Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas

Jovany Marcos Ramírez, Maya Carillo Ruíz, María Josefa Somodevilla

Benemérita Universidad Autónoma de Puebla
jo.va.ny@hotmail.com
{crrllrzmy, mariajsomodevilla}@gmail.com

Resumen: En este artículo se propone la utilización de la representación holográfica reducida (HRR) en la tarea de Atribución de Autoría (AA). Dicha representación permite combinar información léxica y sintáctica de los textos. En vectores de dimensión manejable. Para contar con vectores de dimensión apropiada se aplica la metodología de Indexación Aleatoria (RI). Los experimentos realizados muestran que la HRR genera resultados equiparables a los reportados en la bibliografía.

Palabras clave: Atribución de autoría, Representaciones holográficas reducidas, Indexación aleatoria.

1. Introducción

La Atribución de Autoría (AA) es una tarea que busca caracterizar el estilo de escritura de autores, con el fin de asignar de forma automática textos de autoría desconocida al autor correspondiente Es decir busca identificar características textuales que al compararlas, permitan discriminar entre documentos escritos por distintos autores [1].

Los métodos generales de AA extraen marcadores de estilo, que se consideran atributos de los textos y se utiliza para entrenar clasificadores [2]. Estas marcas de estilo, incluyen: frecuencia de caracteres, palabras, frase, n-gramas a nivel carácter, combinaciones de palabras o n-gramas de palabras, por mencionar algunos. Es importante señalar que la AA no debe ser abordada de forma temática, ya que las características textuales más importantes no son de tipo temático, pues el objetivo es modelar el tipo de escritura de cada autor con el fin de discriminarlos, incluso en el mismo contexto.

Hoy en día, la cantidad de información disponible es abrumadora y gran parte de ella está en texto plano (e-mails, blogs, foros en línea). En este contexto, han surgido diversos temas que involucran a la AA, por ejemplo: ciber-bullying, detección de plagio, correo no deseado, informática forense, detección de fraude, autenticidad de documentos [3].

En el presente trabajo se propone un método para combinar características léxicas y sintácticas empleando una representación novedosa conocida como representación holográfica reducidas (HRR). La HRR fue propuesta por Plate [6] como mecanismo para representar estructuras complejas y jerárquicas, que no se limitan al lenguaje, pues este tipo de estructuras se encuentran en otras áreas como el análisis de imágenes.

Por otra parte cuando los documentos se representan utilizando la aproximación de bolsa de palabras, sabemos que la dimensión vectorial es igual al tamaño del vocabulario de la colección. Para optimizar el procesamiento existen métodos que buscan reducir dicha dimensión, uno de los más utilizados en recuperación de información es la indexación semántica latente. Sin embargo éste método hace uso de la descomposición en valores singulares que es un proceso computacionalmente caro. Como alternativa Salgren [9] proponen un método conocido como indexación aleatoria (Random Indexing RI por sus siglas en inglés). En el presente trabajo se utiliza la RI para reducir la dimensión vectorial y optimizar el tiempo de procesamiento. La aportación principal de la presente investigación es la utilización de la HRR a la tarea de AA, donde no ha sido utilizada, de acuerdo a la información que se tiene hasta el momento.

Este artículo está organizado de la siguiente manera. En la sección 2 se presentan algunos trabajos relacionados, la sección 3 explica lo que es la representación holográfica reducida. La sección 4 introduce la metodología de indexación aleatoria (RI).La metodología propuesta se describe en la sección 5. Los corpus utilizados así como algunas características de los mismos y el tratamiento previo que se dio a los documentos obtenidos se presentan en la sección 6, en la sección 7 se describen los experimentos realizados, así como los resultados obtenidos. En la sección 8 están las conclusiones y trabajo futuro.

2. Trabajo relacionado

Existen dos formas para realizar AA, 1) el enfoque basado en el perfil del autor, en este se concatenan todos los documentos de un autor presentes en el conjunto de entrenamiento, para crear su perfil. Esto se realiza extrayendo varias características principalmente de bajo nivel, tales como n-gramas de caracteres. Para predecir la autoría de un documento nuevo, se debe calcular la similitud entre los perfiles de autor generados y las características del nuevo documento. Posteriormente el documento de autoría desconocida, se asignará al autor, cuyo perfil tuvo la mayor similitud con él [3], su estructura típica es mostrada en la Figura 1.

