ISSN 1870-4069

N-gramas de sílabas vs. n-gramas de caracteres para la tarea de atribución de autoría en un corpus multi-tema

Héctor Javier Hernández[1], Hiram Calvo[1], Eduardo López[2],
Juan Pablo Posadas-Durán[2],
Ilya Markov[3], Grigori Sidorov[1]

[1] Instituto Politécnico Nacional,
Centro de investigación en Computación,
México
[2] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Zacatenco,
México
[3] University of Amsterdam,
Holanda

hhernandezs1203@alumno.ipn.mx,
hcalvo@cic.ipn.mx,lopezmareseduardo@gmail.com,
jposadasd@ipn.mx, imarkov@uva.nl, sidorov@cic.ipn.mx

Resumen. Los caracteres de n-gramas como marcadores de estilo han demostrado un buen desempeño para la tarea de atribución de autoría [5]; sin embargo, es difícil dar una interpretación específica de la información que estas características están capturando. En este trabajo se plantea la comparación de n-gramas de caracteres y n-gramas de sílabas como marcadores de estilo para comprobar qué tanto pueden ayudar estas últimas a identificar correctamente la autoría. Partimos de la hipótesis de que los n-gramas de caracteres capturan cierta información morfológica de las palabras, y que por ello el uso de sílabas podría ser equivalente. Realizamos experimentos con diversas características, y encontramos que el desempeño de las sílabas está cercano al que presentan los n-gramas de caracteres, si bien los caracteres permiten realizar la tarea con un mejor desempeño.

Palabras clave: N-gramas de caracteres, n-gramas de sílabas, atribución de autoría.

N-grams of Syllables vs. N-grams of Characters for the Authorship Attribution Task in a Multi-topic Corpus

Abstract. N-gram characters as style markers have yielded a good performance for the task of authorship attribution [5]; however, it is difficult to give a specific interpretation of the information that these characteristics are capturing. In this work, the comparison of character n-grams and syllables n-grams are proposed as style markers to verify to which extent the latter help to

correctly identify authorship. We start from the hypothesis that character n-grams capture certain morphological information from words, and that, because of this, the use of syllables could be equivalent. We conducted experiments with various characteristics, and found that the performance of the syllables is to that of character n-grams for this task, although characters allowed to perform the authorship attribution task with better performance.

Keywords: Character n-grams, syllable n-grams, authorship attribution.

1. Introducción

La atribución de autoría es la línea de investigación que se dedica al problema de identificar el autor de un documento dentro de un conjunto de posibles autores. En esta área, una de las características que ha demostrado obtener mejores resultados han sido los n-gramas de caracteres [1].

A pesar de que los n-gramas de caracteres no corresponden a una característica propia de la lingüística, su eficacia para resolver el problema de atribución se puede explicar en el hecho de que ciertos n-gramas de caracteres corresponden a elementos como sílabas, prefijos o sufijos. Es así que el propósito de esta investigación es determinar si el buen desempeño de los n-gramas se debe a que capturan esta información morfológica. Para ello, realizaremos una comparación entre ambos tipos de características.

Otra estrategia usada frecuentemente en la solución de problemas relacionados con el Procesamiento de Lenguaje Natural es la denominada frecuencia de palabras que consiste en obtener la frecuencia de ocurrencia de cada palabra en un documento. La estrategia presenta la desventaja de que puede ser que las palabras no siempre estén escritas de la misma manera [1], por lo cual las inflexiones que corresponden a una palabra se consideran tan diferentes como dos palabras no relacionadas entre sí.

En este trabajo se hace una evaluación de los n-gramas y de las sílabas como marcadores de estilo en un problema de atribución de autoría usando un corpus en español que incluye diferentes tópicos; además se propone y evalúa un método basado en el uso de las siguientes características: n-gramas de caracteres filtrados y sílabas. En la sección 2 presentamos algunos trabajos relacionados con esta propuesta, así como la base de la clasificación de n-gramas de caracteres; en la sección 3 presentamos nuestra propuesta; en la sección 4 presentamos el corpus construido; en la sección 5 presentamos algunos experimentos; en la sección 6 nuestros resultados, y finalmente en la sección 7 nuestras conclusiones.

2. Trabajos relacionados

Los n-gramas de caracteres han demostrado ser exitosos en la tarea de atribución de autoría aunque, como se explica en [1], los n-gramas son probados bajo condiciones controladas, es decir, en donde los textos de un corpus pertenecen a un mismo tema o dominio; sin embargo, en un escenario más real, los textos pueden tener autores que escriban de uno o más temas.

Tabla 1. Categorías de sílabas.

