Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 97-104 recibido 30-01-12 revisado 12-03-12 aceptado 18-03-12

Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes

Analysis of lexical richness in the context of latent demographic user attributes classification

John A. Roberto, M. Antònia Martí
CLiC,
Universidad de Barcelona.
Gran Via 585, 08007 Barcelona
roberto.john@ub.edu, amarti@ub.edu

Maria Salamó
Depto. Matemática Aplicada y Análisis,
Universidad de Barcelona.
Gran Via 585, 08007 Barcelona
maria@maia.ub.es

Resumen: En este artículo analizamos la utilidad que tiene el cálculo de la riqueza léxica para predecir atributos demográficos latentes en textos de opinión del español. Nuestro objetivo es determinar hasta qué punto la riqueza léxica permite predecir el sexo, la edad y la procedencia de los autores de este tipo de textos. Para ello hemos analizado 32 métricas de la riqueza léxica en 1911 textos de opinión previamente etiquetados con información demográfica. Esta aproximación tiene como principales ventajas la independencia del dominio y la reducción del coste computacional.
Palabras clave: Sistemas de Recomendación, categorización de textos, riqueza léxica

Abstract: In this paper we analyse the utility of Lexical richness estimations to predict latent user attributes shown in Spanish opinionated texts. Our aim is to establish how useful could be the Lexical richness to predict user's gender, age and regional origin. Because of this goal, we applied 32 lexical richness measures to 1911 previously labeled texts with demographic information. This approach has the advantage that it is domain-independent with modest computational cost.
Keywords: Recommender Systems, text categorization, lexical richness

1. Introducción

Los Sistemas de Recomendación (SR) utilizan el conocimiento que tienen del usuario para personalizar las recomendaciones. La generación de los perfiles de usuario es, por lo tanto, una tarea fundamental en el diseño de estos sistemas. La información demográfica, p.e. edad o sexo, suele estar en la base de todo perfil. Actualmente se buscan alternativas no intrusivas para la obtención de dichos perfiles mediante la interpretación de las acciones y comportamientos del usuario. Analizar los textos que produce el usuario (comportamiento verbal) para predecir sus atributos demográficos, puede constituir una vía indirecta pero efectiva de crear y mantener los perfiles.

El análisis de textos con el propósito asignarles un atributo demográfico predefinido, se ha considerado habitualmente una tarea típica de la categorización de textos (Sebastiani, 2002). No obstante, tratar la asignación de atributos demográficos como un caso de categorización de textos presenta una doble problemática. Por un lado, comporta elevados costes computacionales: el uso de n-gramas exige la habilidad para manejar una alta dimensionalidad de espacio de rasgos. Por otro, la categorización de textos suele estar determinada por el contenido temático del documento, lo cual hace que sea dependiente del dominio (Sarawgi, Gajulapalli, y Choi, 2011).

En este artículo evaluamos el uso de la riqueza léxica para la extracción de atributos demográficos latentes. A diferencia de los métodos para la categorización de textos, el cálculo de la riqueza léxica tiene un coste computacional muy bajo y es totalmente independiente del dominio. Esto se debe a que la riqueza léxica evalúa los textos en términos de la variedad y la cantidad del vocabulario que contienen de acuerdo con la relación types/tokens.

La evaluación se ha llevado a cabo mediante una serie de experimentos que usan una aproximación basada en aprendizaje automático. En los experimentos hemos aplicado métricas estándar de riqueza léxica sobre 1911 textos de opinión en español agrupados

ISSN 1135-5948

por atributos demográficos y clases binarias.

El artículo está organizado en 5 secciones. En la sección 2 se describe brevemente los trabajos relacionados. En la sección 3 presentamos el método y el corpus que hemos usado para el análisis. En la sección 4 describimos los experimentos y los resultados obtenidos. Finalmente, en la sección 5 ofrecemos las conclusiones y el trabajo futuro.

2. Trabajos relacionados

La extracción de atributos demográficos latentes –atributos demográficos no expresados de forma explícita por el autor de un texto– es una tarea orientada a la detección de perfiles de usuario (*user profiling*). Esta tarea está relacionada con la clasificación automática de textos y la estilometría (Koppel, Argamon, y Shimoni, 2003) pero se diferencia de aquellas en que no clasifica textos por su contenido temático ni por el estilo individual de su autor[1].