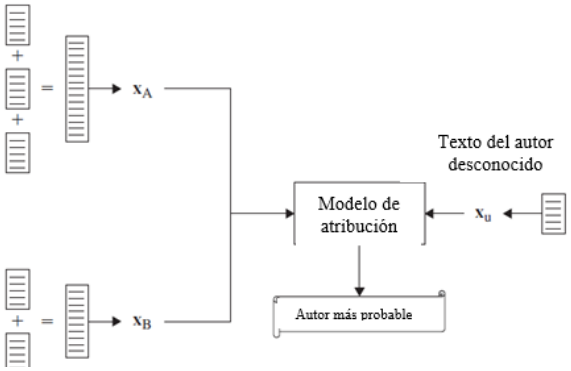

Fig. 1.- Enfoque basado en perfil del autor

2) En contraste el enfoque basado en máquina de aprendizaje, utilizan una representación vectorial, donde cada documento se representa de forma individual por un conjunto de características. Dichos vectores serán utilizados para entrenar un algoritmo de aprendizaje automático. Estos vectores suelen contener características variadas, desde caracteres, longitud de palabras, n-gramas de caracteres, n-gramas de palabras, y partes de la oración (POS). [3], su estructura típica podremos observarla en la Figura 2.

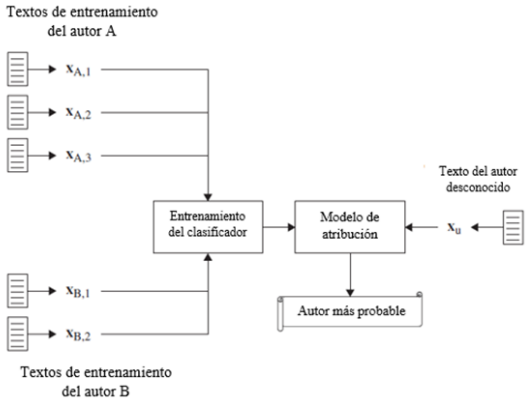

Fig 2. Enfoque basado en instancias

Existen diversas investigaciones en las cuales se han buscado métodos eficientes para la AA, entre ellos podemos mencionar: *Authorship Attribution Using Probabilistic Context-Free Grammars* de Raghavan et al. [2]. Raghavan plantea construir una gramática libre de contexto probabilística para cada autor y el uso de esta gramática como un modelo de lenguaje para la clasificación. Grigori Sidorov et al [4] propone en su investigación el uso de n-gramas, pero no de la manera tradicional, si no obteniendo los n-gramas en función del orden como se presentan en los árboles sintácti-

cos, es decir, seguir el camino del árbol sintáctico para crear los n-gramas, dando a estos el nombre de n-gramas sintácticos (sn-gramas) [4], entre otros.

3. Representación holográfica reducida

Las representaciones holográficas reducidas (HRR´s), son vectores cuyos elementos siguen una distribución normal, con media = 0 (μ=0) y desviación estándar = 1 (σ = 1). Hace uso dl operador de convolución circular ($\otimes$), para combinar la información léxica $x=(x_o, x_1,\ldots,x_{n-1})$ y sintáctica $y=(y_o,y_1,\ldots,y_{n-1})$ en $z = (z_0,z_1,\ldots,z_{n-1})$, . Así z se define como $z = x \otimes y$ [5]

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \qquad i = 0 \text{ a } n\text{-}1 \text{ (subíndices son módulo } n) \textbf{ (1)}$$

4. Indexación aleatoria

Con el uso de la indexación aleatoria (RI) se pretende reducir el espacio vectorial con el cual se trabajará, es decir, cada contexto (documento o palabra) se representa como un vector de tamaño fijo conocido como vector índice (VI). Las entradas de estos vectores serán ceros, con un pequeño número de elementos diferentes de cero, que serán 1's y -1's, en igual proporción. Por ejemplo si los vectores tienen veinte elementos distintos de ceros en un espacio vectorial de 1024, éste tendrá diez 1´s y diez -1´s, estos vectores servirán como etiquetas para las palabras o documentos [7].

La indexación aleatoria se lleva a cabo de acuerdo a los siguientes pasos:

1.- En primer lugar cada contexto (por ejemplo, cada documento o cada palabra) se le asigna una representación única y generada aleatoriamente, se le llama vector índice. Estos vectores de índice sin escasos, de alta dimensión y ternarios, lo que significa que su dimensionalidad (d) es del orden de miles, y que consiste en un pequeño número de 1´s y -1´s distribuidos al azar, el resto de los elementos se establecen en 0's.

2.- A continuación, los vectores de contexto se producen mediante el escaneo a través del texto, y cada vez que una palabra se produce en el contexto (por ejemplo, en un documento, o dentro de una ventana de contexto por deslizamiento), se añade el vector de contexto para la palabra en cuestión. Las palabras son así representadas por vectores de dimensión d que son efectivamente la suma de los contextos de las palabras.

