
        0.064061 , -0.13635  ,  0.057695 ,  0.026385 , -0.22356  ,
        0.0013576,  0.034052 , -0.20792  ,  0.046271 , -0.085871 ,
       -0.29042  , -0.35169  ,  0.0028087,  0.41498  , -0.21869  ,
```

Vector de 300 dimensiones con números entre -1 y 1

Esquema word embeddings

Es necesario generar una representación única para cada glosa

Varias alternativas:

- Media de cada una de las dimensiones
- Mínimo y máximo de cada dimensión (vector de 600 dimensiones)
- Media ponderada por tf-idf

extracción producción cobre

```
##                  [,1]       [,2]       [,3]       [,4]       [,5]
## extraccion -0.1261270 -0.1392765 -0.4185846  0.7903040 -0.9150484
## produccion -0.8848647  0.3217051  0.0764309 -0.4266615 -0.4528008
## cobre       0.2953896 -0.3367866 -0.5318304  0.5798521  0.5074388
```

Estamos utilizando mínimos y máximos

Nuestra metodología

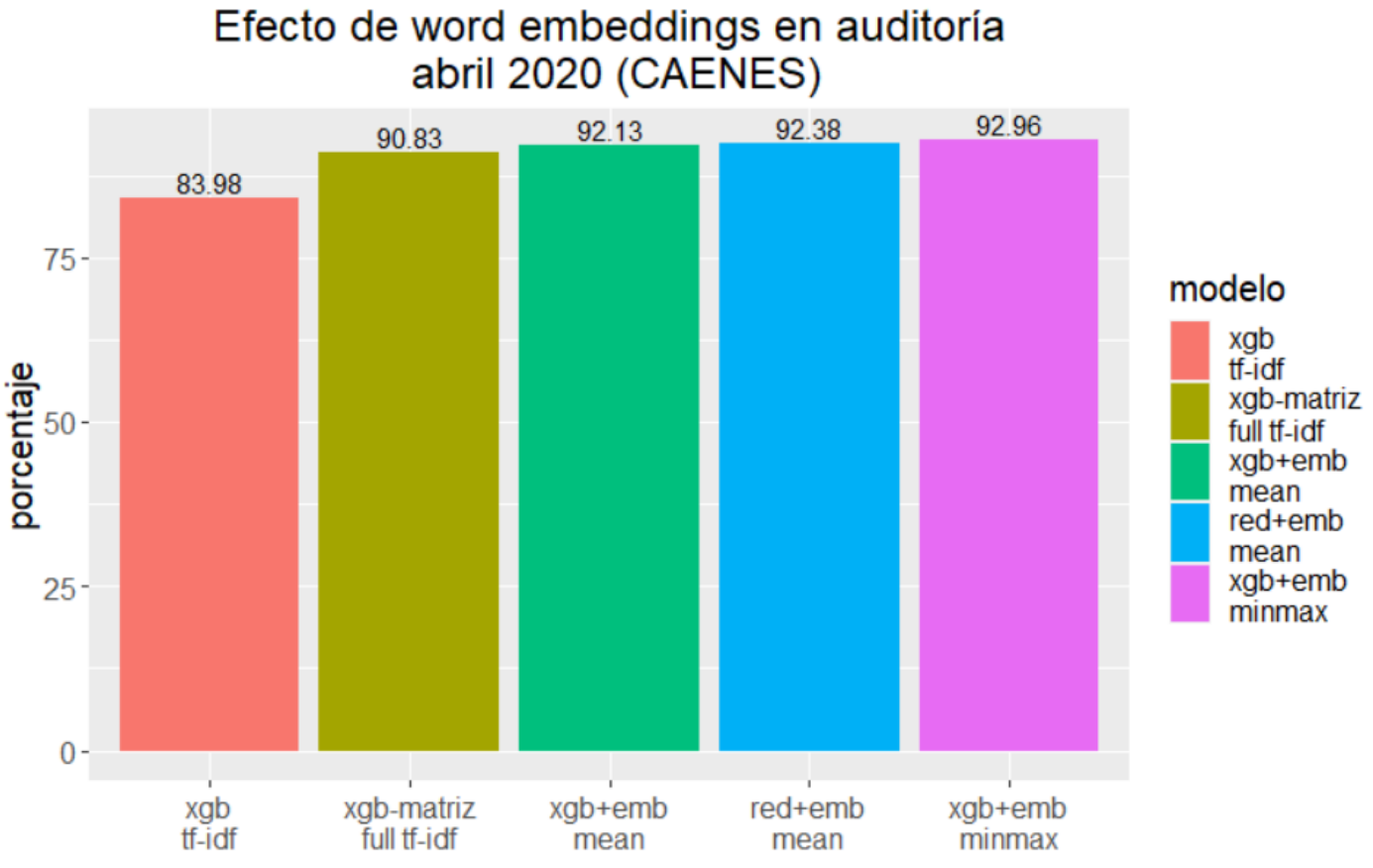
- Representación de textos: *Word embeddings*
- Mínimos y máximos de cada dimensión
- Algoritmos usados: **xgboost** o **redes neuronales**, dependiendo de las características del dataset de entrenamiento

La prueba para cualquier algoritmo es el desempeño en datos que nunca ha visto

Contamos con datos (1200 aproximadamente) etiquetados de abril 2020 para CAENES (Actividad Económica)

Provienen de una auditoría realizada manualmente

Resultados CAENES

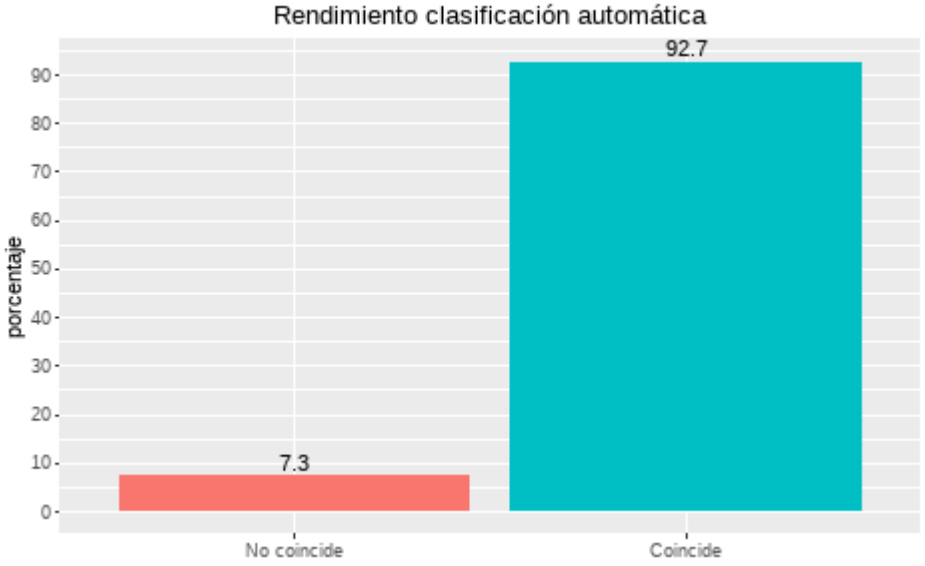


Utilizar word embeddings mejora el rendimiento de un algoritmo

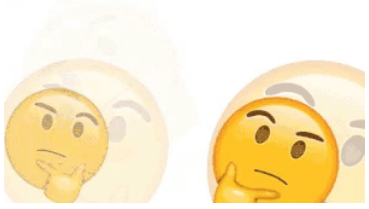
Lo más importante es que mejora la estabilidad de la predicción

Resultados CCIF

La Encuesta de Presupuestos Familiares (EPF) utiliza un clasificador llamado CCIF (COICOP), que tiene más de 1.000 categorías



¿Cómo podemos mejorar el resultado?