No obstante, por su proximidad con la clasificación de textos y la estilometría, para la extracción de atributos demográficos latentes se aplican métodos pertenecientes a ambos campos (Koppel, Schler, y Argamon, 2009). Por ejemplo, se suele emplear grandes conjuntos de rasgos que reflejan el contenido de un documento (p.e. palabras clave) en combinación con métodos de selección automática de rasgos, los métodos estadísticos como el análisis multivariante o las técnicas de aprendizaje automático. Los trabajos que citamos a continuación son una muestra de esta combinación de enfoques.

En Schler et al. (2006) se usan rasgos de estilo y de contenido para clasificar documentos por la edad y el sexo de su autor. Dentro del primer grupo de rasgos los investigadores seleccionan secuencias de clases de palabras (*part of speech*, POS), palabras funcionales, neologismos y palabras que aparecen con una alta frecuencia en blogs. Como rasgos de contenido seleccionan las palabras con contenido léxico y un listado de palabras especiales tomadas del programa para el análisis de textos LIWC[2]. Para cada rasgo se midió la frecuencia de aparición en el corpus por sexo y por edad.

También para predecir la edad y el sexo, Kabbur, Han, y Karypis (2010) usan una aproximación basada en aprendizaje automático. Los autores representan cada documento (páginas web) con un conjunto de rasgos de contenido (términos) y estructurales (tags HTML).

Recientemente, Sarawgi, Gajulapalli, y Choi (2011) combinan técnicas estadísticas y de aprendizaje automático para predecir el sexo de los autores de blogs y de artículos científicos. Su principal objetivo consiste en hacer la clasificación sin tener que recurrir a información relacionada con el tema de los documentos. Los rasgos que emplean se basan en patrones morfológicos, léxico-sintácticos y relaciones de dependencia de larga distancia.

Como ejemplo de predicción de la procedencia del usuario tenemos el trabajo de Koppel, Schler, y Zigdon (2005). En su trabajo, utilizan una serie de rasgos estilísticos para clasificar los autores de textos en inglés según su procedencia. Los investigadores agrupan en tres categorías dichos rasgos: palabras funcionales, n-gramas de caracteres y errores ortográficos y sintácticos. Un trabajo similar a este, sólo que usando rasgos léxicos y estructurales, es el de Estival et al. (2007). Finalmente, una investigación que intenta ampliar el conjunto de rasgos de estos dos modelos es el artículo de Jojo y Dras (2009). Los investigadores recurren a errores sintácticos causados por interferencia con la lengua de origen para predecir la procedencia de autores de textos en inglés.

A diferencia de los trabajos anteriores que para identificar atributos demográficos se basan en la cantidad y complejidad de los rasgos, en esta investigación usamos conjunto reducido de rasgos que obtenemos mediante el cálculo de la riqueza léxica. Los beneficios de esta aproximación son la independencia del dominio y del idioma, la facilidad para extraer los rasgos y el bajo coste computacional.

3. Análisis

El objetivo de este artículo es analizar la viabilidad del empleo de la riqueza léxica para la extracción de atributos demográficos latentes. A continuación, presentamos las métricas de riqueza léxica y los datos utilizados para cumplir con este objetivo.

[1] Según Koppel, Argamon, y Shimoni (2003), si consideramos que los autores individuales exhiben hábitos más consistentes en cuanto a estilo que los autores agrupados por clases, la detección de perfiles es una tarea más ardua que la atribución de autoría.

[2] *Linguistic Inquiry and Word Count* (http://www.liwc.net/)

3.1. Las métricas

La riqueza léxica –en adelante nos referiremos como RiqLex– (Read, 2005) se compone de tres dimensiones que son la densidad (DenLex), la sofisticación (SofLex) y la variación (VarLex). La DenLex (Ure, 1971) calcula la proporción entre el número de palabras léxicas y el número total de palabras (léxicas y gramaticales) de un texto. La SofLex (Read, 2005), también conocida como singularidad léxica, mide la proporción de palabras "sofisticadas o avanzadas" presentes en un texto. Finalmente, la VarLex (Granger y Wynne, 2000), denominada diversidad léxica (Malvern et al., 2004) y ámbito léxico (Crystal, 1982), mide el número de *types* o palabras diferentes que hay en un texto.

