ISSN 1870-4069

Uso de *n*-gramas sintácticos en la atribución de traductor

Christian Caballero, Hiram Calvo, Ildar Batyrshin

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{cdch10, batyr1}@gmail.com, hcalvo@cic.ipn.mx

Resumen. La atribución de traductor es la tarea de atribuir un texto traducido a su traductor. La tarea es similar a la más conocida tarea de atribución de autoría en procesamiento de lenguaje natural, pero, a diferencia de esta tarea, la atribución de traductor puede ser más difícil ya que el tema está determinado ya por el autor del texto original, lo que restringe el vocabulario a usar por el traductor. En este trabajo se presentan los resultados de representar los textos usando *n*-gramas sintácticos para atribuir los textos traducidos a su traductor usando algoritmos de aprendizaje automático.

Palabras clave: Atribución de traductor, *n*-gramas sintácticos, aprendizaje automático, estilometría.

Use of Syntactic *n*-grams in Translator Attribution

Abstract. Translator attribution is the task of attributing a text to its translator. This task is related to the similar and well-known task in natural language processing of Authorship Attribution. However, unlike the latter, translator attribution might be a more difficult task since the topic is predetermined by the original author, which limits the vocabulary to be used by the translator. This work presents the results of representing the texts using syntactic *n*-grams to attribute translated texts to their translator using machine learning algorithms.

Keywords: Translator attribution, syntactic *n*-grams, machine learning, stylometry.

1. Introducción

Atribuir un fragmento de texto a un autor (es decir, atribución de autoría) es una tarea bien establecida en el procesamiento del lenguaje natural [8,9,10,17]. Sin embargo, atribuir un documento traducido a su *traductor* (se podría llamar

a esto atribución de traductoría) es una tarea poco estudiada. Hay tan solo unos pocos artículos abordando este problema y solo un pequeño subconjunto de ellos reportan resultados exitosos.

En [1], Mona Baker propuso hace dos décadas un esquema del marco de trabajo para estudiar el estilo de un traductor o de un grupo de traductores. Primeramente, cabe mencionar que la palabra *estilo* es una palabra "cargada" en el sentido de que su significado varía con respecto a quién la usa y en qué contexto [3]. Para Baker en ese estudio (y en el presente trabajo), el *estilo* "es una cuestión de patrones: involucra el describir patrones preferidos o recurrentes de comportamiento lingüístico en lugar de instancias individuales o extraordinarias de intervención" [1]. Esta definición incluiría, por supuesto, elecciones tanto conscientes como inconscientes por parte del escritor.

Mona Baker concluye su artículo sugiriendo que, en lugar de analizar varias traducciones del mismo traductor (como ella hizo para dos traductores), tal vez se debería comparar el mismo texto fuente traducido por diferentes traductores con la finalidad de encontrar el estilo del traductor, de esa forma se controla por las características del trabajo original. Este enfoque es usado en [5] y [11] y es el seguido en este trabajo para poder comparar los resultados.

Baker también menciona en [1] que entre las opciones para buscar el estilo individual del traductor están los patrones sintácticos, dispositivos de cohesión e incluso la puntuación. Sin embargo, ella solo utiliza características más tradicionales (en el campo de la atribución de autoría) como son razón de tipo a *token* y longitud promedio de oración, y frecuencia de uso de la palabra *say*, que es el verbo para reportar más común en inglés.

A continuación se menciona el trabajo que se ha realizado en este problema. Después, se da detalle sobre el corpus usado para realizar los experimentos y los métodos empleados. En seguida, se presentan los resultados, se discuten y presentan las conclusiones y la bibliografía referida en este trabajo.

2. Trabajo relacionado

2.1. Trabajos con textos traducidos

Como se mencionó en la introducción, Mona Baker llamó la atención al problema de identificar el estilo personal de un traductor. Su equipo de investigación produjo un corpus de textos traducidos, llamado el Translational English Corpus (TEC), que contiene alrededor de diez millones de palabras de texto traducido al inglés de distintas lenguas fuente. De este corpus, ella tomó textos de dos traductores para su artículo [1]: seis textos (cinco del español y uno del portugués) de un traductor y tres del árabe del segundo traductor.