Resultados CCIF

Idea: Puedo codificar los registros difíciles a mano y dejarle los fáciles a mi algoritmo

Estrategia mixta: parte automática y parte manual

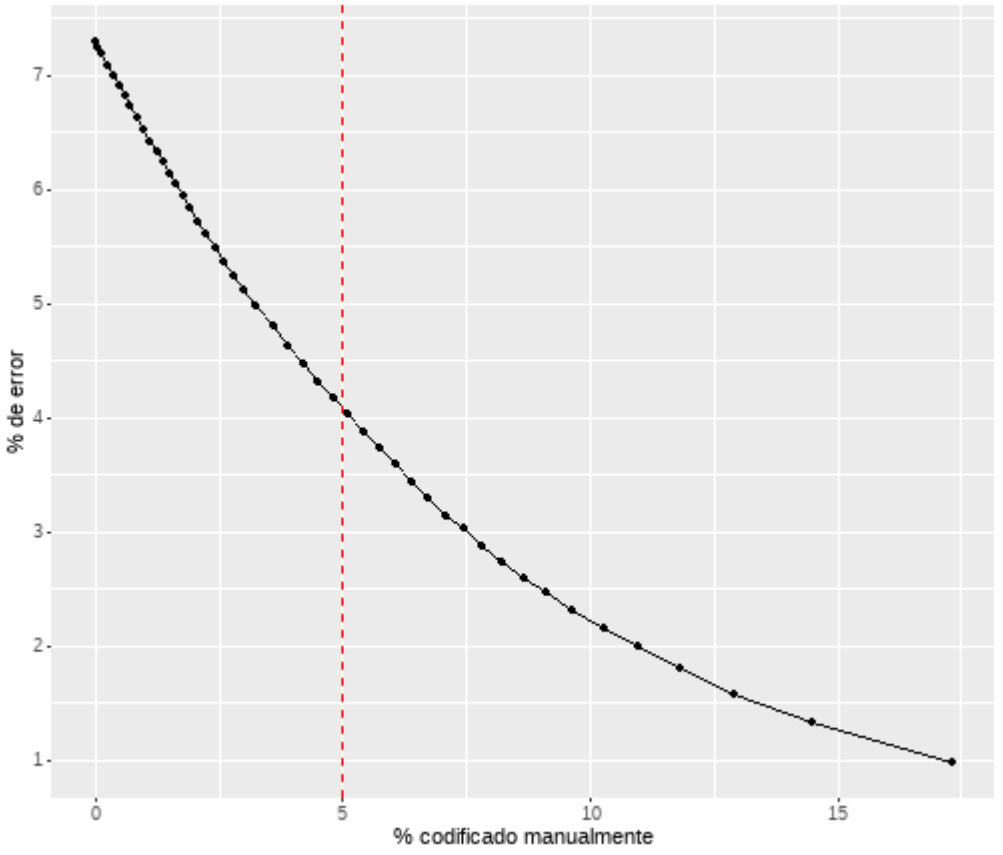
¿Cómo puedo elegir qué registros debo codificar manualmente?

Apartar registros complejos

Receta:

- **paso 1:** Rankear los registros de mayor a menor dificultad (probabilidad asignada por la red)
- **paso 2:** Elegir un punto de corte
- **paso 3:** Codificamos automáticamente solo lo que está sobre el punto de corte (más fáciles)
- **paso 4:** Calcular el error

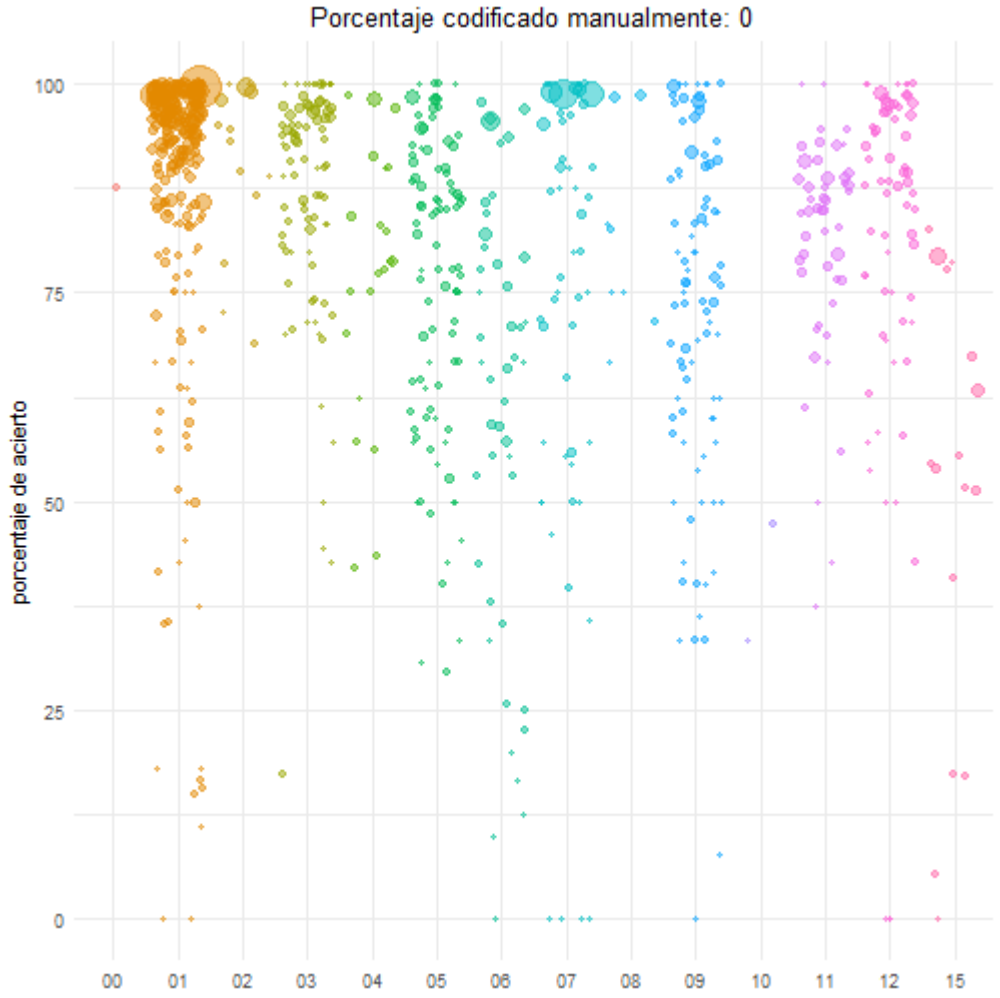
Resultados CCIF



Si clasificamos el 5% a mano, mejoramos más de 3 pp