Adicionalmente, Laufer y Nation (1995) reconocen una dimensión más, la originalidad léxica. No obstante, el índice de originalidad léxica (OriLex) es una medida individual: se emplea para comparar la producción escrita de una persona con relación al grupo. La OriLex calcula el porcentaje de palabras que usa la persona y que no son usadas por nadie más dentro de ese mismo grupo. Por este motivo no se emplea en esta investigación.

La mayoría de las métricas que hemos seleccionado para predecir los atributos edad, sexo y procedencia, son transformaciones algebraicas que trabajan con índices léxicos para compensar la variación en la longitud de los textos. Hemos optado por este método ya que consideramos que representa un nivel más avanzado en el calculo de la riqueza léxica. Además, con ello evitamos tener que recurrir a métodos más simples de estandarización como, por ejemplo, dividir todos los textos en segmentos de la misma longitud (Thordardottir y Weismer, 2001; Jarvis, 2002) o hacer una selección aleatoria de palabras (Breeder, Extra, y van Hout, 1986). Estos últimos se consideran procedimientos "derrochadores" (*wasteful*) pues prescinden de datos potencialmente relevantes al tiempo que dificultan la reproducción de los experimentos y el contraste de los resultados (Malvern et al., 2004; Lu, 2011).

En el Cuadro 1 presentamos el conjunto de métricas que hemos aplicado. En total hay una medida de la DenLex (1), cinco de la SofLex (2-6) y 22 de la VarLex (7-29). El número real de métricas aplicadas es de 32 pues en las fórmulas 16, 20, 25 y 29 el subíndice $_G$ puede ser interpretado como n (nombre), v (verbo), a (adjetivo) o r (adverbio) puesto que son las categorías de base léxica. Adicionalmente, para cada medida se especifican los siguientes valores (según columna): la dimensión a la que pertenece (Dim.), un identificador de la métrica (Id.), el nombre de la métrica, una etiqueta que facilita su posterior identificación en la sección de resultados, la fórmula y una referencia bibliográfica. En la parte inferior de la tabla aparecen las convenciones necesarias para interpretar correctamente las fórmulas.

En la definición de las métricas hemos asumido que las palabras de contenido léxico son los nombres, los adjetivos, los verbos y los adverbios. Dentro de las palabras de contenido gramatical estarían las preposiciones, las conjunciones, el artículo y los pronombres. De la misma forma, para determinar los índices de sofisticación nos hemos apoyado en una lista de las 5000 formas más frecuentes del español de acuerdo con la RAE[3]. Según Laufer y Nation (1995), las palabras sofisticadas no están en la lista de las 1000 o 2000 palabras más frecuentes de un idioma. En nuestro caso hemos usado las 5000 ya que la lista de la RAE contiene tanto palabras léxicas como gramaticales.

3.2. Los datos

Los textos sobre los cuales aplicamos las métricas de la RiqLex pertenecen al corpus Hopinion. Este corpus contiene más de 18.000 textos de opinión en castellano provenientes de la web de TripAdvior (*www.tripadvisor.es*). Cada texto tiene una extensión aproximada de 150 palabras. De este conjunto de textos seleccionamos 1911 por estar anotados con información morfológica y demográfica.

En primer lugar, mediante la anotación morfológica a cada palabra del texto se le asigna automáticamente un lema y una categoría gramatical (POS). Los errores originados en el proceso automático de anotación morfológica, han sido corregidos manualmente con el objetivo de detectar variantes ortográficas propias del registro coloquial.