Categoría	Subcategoría	Definición
Afijo	Prefijo	Primera sílaba de una palabra.
Afijo	Sufijo	Última sílaba de una palabra.
Palabra	Palabra completa	Palabras de una sola sílaba
Palabra	Palabra al centro	Las sílabas que se encuentren al centro de una palabra
Palabra	Multi palabra	Contiene la última sílaba de una palabra y la primera de la siguiente, siempre y cuando las dos palabras tengan más de una sílaba.
Puntuación	Puntuación inicial	Bigrama compuesto por: signo de puntuación + sílaba
Puntuación	Puntuación central	Trigrama compuesto por: sílaba + signo de puntuación + sílaba.
Puntuación	Puntuación final	Bigrama compuesto por: sílaba + signo de puntuación

En trabajos previos, los n-gramas de caracteres son tratados de la misma manera, sin importar la estructura o la posible información morfológica/sintáctica que estos n-gramas nos puedan proporcionar. No es hasta que en [2] se propone una clasificación de n-gramas de caracteres, y el resultado de esta clasificación muestra que hay n-gramas en específico que parecen aportar más información y por tanto mejorar el desempeño de la clasificación. En este trabajo usaremos la misma clasificación, con la diferencia de reemplazar los n-gramas de caracteres por n-gramas de sílabas.

3. Sílabas como marcadores de estilo

Se propone usar las sílabas como marcadores de estilo, además de esto se propone seguir la clasificación de n-gramas de caracteres propuesta en [2], pero aplicada a los n-gramas de sílabas.

3.1. Extracción de sílabas

La división silábica *per se* es un problema complejo en cualquier idioma, en este trabajo se limita a implementar soluciones existentes a este problema. Para poder obtener las sílabas de una palabra empleamos el siguiente proceso para obtener la división silábica de las palabras, como resultado de una búsqueda, obtuvimos un diccionario de división silábica del sitio web [4], el cual fue la primera fuente de consulta para obtener una división correcta, si la palabra no se encuentra en la

Tabla 2. Documentos que componen el corpus CCAP-s1.

Autor	Cine	Comida	Fotografía	Arte	Estilo de vida	Total
Alejandro Arroyo C.	52	16	16	9	34	149
Alejandro López	51	16	17	12	54	181
Andrea Méndez B.	26	5	21	44	33	147
Daniela Fernandez	6	4	9	5	14	51
Iván Montejo	14	7	4	8	12	58
Mafer Fernández L.	9	2	7	12	15	57
Total por categoría	**158**	**50**	**74**	**90**	**162**	**643**

colección, entonces se empleamos una librería para python llamada *pyphen* [3] la cual está bajo la licencia GPL 2.0+/LGPL 2.1+/MPL 1.1 tri-license.

4. Corpus

Para el desarrollo de este proyecto, se necesita usar un corpus que refleja el uso del lenguaje en español en escenarios reales. Una de las características importantes que este recurso debe cumplir es que debe de ser multi-tema; es decir, los diferentes autores deben de tener documentos de diferentes temas; por ejemplo, el autor uno debe tener documentos de las categorías "A","B", "C" y "D", de igual forma el resto de los autores que integran el corpus. Esta característica presento un reto complicado ya que en un primer momento se intentó obtener el corpus de periódicos nacionales y de revistas, sin embargo, surgió el problema de que la gran mayoría de los autores (periodistas) de estos medios, sólo se centraban en una categoría, por lo que estos medios de comunicación no pudieron ser utilizados. Como resultado de la continuación de la búsqueda, se determinó utilizar el sitio web llamado "Cultura Colectiva", el cual es una plataforma de comunicación de contenido original en español; y contiene publicaciones de un gran número de autores y variedad de temas como: cine, comida, arte, diseño, historia, entre otras. La importancia de un corpus multi-tema estriba en que, al realizar los experimentos en un corpus de este tipo, representa un escenario más realista, ya que, en situaciones reales, los factores como número de autores, número de documentos por autor, y el tema de cada documento puede variar, y el objetivo de este trabajo es demostrar que se puede resolver esta tarea a través de un análisis de información morfológica, en un entorno más complejo y desbalanceado. El corpus recopilado (CCAP) consta de 4,134 documentos. Estos documentos pertenecen a una colección de 11 categorías diferentes, y 21 autores diferentes, para este trabajo se tomó la decisión de acotar este corpus a solo 643 textos, de 6 autores, y 6 temas diferentes (CCAP-s1). El detalle de la distribución de dichos textos en CCAP-s1 puede apreciarse en la Tabla 2. Ambos corpus pueden encontrarse en http://idic.likufanele.com/~ccap.

5. Experimentos

En este trabajo se propone un escenario más real al entrenar SVM usando un conjunto de textos pertenecientes a una categoría y haciendo pruebas con textos de otra categoría, en este trabajo se han probado diferentes combinaciones de características, pero solo nos enfocaremos en las que presentaron mejores resultados.