Resultado CCIF por categoría

¿Cuál es el resultado a nivel de categoría?



Problemas durante el entrenamiento

Uno de los grandes problemas para la inteligencia artificial son los datos de entrenamiento

Retomemos la pregunta respecto a la diferencia entre los datos de entrenamiento y los predichos

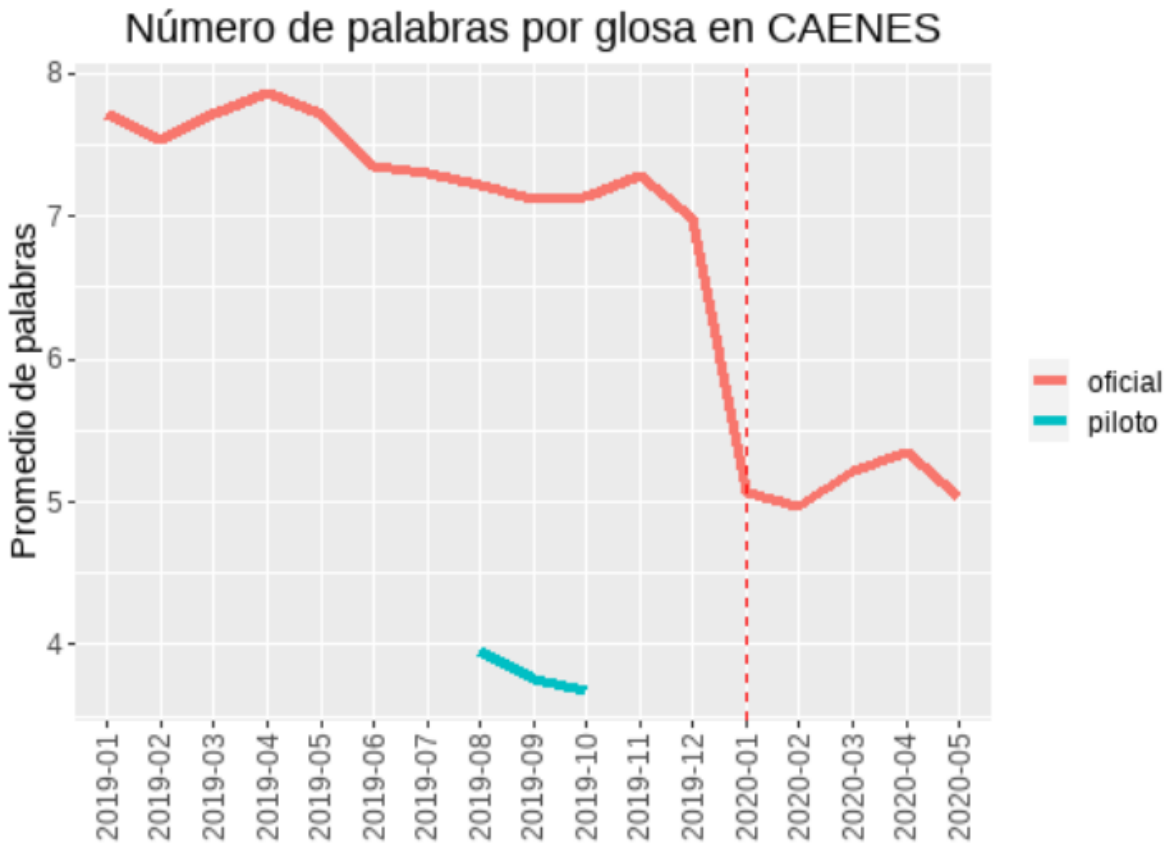
En enero de 2020, la Encuesta de Empleo comenzó a transitar hacia un levantamiento vía dispositivo móvil

Esto es algo positivo, pero introduce desafíos para la codificación

Si los textos levantados mediante tablet han cambiado respecto a los de papel, la predicción se dificulta

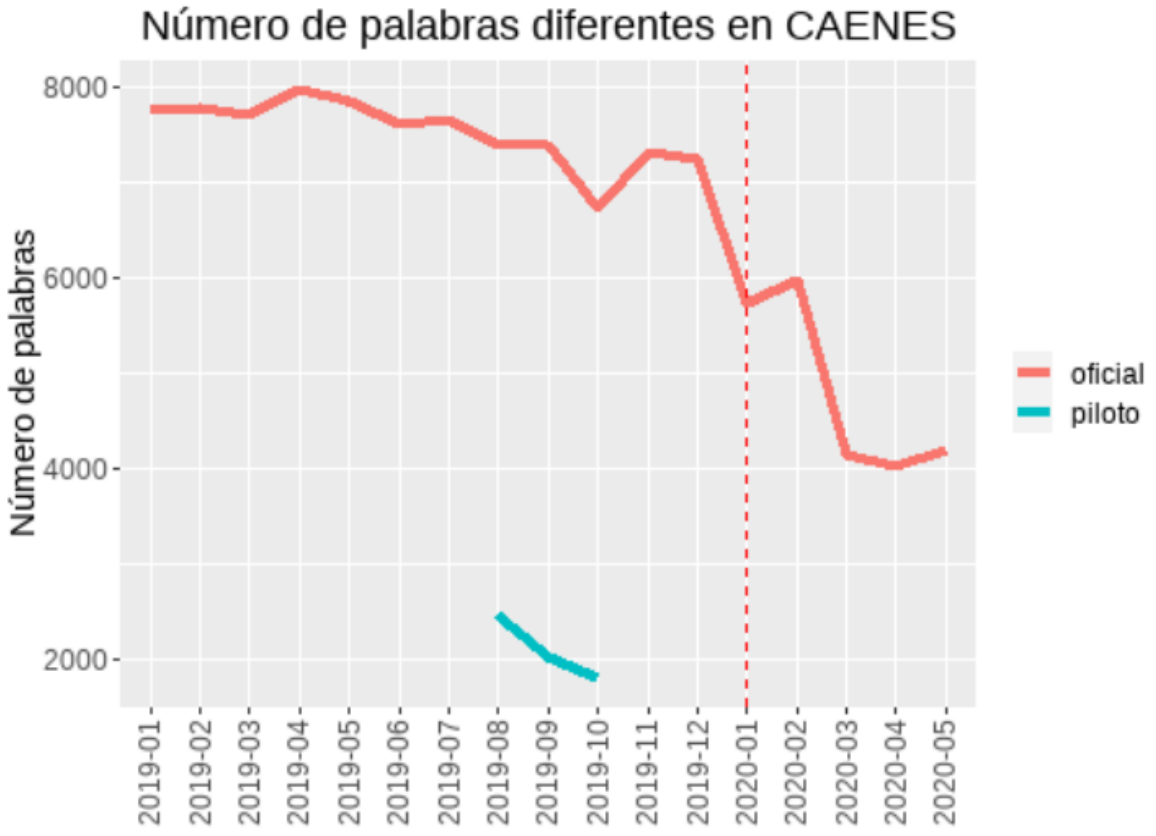
Cambios en el registro