En segundo lugar, usando los metadatos del perfil publico de cada usuario en TripAdvisor, a cada texto se le asignan tres etiquetas: edad, sexo y procedencia. En los pocos casos en los que el usuario no ha declarado uno de estos tres atributos demográfi-

[3] http://corpus.rae.es/lfrecuencias.html

Dim.	Id	Métrica	Etiqueta	Fórmula	Referencia
DenLex	1	Densidad léxica	DL	$\frac{N_{lex}}{N}$	(Engber, 1995)
SofLex	2	Sofisticación léxica	SL	$\frac{N_{slex}}{N_{lex}}$	(Linnarud, 1986; Hyltenstam, 1988)
	3	Perfil de Frecuencia Léxica	PFL5	$\frac{T_s}{T}$	(Laufer y Nation, 1995)
	4	Sofisticación Verbal I	SV-I5	$\frac{T_{vs}}{N_v}$	(Harley y King, 1989)
	5	Sofisticación Verbal II	SV-II5	$\frac{T^2_{vs}}{N_v}$	(Chaudron y Parker, 1990)
	6	Sofisticación Verbal Corregida	SVC5	$\frac{T_{vs}}{\sqrt{2N_v}}$	(Wolfe-Quintero, Inagaki, y Kim, 1998)
VarLex	7	Type/Token ratio	TTR	$\frac{T}{N}$	(Templin, 1957)
	8	Root TTR	RTTR	$\frac{T}{\sqrt{N}}$	(Guiraud, 1960)
	9	TTR Bilogarítmico	TTRB	$\frac{LogT}{LogN}$	(Herdan, 1960)
	10	TTR Corregido	TTRC	$\frac{T}{\sqrt{2N}}$	(Carroll, 1964)
	11	a^2	ac	$\frac{logN - logT}{log^2 N}$	(Maas, 1972; Tweedie y Baayen, 1998)
	12	Indice de Uber I	UI	$\frac{(logN)^2}{logN - logT}$	(Dugast, 1979; Tweedie y Baayen, 1998)
	13	K de Yule	YuleK	$10^4 \times (\sum i^2 T_i - N_{lex})/N^2_{lex}$	(Yule, 1944; Smith y Kelly, 2002; Miranda-García y Calle, 2005; Tweedie y Baayen, 1998)
	14	Z de Zipf	ZIPF	$\frac{Z \times N \times log(N/Z)}{(N-Z)log(p \times Z)}$	(Smith y Kelly, 2002; Tweedie y Baayen, 1998)
	15	Variación de Palabras Léxicas VPL		$\frac{T_{lex}}{N_{lex}}$	(Engber, 1995)
	$16_{n,v,a,r}$	Variación G I	VG-I	$\frac{T_G}{N_{lex}}$	(McClure, 1991; Harley y King, 1989)
	$20_{n,v,a,r}$	Variación G II	VG-II	$\frac{T_G}{N_G}$	
	24	Variación Mod.	VM	$\frac{(T_a + T_r)}{N_{lex}}$	
	$25_{n,v,a,r}$	Variación G Cuadrada	VG2-II	$\frac{T^2_G}{N_G}$	(Wolfe-Quintero, Inagaki, y Kim, 1998)
	$29_{n,v,a,r}$	Variación G Corregida	VGC-II	$\frac{T_G}{\sqrt{2N_G}}$	

Convenciones:

N = tokens
T = types
$_{lex}$ = unidades léxicas
$_s$ = unidades sofisticadas
$_G$ = categoría gramatical (n, v, a, r)
$_n$ = nombre
$_v$ = verbo
$_a$ = adjetivo
$_r$ = adverbio
T_i = número de Types léxicos que ocurren i veces
Z = una medida de la riqueza léxica
p = Token más frecuente dividido por la longitud del texto

Cuadro 1: Métricas empleadas para el cálculo de la riqueza léxica.

cos, hemos recurrido a la información contextual para deducirlo. Por ejemplo, el alias "ANA1983Madrid" nos dice que se trata de una mujer, menor de 34 años y que vive en Madrid.