5. Metodología propuesta

La metodología seguida se describe a continuación:

1. Primeramente se preprocesaron los documentos del conjunto de entrenamiento y pruebas, eliminando símbolos no alfanuméricos, así como los valores numéricos.

2. Una vez preprocesados los documentos, con ayuda del etiquetador de partes de la oración de Stanford [8], se procedió a etiquetar cada uno de los documento del conjunto de entrenamiento., Etiquetados todos los documentos se identificaron las etiquetas sintácticas únicas contenidas en estos documentos. El número total de etiquetas para los corpus utilizados se presentan en la Tabla 1.

3. Se utilizó RI para reducir el espacio vectorial, representando todo el vocabulario como vectores de ceros, unos y menos unos. Las etiquetas sintácticas se representaron como HRR y se asociaron mediante a convolución circular a los vectores generados por RI. La dimensión del espacio vectorial para todos los experimentos fue de 2048.

4. Se crearon las representaciones para la aproximación basada en instancia y en perfil

5. Obtenidos los vectores tanto para el conjunto de entrenamiento como el de prueba, para el enfoque basado en instancia se experimentó con los clasificadores J48, Naive Bayes, SVM. Para los enfoques basados en perfil del autor se hizo uso de la distancia euclidiana.

A continuación se ejemplifican los pasos para representar a los documentos. La salida del etiquetador Stanford, está en la Figura 3, en la cual se muestra una fracción de uno de los documentos etiquetados.

```
the_DT one_CD highly_RB visible_JJ success_NN of_IN the_DT stimulus_NN
program_NN has_VBZ been_VBN the_DT cash_NN for_IN clunkers_NNS
```

Fig 3. Etiquetado del texto

En la Figura 4 se ilustra la representación de los documentos. Las etiquetas sintácticas, se representaron como HRRs y las palabras como VI. Estos se combinaron con

la convolución circular. Finalmente los documentos se representaron como la suma de sus palabras representadas como HRRs.