Con esos textos, comparó la variación de la proporción de tipo a *token* con un promedio móvil de 1 000 palabras, longitud promedio de oración y estructuras de diálogo indirecto (realizaciones del verbo *say*). Ella encontró diferencias entre los traductores y consistencias a lo largo de sus producciones individuales.

La conclusión que presenta en ese artículo apunta a usar corpus paralelos de traducciones para mantener fijas las variables del autor y lengua originales.

Alrededor de esa época, Mikhailov y Villika publicaron un artículo [14] en el que buscaban responder si los traductores tienen una huella estilística usando algunos de los métodos más usados en la atribución de autoría, tales como riqueza léxica, palabras más frecuentes y palabras favoritas en un corpus de textos rusos traducidos al finés. Desafortunadamente, sus resultados mostraron que los textos eran más similares a los originales que entre traducciones del mismo traductor. Es decir, traducciones de distintas obras del mismo autor son más similares entre sí que traducciones de diferentes textos por el mismo traductor. Sin embargo, Mikhailov y Villika mencionan que aún piensan que el estilo del traductor se puede encontrar pero usando otras características como palabras modales, conjunciones o formas gramaticales.

Hay más trabajos con traducciones en la última década. Hedegaard y Simonsen [7] consideraron la influencia del traductor como "ruido" en el problema de hacer la atribución de autoría en textos traducidos. Ellos utilizaron semántica de marcos, palabras más frecuentes y n-gramas de caracteres con una máquina de vectores de soporte. Concluyen que si se quisiera hacer la atribución al traductor se deberían usar los n-gramas de caracteres y las palabras más frecuentes ya que los traductores dejan una huella rastreable.

Otros trabajos, como [15,6] utilizan técnicas de agrupamiento en textos traducidos y observan que los textos se agrupan por los autores originales en lugar de hacerlo por su traductor. Ambos trabajos utilizan las palabras más frecuentes y métricas comunes en atribución de autoría como la Delta de Burrows [2].

2.2. Trabajos en estilometría del traductor

Más recientemente, en 2018 y 2019, se publicaron un par de trabajos que lidian con el problema tal como lo planteó Baker en el 2000, usando corpus de traducciones paralelas. Ambos trabajos usan técnicas de aprendizaje automático, pero difieren en las características que utilizan para representar los documentos.

Primero en [11], Lynch y Vogel abordan este problema usando n-gramas de palabras y etiquetas de categoría gramatical (*POS tags*) y también características más clásicas como longitud promedio de oración o proporción de verbos a palabras totales. Utilizan cuatro algoritmos de clasificación (máquina de vectores de soporte, Bayes ingenuo, árbol de decisión y regresión logística) en distintas combinaciones de experimentos ya que cuentan con dos corpus. Sus resultados son buenos con bigramas de palabras y también con longitud promedio de oración, aunque no estaban tan interesados en la clasificación por sí misma sino en qué características de cada traductor eran las que ayudaban más a diferenciar entre traductores.

El segundo trabajo es de El-Fiqi *et al.* [5]. Ellos usan las mismas características que Mikhailov y Villika [14], además de usar motivos de redes para entrenar una máquina de vectores de soporte y un árbol de decisión. Los motivos de redes son subgrafos recurrentes en un grafo de adyacencia de las palabras de un documento. Ellos utilizan dos corpus de traducciones paralelas y encuentran buenos resultados después de hacer selección de características y normalización a las características de los motivos de redes.

3. Métodos

En esta sección se detalla el corpus en el que se realizaron los experimentos así como los pasos para realizarlos (ver figura 1). Se desarrollaron programas en el lenguaje de programación Python[1] para los experimentos. Los programas se apoyaron en el uso de bibliotecas especializadas para manejo de lenguaje natural (spaCy[2]), algoritmos de aprendizaje automático (scikit-learn[3]) y análisis de datos (pandas[4]) para el preprocesamiento, experimentación, evaluación y reporte de resultados.

