

Optimización de viajes compartidos en taxis utilizando algoritmos evolutivos

Gabriel Fagúndez de los Reyes Renzo Massobrio

Facultad de Ingeniería,
Universidad de la República,
Montevideo, Uruguay



Contenido

- 1 Introducción
- 2 Descripción del problema
- 3 Métodos para la resolución del problema
- 4 Evaluación
- 5 Conclusiones y trabajo futuro

Acentuación ortográfica

La acentuación ortográfica en algunas palabras del idioma español elimina la ambigüedad de una oración.

Clasificación de palabras con acento ortográfico

- Sin ambigüedad: única forma correcta de escribir estas palabras, p.ej. *acentuación*.
- Con ambigüedad: cambian su significado si son escritas con o sin acentuación ortográfica, p.ej.: verbos (*canto/cantó, hable/hablé*), sustantivos (*papa/papá, secretaria/secretaría*).

Restauración de acentos ortográficos en palabras con ambigüedad

- No es trivial.
- Involucra aspectos de desambiguación de significado.
- Requiere examinar el contexto de cada palabra.

Adverbios interrogativos

- Gran dependencia con el contexto en el que aparecen.
- Particularmente difíciles de desambiguar.

¿Qué es un adverbio interrogativo?

- Los adverbios son palabras invariables que complementan el significado de un verbo, un adjetivo u de otro adverbio.
- Pueden funcionar de forma interrogativa, p.ej. *¿dónde nació?*
- Pueden formularse de forma directa o indirecta, p.ej. *¿Adónde os marcháis?* o *Dime adónde saldréis.*
- En una frase pueden presentarse adverbios interrogativos y no interrogativos, p.ej. *¿por **qué** algunas enfermedades de origen vírico, **como** los catarros o la gripe, pueden sufrirse en repetidas ocasiones?*

Problema a resolver

Dado un texto del que fueron quitados todos los acentos ortográficos de sus adverbios, se debe clasificar cada palabra en una de las siguientes clases:

- **O**. Toda palabra que no un adverbio.
- **SIN_TILDE**. Si se trata de un adverbio no interrogativo.
- **CON_TILDE**. Si se trata de un adverbio interrogativo.

Corpus de trabajo

- Basado en la unión del corpus CESS Treebanks y CoNLL 2002.
- Consta de un total de aprox. 560,000 tokens, de los cuales:
 - 18,000 son adverbios no interrogativos.
 - 240 son adverbios interrogativos.
- Gran mayoría de los tokens del corpus no son adverbios.
 - 99,96 % de éxito en clasificador de línea base.

Clasificadores propuestos

- Se aborda el problema de la restauración de acentos ortográficos como un problema de clasificación.
- Según nuestro conocimiento no existen antecedentes de trabajos previos orientados a resolver el problema planteado.
- Se presentan dos técnicas de aprendizaje automático para la resolución del problema:
 - Clasificador basado en Support Vector Machines (SVM).
 - Clasificador basado en Conditional Random Fields (CRF).

Clasificador basado en SVM

- Implementado utilizando la herramienta SVM^{light} y SVMTool.
 - SVMTool es un generador de etiquetadores de secuencias.
- Atributos utilizados para la clasificación.
 - Ventana de 5 tokens centrada en el token a etiquetar.
 - Etiquetas de los dos tokens previos al token a etiquetar.
 - Bigramas y trigramas de tokens y etiquetas de tokens.
 - (t_{-2}, t_{-1}) , (t_0, t_{+1}) , (e_{-2}, e_{-1}) , (t_{-1}, t_{+1}, t_{+2}) , etc.
 - Información de puntuación en la oración.
 - Información tipográfica, p.ej.: mayúsculas, minúsculas, etc.

| | | | | |
|----------|----------|-------|----------|----------|
| e_{-2} | e_{-1} | | | |
| t_{-2} | t_{-1} | t_0 | t_{+1} | t_{+2} |

¿Cómo no nos **habremos** dado cuenta?

Clasificador basado en CRF

- Implementado utilizando la herramienta MALLET.
- Atributos utilizados para la clasificación.
 - Conjunciones de atributos del token anterior y siguiente.
 - $\text{conj}(-1,0)$ y $\text{conj}(0,+1)$.
 - Atributo que marca la presencia de un token que generalmente precede o sucede a un adverbio.
 - PREV-SINT, NEXT-SINT, PREV-CONT y NEXT-CONT.
 - Información de puntuación en la oración.
 - Información tipográfica, p.ej. mayúsculas, minúsculas, etc.

$\text{conj}(-1,0)$
¿Cómo no nos **habremos** dado cuenta?
 $\text{conj}(0,+1)$

Metodología de evaluación

- Se dividió el corpus en 10 partes de tamaño similar.
- Se realizaron 10 entrenamientos con $\frac{9}{10}$ del corpus y se evaluó utilizando el $\frac{1}{10}$ restante.
- Se utilizaron métricas clásicas para la evaluación.
 - Precisión, Recall y Medida-F.