Analizamos los registros de actividad económica a lo largo de varios meses



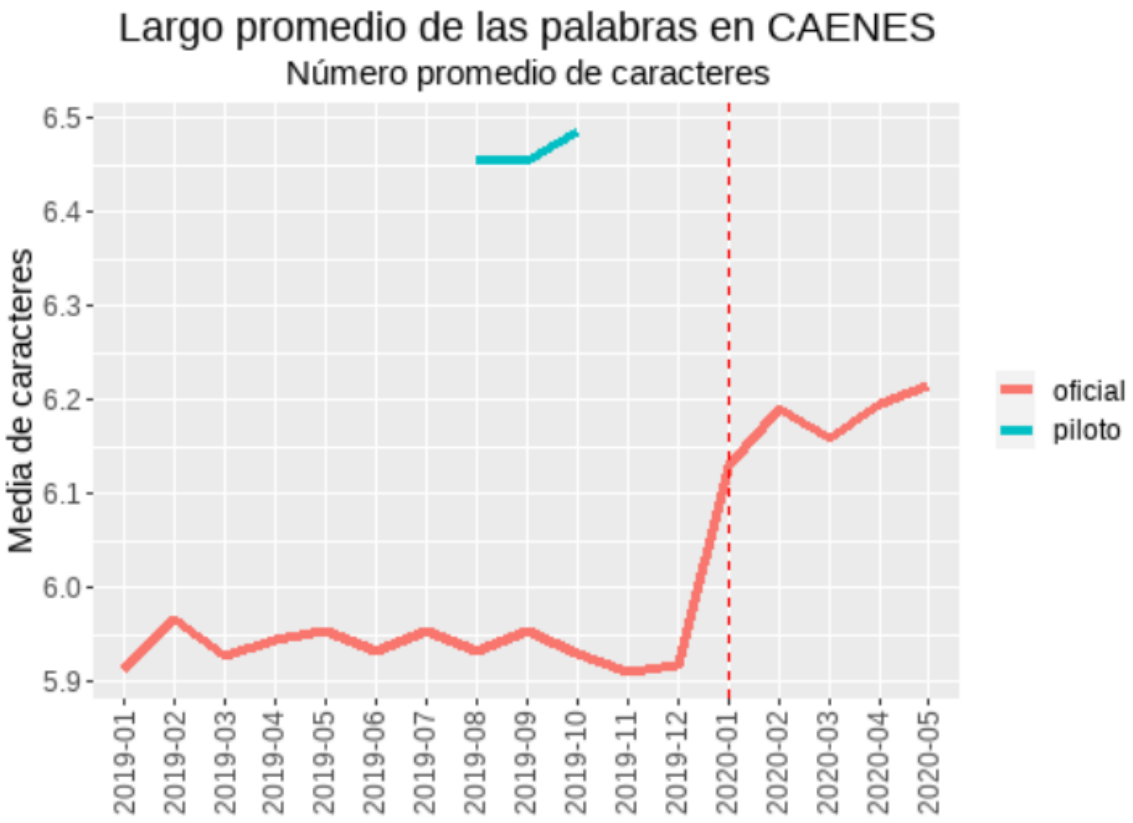
Tenemos menos palabras por glosa

Cambios en el registro



Menor variedad de palabras

Cambios en el registro



Palabras más largas

Cambios en el registro

- Glosas más cortas
- Palabras más largas
- Menor variedad de palabras

Hipótesis:

Los encuestadores han transitado hacia un lenguaje más sucinto y con palabras que posiblemente estén más cargadas de significado.

En principio, esto no es bueno ni malo. El lenguaje es dinámico

Debemos implementar técnicas flexibles que se ajusten a esta realidad

Conclusiones

Existe espacio para implementar nuevas mejoras en el proceso

Es importante monitorear las transformaciones en el modo de registro

Actualización periódica de los datos de entrenamiento

- Actualmente, nuestra institución se encuentra desarrollando esta tarea

Es razonable utilizar una estrategia mixta de codificación

- Identificar los registros de mayor complejidad
- Codificar la mayor parte automáticamente y dejar lo difícil para un codificador entrenado

Existe un proyecto transversal en la institución que actualmente está estudiando estos temas

Proyecto Estratégico de Servicios Compartidos

Experiencia codificación automática INE Chile

Red de transmisión del conocimiento CEPAL

Codificación automatizada

Enero 2020