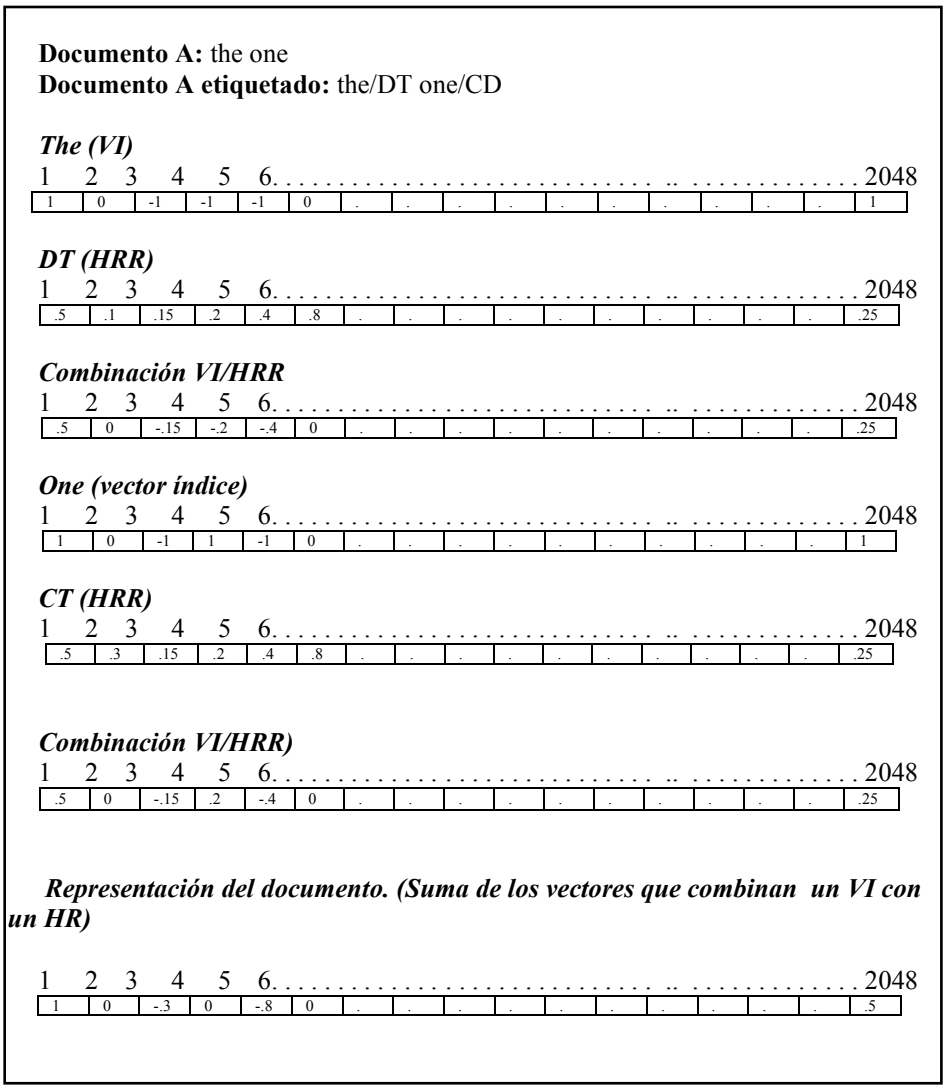

Fig 4. Representación de un documento como HRR

6. Conjunto de datos utilizados

Para los experimentos reportados a continuación, se utilizó un conjunto de 3 colecciones (Poetry, Business, NFL), con un total de 15 autores (6 para Poetry, 6 para Business y 3 para NFL), cabe destacar y hacer énfasis en que el conjunto de documentos de cada autor es pequeño. La Tabla 1, muestra información relevante de dichos corpus.

Tabla 1: Estadísticas sobre los corpus utilizados en los experimentos.

	Vocabulario	Etiquetas Sintácticas	Documentos Train	Documentos Test	Autores
Poetas	6940	34	146	55	6
Negocios	8492	34	85	90	6
NFL	4982	34	48	45	3

7. Experimentos y resultados

Se experimentó con los dos tipos de enfoques mencionados anteriormente: basados en instancias y los basados en perfil. Cabe destacar que para estos experimentos las colecciones fueron normalizadas por número de documentos y por el tamaño del vocabulario.

7.1. Resultados

A continuación se reportan los resultados obtenidos. La aproximación basada en instancias como podrá observase generó resultados muy pobres por lo que se descartó y únicamente se experimentó con la aproximación de perfil de autor.

7.1.1. Resultados del enfoque basado en instancias.

Como se aprecia en la Tabla 2, los resultados obtenidos en el enfoque basado en instancias son en su mayoría bajos y en ocasiones no se obtiene ningún resultado favorable, es decir, no se logra identificar o predecir de forma correcta los documentos que pertenecen al autor en específico, como es el caso del Autor 3 y el autor 6.

Tabla 2: Corpus Poetry, utilizando: J48, Naive Bayes, SVM

	Precisión	Recuerdo	Medida F
Autor 1	0.40	**0.90**	0.56
Autor 2	0.28	0.40	0.33
Autor 3	0.00	0.00	0.00
Autor 4	**0.66**	0.40	0.50
Autor 5	0.54	0.60	**0.57**
Autor 6	0.00	0.00	0.00

7.1.2. Resultados del enfoque basado en perfil del autor.

Los resultados obtenidos del enfoque basado en perfil del autor, fueron más favorables como podrá observarse.

Para el corpus Poetry los resultados se muestran en la Tabla 3.

Tabla 3: Métricas obtenidas para el corpus Poetry

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.70	0.70	0.70	**0.89**
Autor 2	0.50	0.40	0.44	**0.81**
Autor 3	0.55	0.50	0.52	**0.83**
Autor 4	0.36	0.40	0.38	**0.76**
Autor 5	0.45	0.50	0.47	**0.80**
Autor 6	0.33	0.40	0.36	**0.87**

En la Tabla 4 se muestran los resultados para el corpus NFL.

Tabla 4. Métricas obtenidas para el corpus NFL.

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.88	0.90	0.93	**0.95**
Autor 2	0.66	**0.80**	0.72	**0.80**
Autor 3	**0.90**	0.60	0.72	0.84

La tabla 5 muestra los resultados obtenidos del corpus de Business, con mejores resultados que los anteriores. Esto probablemente a que en este corpus los documentos de entrenamiento y pruebas están balanceados, es decir, la misma cantidad de documentos para los autores.

Tabla 5. Métricas para el corpus Business

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.86	0.86	0.86	**0.95**
Autor 2	0.68	0.86	0.76	**0.91**
Autor 3	0.81	0.86	0.83	**0.94**
Autor 4	0.78	0.73	0.75	**0.92**
Autor 5	0.90	0.66	0.76	**0.93**
Autor 6	0.86	0.86	0.86	**0.95**

Se compararon nuestros resultados con resultados tomados del *artículo Authorship Attribution Using Probabilistic Context-Free Grammars [4]*. Se debe tener en cuenta que los corpus utilizados en este artículo fueron completados con secciones de Penn Treebank [11].