5.1. Desambiguación de los n-gramas de caracteres y de sílabas

Dentro de un conjunto de n-gramas de caracteres existe la posibilidad de que un n-grama pertenezca a más de una categoría [2], por ejemplo el bigrama *de*, se puede considerar como un n-grama que bien podría estar dentro de la categoría de *n-grama como palabra completa*, sin embargo, este bigrama también lo podemos encontrar al inicio, al centro o al final de una palabra (*destello*, *clandestino*, *jade*), es por esto que en este trabajo se emplea el concepto de n-gramas *etiquetados* y *sin etiquetas*; con esto pretendemos desambiguar el sentido de un n-grama, y pretendemos identificar los n-gramas que puedan brindar información más relevante. Así mismo aplicamos esta misma idea para el uso de sílabas como marcadores de estilos, una silba podría pertenecer a más de una categoría.

5.2. Evaluación de los métodos propuestos

Para la evaluación del método propuesto, utilizamos la precisión de la predicción que el modelo hace; Los experimentos se realizaron entrenando máquinas de vectores soporte usando WEKA con su configuración por defecto.

Para todos los experimentos usamos n-gramas de caracteres con n=3, la razón de esta decisión está basada en los resultados de que se reportan en trabajos previos [2].

5.3. Configuración para el corpus cultura colectiva

La configuración para los experimentos con el corpus de cultura colectiva consistió en tomar los documentos de una categoría y hacer pruebas con otra, cabe aclarar que los conjuntos de entrenamiento y de pruebas no están estrechamente correlacionados, ya que el lenguaje como el contexto será diferente. Al tratarse de 6 diferentes categorías, se realizaron 30 combinaciones diferentes, de entrenamiento-prueba, de estas combinaciones se presentan la que dio un mejor resultado.

6. Resultados

El corpus en español que representa un escenario mucho más interesante y complejo, el esquema que se siguió, en esencia es el mismo, se tiene que dividir el corpus en una sección para el entrenamiento y otra para las pruebas, más sin embargo, lo interesante de estos experimentos estriba en que, los conjuntos de entrenamiento y pruebas no estarán estrechamente relacionados, es decir, el uso del lenguaje en dichos conjuntos será distinto, ya que ambos pertenecen a diferentes temas, lo interesante es poder observar, aquellas características que prevalecen y otorgan información sobre el estilo

Tabla 3. N-gramas vs. sílabas (P=Porcentaje de precisión).

Tipo de Característica	Características sin etiquetar		Características etiquetados		Combinación de afijos + puntuación	
	P	característica s	P	características	P	características
n-gramas	56.0	4,913.7	56.3	5734.7	56.5	30,047
sílabas	54.2	5,801.7	54.8	6419.8	52.9	1821.3

que una persona tiene al escribir, sin importar el tema del que esté hablando. Para estos experimentos al tratarse de 6 categorías se realizaron las 30 posibles combinaciones de entrenamiento-prueba, es decir, arte vs. cine, arte vs. comida, arte vs. diseño, arte vs. estilo, arte vs. fotografía, y viceversa, para finalmente obtener el promedio de éstas. Para los experimentos con n-gramas de caracteres empleamos el siguiente set de características: (*Prefijos, Sufijos, Espacio + prefijo, Espacio + sufijo*) y para los experimentos con sílabas empleamos el siguiente set de características (*Prefijos, Sufijos, Multi palabra, Toda la palabra, Palabra al centro*). A continuación, se muestran los resultados obtenidos para este corpus.

7. Conclusiones y trabajo a futuro

Los n-gramas basados en caracteres han demostrado tener un buen desempeño para la tarea de atribución de autoría, y tal parece que parte de este éxito se debe a que estos n-gramas están capturando información morfológica de las palabras lo cual se ha podido mostrar parcialmente mediante los experimentos presentados en este trabajo, sin embargo, este trabajo presenta varias áreas de oportunidad, como mejorar la manera en la que se obtienen las sílabas, ya que no podemos obtener una división silábica que sea cien por ciento precisa, y al no tener la división, perdemos información.

Como trabajo futuro se plantean las siguientes acciones:

- Robustecimiento de recursos léxicos para el idioma español,
- Plantear un algoritmo o metodología más eficiente para la división silábica,
- Probar diferentes algoritmos de aprendizaje automático para resolver esta tarea.

Referencias

1. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy, 21, pp. 421–439 (2013)
2. Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not all character N-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–102 (2015)
3. Kozea, A.: Community project. Pyphen de A Kozea Community Project Sitio https://pyphen.org/ (2008)

4. OSLIN: Diccionario de división silábica. OSLIN https://web.archive.org/web/20160203184525/es.oslin.org/syllables.php. (2015)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. JASIST, 60, pp. 9–26 (2009)