En el Cuadro 2 presentamos la distribución de los datos atendiendo al atributo (columna 1), las clases (binarias) en las que cada atributo ha sido agrupado (columna 2), el número de muestras que han sido etiquetadas bajo cada clase (columna 3) y el número total de muestras por atributo (columna 4). Con el propósito de contar con una distribución alternativa de los datos, para edad hemos agrupado las muestras en dos subconjuntos: Edad (Ed) y Edad2 (Ed2). Un primer subconjunto con las clases $\geq$35 (1044 textos) y $\leq$34 (867 textos) y un segundo subconjunto con las clases $\geq$50 (199 textos) y $\leq$24 (86 textos). De este último subconjunto hemos descartado 1626 muestras que corresponden a la franja de edad comprendida entre los 25

y los 49 años.

Atributos	Clases	Muestras	Total
Sexo (Sx)	Hombres	948	1911
	Mujeres	963	
Edad (Ed)	≥35	1044	1911
	≤34	867	
Edad2 (Ed2)	≥50	199	1911
	≤24	86	
	≥25-≤49	**1626**	
Procedencia (Pr)	España	1450	1911
	América	461	

Cuadro 2: Distribución de las muestras

Finalmente, las 32 medidas de riqueza léxica fueron aplicadas sobre los textos, lo que devuelve una serie de valores que se usaron como atributos para entrenar diferentes clasificadores. Las pruebas se confeccionaron escogiendo 90 % de datos para el training y 10 % para validar las pruebas (*test*). Todos los clasificadores fueron entrenados usando validación cruzada ($ten - fold\ cross - validation$). Para realizar los experimentos usamos Weka (Witten y Frank, 2000). Los acronimos de los algoritmos usados se pueden encontrar al pie de la tabla y figura de resultados.

El corpus utilizado junto con las métricas y los resultados de los experimentos están disponibles en la web de CLiC[4].

4. *Experimentos y resultados*

Los experimentos que realizamos están enfocados a determinar hasta qué punto el cálculo de la RiqLex ayuda a inferir el sexo (Sx), la edad (Ed) y la procedencia (Pr) de los autores de textos de opinión. Para ello hemos realizado tres experimentos.

4.1. Experimento 1

En el primer experimento hemos realizado un análisis preliminar aplicando las 32 métricas de RiqLex (ver Cuadro 1) directamente sobre los textos agrupados por atributos (Sx, Ed y Pr) y clases. Los valores que retornaron cada una de las métricas se usaron como conocimiento para entrenar una serie de clasificadores (Cuadro 3). Nuestro objetivo era medir el rendimiento general de las métricas de RiqLex para clasificar los textos por Sx, Ed y Pr.

El Cuadro 3 presenta un resumen de este experimento. Cada columna contiene la precisión (*Prediction Accuracy*, PA) obtenida al predecir cada atributo de acuerdo con sus respectivas clases binarias: Sx (hombre/mujer), Ed (≥35/≤34), Ed2 (≤24/≥50) y Pr (español/latinoamericano). En la primera fila tenemos el PA de los mejores clasificadores, en la segunda fila el PA de los peores y en la última fila el PA promedio de todos los clasificadores. Al final del cuadro está la lista de los algoritmos empleados.

Como podemos ver en el Cuadro 3, el rendimiento promedio de todos los clasificadores para el atributo Sx es de 55.6 % y para Ed de 54.9 %. Los mejores PAs para estos mismos atributos están en torno al 60 %. Estos resultados son muy discretos. Por el contrario, el rendimiento de los clasificadores mejora cuando los atributos a predecir son Ed2 y Pr, su promedio de PA supera el 70 %. De hecho, el clasificador con mejor desempeño en el caso de Ed2 alcanza el 98.8 % de precisión y es del 77.3 % para Pr. Estos resultados indican que existen diferencias léxicas importantes asociadas con la procedencia y, en menor medida, la edad de los usuarios.

	Sx	**Ed**	**Ed2**	**Pr**
mejor	60	57.3	98.8	77.3
peor	53.1	52.6	67	58.4
promedio	55.6	54.9	81	73.7

bayes (BayesNet, BayesianLogisticRegression, NaiveBayes, NaiveBayesSimple, NaiveBayesUpdateable), lazy (IB1, IBk, KStar, LWL), misc (HyperPipes, VFI), rules (ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor, ZeroR), trees (ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, REPTree, RandomForest, RandomTree, SimpleCart)

Cuadro 3: Precisión (PA) obtenida en el primer experimento.