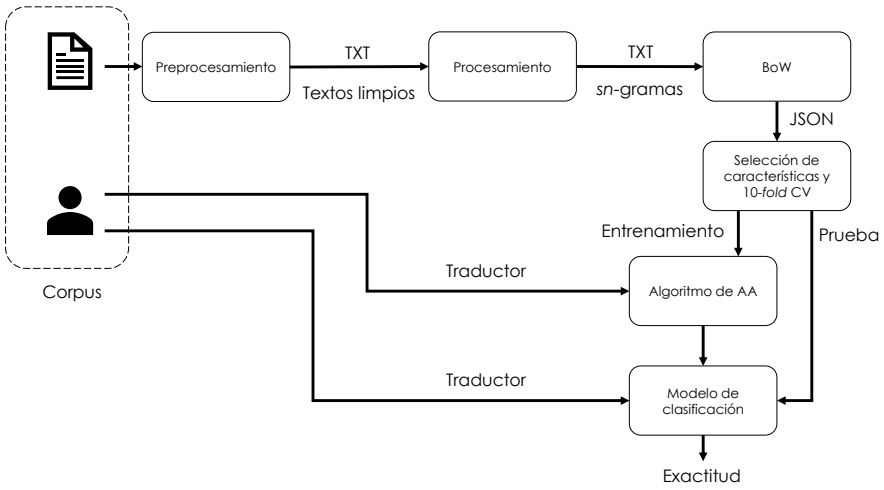

Fig. 1. Proceso para obtener la exactitud del modelo de clasificación usando validación cruzada.

3.1. Corpus

Los autores de [5] utilizaron dos corpus para sus experimentos y los hicieron disponibles en Internet. Se tomó la decisión de usar su corpus más extenso: el de *Don Quijote*. Este corpus está compuesto por tres traducciones al inglés de las dos partes de la novela de *Don Quijote*, 52 capítulos de la primera parte y 74 de la segunda. La tabla 1 muestra los detalles de los nombres de los traductores y el número de palabras mínimo y máximo por capítulo para cada traductor.

[1] https://www.python.org
[2] https://spacy.io
[3] https://scikit-learn.org
[4] https://pandas.pydata.org

Table 1. Número de palabras por capítulo del corpus de *Don Quijote*.

Traductor	Parte I	Parte II
Charles Jarvis	1 700 - 8 202	797 - 5 530
John Ormsby	1 648 - 8 304	823 - 5 680
Thomas Shelton	1 820 - 8 921	759 - 5 158

3.2. Preprocesamiento

En la etapa de preprocesamiento, se reemplazaron los caracteres especiales por sus contrapartes (por ejemplo, *ë* por *e*); se removieron los números que se encontraban entre corchetes, usados para indicar notas al pie, y se colapsaron dos o más espacios en blanco (incluyendo saltos de línea) a un solo espacio simple.

3.3. Procesamiento

El corpus provisto por [5] cuenta con los textos ya separados por capítulo y por traductor. Así que después de preprocesarlos, se analizaron usando la biblioteca especializada de procesamiento de lenguaje natural spaCy. Esta biblioteca cuenta con modelos de lenguaje basados en redes neuronales convolucionales de distintos tamaños para una variedad de lenguas que proporcionan la capacidad de hacer tokenizado, etiquetado de categoría gramatical (*POS-tagging* en inglés), análisis de dependencias sintácticas y etiquetado de entidades nombradas.

Ya analizados los textos, se generaron archivos de dependencias usando el formato de analizador de Stanford enmascarando todos los nombres propios con su etiqueta de categoría gramatical "PROPN", pasando todo a minúsculas y omitiendo las relaciones de dependencias de los signos de puntuación, para después usar un programa provisto por [16] para extraer n-gramas sintácticos. Los n-gramas sintácticos de palabras son n palabras adyacentes en el árbol sintáctico de las oraciones del texto.