La comparación de nuestros resultados para el corpus Poetry, con los reportados en [4] se presenta en la tabla 6, donde: **MaxEnt** y **NB** corresponden a los clasificadores de Máxima Entropía y el clasificador Naive Bayes, **Bigram-I** se refiere al modelo de lenguaje de bigramas con suavizado, **PCFG** es el método propuesto en [4] que utiliza gramática libre de contexto probabilística para modelar el estilo de cada autor, **PCFG-I** corresponde al modelo mencionado anteriormente con interpolación, **PCFG-E,** corresponde a la combinación de máxima entropía y PCFG y finalmente **MaxEnt + Bigram-I**, corresponde a la combinación del clasificador de máxima entropía y bigramas con interpolación. La última columna corresponde a la exactitud obtenida con los HRR.

Corpus Poetry

Tabla 6. Exactitud obtenida para el corpus Poetry

	Artículo							
	MaxEnt	NB	BIgram -I	PCFG	PCFG-I	PCFG-E	MaxEnt+ Brigram-I	HRR
Poetry	56.36	78.18	70.90	78.18	83.63	87.27	76.36	82.00

Se puede observar que el método de PFG no supera a los HRR. Quienes son superados únicamente por la combinación de más de uno de los métodos reportados en [4].

Corpus NFL

Tabla 7. Exactitud obtenida para el corpus NFL

	Artículo							
	MaxEnt	NB	BIgram -I	PCFG	PCFG-I	PCFG-E	MaxEnt+ Brigram-I	HRR
NFL	84.45	86.67	86.67	93.34	80.00	91.11	86.67	88.10

Para NFL, PCFG supera a los HRR, Tabla 7.

La Tabla 8 muestra los resultados paras Business, donde los HRR claramente superan al método de PCFG y aun a la combinación de éste con otros métodos.

Corpus Business

Tabla 8. Exactitud obtenida para el corpus Business

	Artículo							
	MaxEnt	NB	BIgram -I	PCFG	PCFG-I	PCFG-E	MaxEnt+ Brigram-I	HRR
Business	83.34	77.78	90.00	77.78	85.56	91.11	92.22	93.30

8. Conclusiones y trabajo a futuro

En base a los resultados obtenidos de los experimentos realizados con anterioridad podemos concluir lo siguiente:

Los enfoques basados en instancias combinando información léxica y sintáctica, con el uso de los HRR a través del operador de convolución circular, produce resultados poco favorables.

En corpus equilibrados mayores de 50 documentos los HRRs paracen producir resultados adecuados.

Los resultados generados por los HRRs son equiparables a los reportados en la bibliografía.

Como trabajo futuro se experimentará con corpus balanceados y de mayor tamaño para validar que el método probado se comporta de forma favorable. Así mismo se pretende representar como HRRs estructuras lingüísticas mayores que palabras aisladas.

Agradecimientos

Agradecemos a la Vicerrectoría de Investigación y Estudios de Posgrado por el soporte ofrecido para la realización de este trabajo a través del proyecto "Utilización de expresiones lingüísticas para el análisis de sentimientos".

Referencias

1. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the AmericanSociety for information Science and Technology 60(3), pp. 538–405 (2009)
2. Sindhu Raghavan Adriana Kovashka Raymond Mooney: Authorship Attribution Using Probabilistic Context-Free Grammars. In: Proceedings of the ACL 2010 Conference Short Papers, ACLShort'10. pp.1-3 (2010)
3. Adrian Pastor et al. A New Document Author Representation for Authorship Attribution. Pattern Recognition, Springer Berlin Heidelberg pp. 283-292 (2012).
4. Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic Dependency-Based N-grams as Classification Features. In: Mendoza, M.G. (ed.)MICAI 2012, Part II. LNCS (LNAI), vol. 7630, pp. 1–11. Springer, Heidelberg (2013)
5. Maya Carrillo. Representando Estructura y Significado en Procesamiento de Lenguaje Natural. , Tratamiento del Lenguaje y del Conocimiento, BUBOK PUBLISHING S.L.,(2013)
6. Tony Plate. Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. In John Mylopoulos and Ray Reiter, editors, Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI), pp30-35, Morgan Kaufmann, San Mateo, CA, 1991, 6 pages.
7. Maya Carrillo Ruiz et al. Exploring the Use of Random Indexing for Retrieving Information (2009)
8. Etiquetador Stanford. http://nlp.stanford.edu/software/pos-tagger-faq.shtml (2014)
9. Magnus Sahlgren. An Introduction to Random Indexing. In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE (2005).