Cuadro: Matriz de confusión del clasificador basado en SVM.

| | O | SIN_TILDE | CON_TILDE |
|------------------|----------|------------------|------------------|
| O | 54322,1 | 0,0 | 0,0 |
| SIN_TILDE | 0,0 | 1867,4 | 0,3 |
| CON_TILDE | 0,0 | 19,3 | 4,5 |

Cuadro: Matriz de confusión del clasificador basado en CRF.

| | O | SIN_TILDE | CON_TILDE |
|------------------|----------|------------------|------------------|
| O | 54322,1 | 0,0 | 0,0 |
| SIN_TILDE | 0,0 | 1867,7 | 0,8 |
| CON_TILDE | 0,1 | 16,0 | 7,7 |

Análisis experimental

- Confusión en la clasificación de adverbios interrogativos.
 - 98,46 % de errores del clasificador basado en SVM.
 - 95,26 % de errores del clasificador basado en CRF.
- Mayores causas de este tipo de error.
 - Adverbios interrogativos indirectos.
 - Adverbios no interrogativos contenidos en frases interrogativas.

| Etiqueta | Precisión | | Recall | | $F_{0,5}$ | |
|-----------|-------------|------|--------|-------------|-----------|-------------|
| | SVM | CRF | SVM | CRF | SVM | CRF |
| O | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SIN_TILDE | 0.99 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 |
| CON_TILDE | 0.94 | 0.91 | 0.18 | 0.33 | 0.31 | 0.48 |

- Ambos clasificadores presentan una alta *Precisión*.
 - Se comete una cantidad muy pequeña de errores de Tipo I.
- El clasificador basado en CRF presenta mejores resultados para la métrica de *Recall*.
 - El clasificador basado en CRF comete una menor cantidad de errores de Tipo II.

| Adverbio | Cantidad | Precisión | | Recall | | $F_{0,5}$ | |
|----------|----------|-------------|-------------|--------|-------------|-----------|-------------|
| | | SVM | CRF | SVM | CRF | SVM | CRF |
| qué | 142 | 0.94 | 0.92 | 0.23 | 0.42 | 0.36 | 0.57 |
| cómo | 72 | 1.00 | 0.89 | 0.14 | 0.23 | 0.24 | 0.36 |
| dónde | 18 | 0.67 | 1.00 | 0.11 | 0.17 | 0.19 | 0.29 |
| cuándo | 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| cuánto | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

- Mejores resultados con mayor cantidad de ejemplos.
- Adverbio *qué* cuenta con la mayor cantidad de ocurrencias y es con el que se obtienen los mejores resultados.
 - ¿Es necesario aumentar el tamaño del corpus?

Trabajo realizado y resultados obtenidos

- Se presentó el problema de la restauración automática de acentos ortográficos en adverbios interrogativos.
- Se construyó un corpus de trabajo.
- Se propusieron dos implementaciones para su resolución: clasificador basado en SVM y clasificador basado en CRF.
- En promedio valores de $F_{0,5}$ para adverbios interrogativos de 0,48 con CRF y 0,31 con SVM.
- Resultados mejoran al aumentar ejemplos de entrenamiento.
- Llegando a valores de $F_{0,5}$ de hasta 0,57 con CRF.

Principales líneas de trabajo a futuro

- Realizar un análisis estadístico de los resultados obtenidos.
- Construcción de un corpus de mayor porte.
 - ¿mejoran los resultados al aumentar la cantidad de ejemplos?
 - tokens que preceden y suceden a adverbios utilizados en el clasificador CRF, ¿son extensibles a otros corpus?
- Agregación de otras técnicas para aumentar la información de contexto.
 - p.ej.: etiquetado gramatical, análisis morfosintáctico, etc.

Gracias por su atención



Tipos de errores

- *Positivos Verdaderos* (PV) son elementos de la clase buscada que fueron correctamente identificados.
- *Negativos Verdaderos* (NV) son elementos que no pertenecen a la clase buscada que fueron correctamente ignorados y clasificados en una clase diferente.
- *Falsos Positivos* (FP), o errores de Tipo I, son elementos que pertenecen a otra clase y que fueron incorrectamente clasificados en la clase buscada.
- *Falsos Negativos* (FN), o errores de Tipo II, son elementos que pertenecen a clase buscada y que fueron incorrectamente clasificados en otra clase.

$$P = \frac{PV}{PV + FP} \quad (1)$$

$$R = \frac{PV}{PV + FN} \quad (2)$$

$$F_{\alpha} = \frac{P \times R}{(1 - \alpha)P + (\alpha)R} \quad (3)$$

$$F_{0,5} = \frac{2 \times P \times R}{P + R} \quad (4)$$

- *Si de él careciéramos, ¿para qué/SIN_TILDE unas tareas que/SIN_TILDE requieren esfuerzo, dedicación, capacidad y que/SIN_TILDE —además— no mejoran ninguna economía?*
- *¿No se debate permanentemente —como/CON_TILDE toda religión y toda demencia— en el conflicto entre lo real y lo ficticio, lo percibido y lo proyectado, lo que/SIN_TILDE constriñe y lo que/SIN_TILDE exalta, los milagros y las bromas pesadas?*
- *-Que/CON_TILDE cómo/SIN_TILDE va a llamarse el chiquillo?*
- *"Un experimento probará la preparación y las capacidades para el contexto militar del futuro, y sirve para que/SIN_TILDE veamos cómo/SIN_TILDE cada una de las fuerzas se desempeñará en una guerra".*