Después de tener los n-gramas sintácticos con $n \in \{2, 3\}$ para cada oración, se contabilizaron las ocurrencias de cada bigrama y trigrama sintáctico para cada documento para tener dos representaciones de bolsa de palabras de los documentos. Esta información se guardó a disco a archivos JSON[5] para su fácil conversión a matrices dispersas usando scikit-learn[6].

3.4. Entrenamiento de los clasificadores

Se seleccionaron los cuatro algoritmos de clasificación usados en los trabajos del estado del arte [5,11]: regresión logística, máquina de vectores de soporte, Bayes ingenuo y árbol de decisión; así como la misma metodología de evaluación cruzada con diez iteraciones (*10-fold cross-validation*).

[5] https://www.json.org
[6] https://scikit-learn.org

Esta metodología de evaluación divide el conjunto de datos en diez subconjuntos y utiliza nueve partes para el entrenamiento de los clasificadores y la restante para su validación [13]. Este proceso se repite diez veces de tal forma que, al final, se habrán usado todos los datos tanto para entrenamiento como para validación. El resultado de la validación es el promedio aritmético (media) de alguna medida de evaluación. La métrica de evaluación seleccionada fue la exactitud, definida como el número de instancias correctamente clasificadas entre el total de las instancias [4], ya que el número de instancias entre clases está balanceado.

Para reducir el número de dimensiones (número de n-gramas sintácticos distintos) con los que se representan los documentos, se seleccionaron las 45 características más relevantes para la clasificación usando el estadístico χ^2, un método común de selección de características en el campo de la recuperación de información [12].

Se utilizaron los valores por defecto de los hiperparámetros de los cuatro clasificadores. Para la máquina de vectores de soporte se utilizó un *kernel* lineal y un término de regularización en ℓ_2 y una estrategia *uno contra todos* para k clasificadores en un problema de clasificación con k clases. Para el clasificador de Bayes ingenuo multinomial se utilizó un término de suavizado de Laplace $\alpha = 1$. El árbol de decisión utiliza la función de impureza "Gini". Por último, el clasificador de regresión logística tiene un término de regularización en ℓ_2 y una función de pérdida multinomial cuando se le pasa un problema multiclase.

4. Resultados experimentales

Los resultados de los experimentos se muestran en el cuadro 2. Los cuatro clasificadores son máquina de vectores de soporte (SVC), Bayes ingenuo multinomial (NB), árbol de decisión (DT) y regresión logística (LR). Los valores de exactitud para cada conjunto de características están en negritas.

Table 2. Resultados de la clasificación

Características	SVC	NB	DT	LR
bigramas sintácticos	96,29 %	95,77 %	85,72 %	**97,35** %
trigramas sintácticos	88,89 %	**91,01** %	76,17 %	89,68 %

En [5], los autores, que utilizaron este mismo corpus, reportan un valor de exactitud promedio de 95,10 % usando motivos de red de tamaño tres como características. Obtienen este valor después de haber hecho selección de características y empleando un árbol de decisión como algoritmo de clasificación.

5. Discusión de resultados

Los autores de [5] argumentan que esas características de motivos de redes que utilizan logran capturar información sintáctica ya que las extraen de redes de adyacencia de lemas de los textos. Nosotros usamos verdadera información sintáctica ya que los n-gramas sintácticos se obtienen a partir de dependencias sintácticas de las oraciones. Por lo tanto, nuestros resultados son comparables con los suyos.

Una ventaja de su método sobre el nuestro es que no necesitan un modelo de lenguaje para obtener las relaciones sintácticas. Sin embargo, si este tipo de método se aplica a una lengua que cuente con un modelo de lenguaje (típicamente, una lengua natural con grandes recursos) no hay mayor complicación en realizar el análisis sintáctico. La biblioteca empleada en este trabajo es bastante rápida: analiza el corpus entero–más de 400 000 palabras por traductor– en menos de 15 minutos en una computadora personal y cuenta con modelos para varias lenguas (inglés, alemán, francés, español, portugués, italiano, holandés, griego, noruego, lituano).

Por otro lado, como ya se mencionó, el usar n-gramas sintácticos captura *verdadera* información sintáctica. Además, se pueden usar variedades de estos n-gramas: con pares palabra-categoría gramatical o pura categoría gramatical, por ejemplo. El análisis de los textos solo se realiza una vez y toda la información del modelo se mantiene disponible.

6. Conclusiones y trabajo a futuro

En este trabajo se presentó el poco estudiado problema de atribución de traductor, que consiste en hacer la atribución de un texto traducido a su traductor en lugar de a su autor original. El problema está relacionado con el más conocido problema en procesamiento de lenguaje natural de atribución de autoría.

La metodología seguida es la propuesta en [1] por la profesora de Estudios de Traducción Mona Baker en el año 2000 de usar corpus paralelos de traducciones (traducciones de un mismo texto por varios traductores a la misma lengua meta).

En [5], El-Fiqi *et al.* usan esta misma metodología y prueban con técnicas de aprendizaje automático en un par de corpus de traducciones paralelas utilizando motivos de redes. Los motivos de redes son patrones recurrentes que aparecen en redes (o grafos) de adyacencia de palabras. Por la naturaleza de las características que emplearon, argumentan que sus características recogen información sintáctica.

En el presente trabajo se presentaron los resultados de usar verdadera información sintáctica por medio de n-gramas sintácticos y se probó en uno de los corpus usado en [5]. Los resultados que obtenemos con bigramas sintácticos usando selección de características con el estadístico χ^2 es mejor que el obtenido en el trabajo citado con el valor añadido de usar información sintáctica.

Como trabajo adicional y líneas a explorar en un futuro quedan probar con otro tipo de características en otros corpus paralelos y semiparalelos (traducciones de diferentes textos por los mismos traductores). Así como realizar

experimentos con otros pares de lenguas y textos de otros géneros para evaluar la sensibilidad de los métodos al género o incluso a la tipología textual.

Otra línea adicional sería probar estas técnicas para discernir entre traductores humanos y traductores automáticos o incluso entre distintos traductores automáticos.

References

1. Baker, M.: Towards a methodology for investigating the style of a literary translator. Target. International Journal of Translation Studies 12(2), 241–266 (2000)
2. Burrows, J.: 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing 17(3), 267–287 (2002)
3. Crystal, D.: Style: the varieties of english. A history of literature in the English language 10 (1987)
4. Eisenstein, J.: Introduction to Natural Language Processing. The MIT Press (2019)
5. El-Fiqi, H., Petraki, E., Abbass, H.A.: Network motifs for translator stylometry identification. PLOS ONE 14(2) (2019)
6. Forsyth, R.S., Lam, P.W.: Found in translation: To what extent is authorial discriminability preserved by translators? Literary and Linguistic Computing 29(2), 199–217 (2014)
7. Hedegaard, S., Simonsen, J.G.: Lost in translation: authorship attribution using frame semantics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 65–70. Association for Computational Linguistics (2011), https://www.aclweb.org/anthology/P11-2012
8. Holmes, D.I.: Authorship attribution. Computers and the Humanities 28(2), 87–106 (1994)
9. Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3), 233–334 (2008)
10. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for information Science and Technology 60(1), 9–26 (2009)
11. Lynch, G., Vogel, C.: The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. Computer Speech & Language 52, 79–104 (2018)
12. Manning, C.D., Raghavan, P., Schütze, H.: An introduction to information retrieval. Cambridge University Press (2008)
13. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. Massachusetts Institute of Technology (1999)
14. Mikhailov, M., Villikka, M.: Is there such a thing as a translator's style? In: Proceedings of the Corpus Linguistics. pp. 378–385. Lancaster, UK (2001)
15. Rybicki, J.: The great mystery of the (almost) invisible translator. Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research 231 (2012)
16. Sidorov, G.: Syntactic n-grams in computational linguistics. Springer (2019)
17. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)