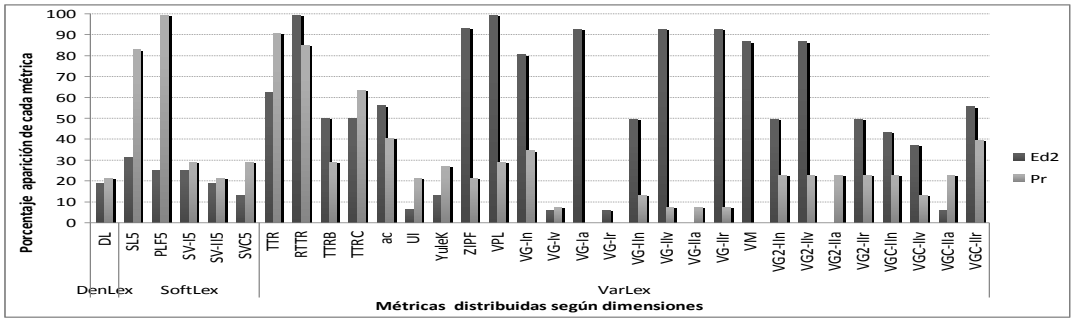
4.2. Experimento 2

Empleando los mismos algoritmos de clasificación del experimento anterior (ver pie Cuadro 3), en este experimento entrenamos de nuevo los clasificadores agrupando las métricas según la dimensión de la RiqLex que evalúan. Nuestro objetivo era determinar si alguna dimensión de la RiqLex predice mejor que otra los atributos demográficos.

Obtenemos, por tanto, tres grupos de clasificadores, uno por cada dimensión. El primer grupo aprende de la única métrica de DenLex que tenemos (DL) y de dos de los atributos demográficos (Ed, Sx y Pr) que incorporamos como parte de su conocimiento[5].

[4] http://clic.ub.edu/

[5] El atributo demográfico omitido es, evidentemen-



(a) DMNBtext, NaiveBayes, IBk, rules.DTNB / OneR / PART, trees.ADTree / FT / J48 / J48graft / LADTree / LMT / RandomForest. (b) Cfs, ChiSquared, Consistency, Filtered, InfoGain, PCA, ReliefF. (c) BestF., GeneticS., LinearForwardS., RankS., Ranker.

Figura 1: Frecuencia de uso de las métricas tras la selección de rasgos: (a) algoritmos, (b) evaluadores y (c) selectores.

El segundo grupo aprende de las métricas de SofLex (2 a la 6 en el Cuadro 1) y, también, del par de atributos demográficos. El tercer grupo extrae el conocimiento de las medidas de VarLex (7 a la 29) e, igualmente, de los atributos demográficos.

En el Cuadro 4 tenemos los promedios de predicción obtenidos con todos los clasificadores agrupando las métricas por dimensiones (VarLex, SofLex, DenLex). En la primera columna de la tabla (RiqLex) mantenemos los promedios de predicción obtenidos en el experimento anterior y que resultan de aplicar todas las métricas simultáneamente (ver Cuadro 3). Los números entre paréntesis señalan los rangos (R), es decir, la importancia de cada dimensión para cada uno de los atributos evaluados (Sx, Ed, Ed2 y Pr): el rango (1) indica el mejor PA y el rango (4) el peor. La suma de los rangos, en la última fila (Rango), refleja la posición definitiva que ocupa cada dimensión.

En el cuadro observamos que la RiqLex obtiene los mejores rangos (1, 2, 1, 3) que cualquiera de las dimensiones que la componen de forma individual. Esto significa que las 32 métricas en conjunto funcionan mejor que por separado. No obstante, de cara a reducir aún más el número de rasgos con los que se trabaja, las métricas de VarLex o de DenLex predicen mejor que las de SofLex. En efecto, si atendemos a los rangos de la SofLex (3, 3, 3, 4) comprobaremos que las cinco métricas que la componen son las menos efectivas para predecir cualquiera de los atributos demográficos.

La conclusión de este análisis es que la RiqLex (suma de todas las dimensiones) es más efectiva para predecir los atributos demográficos. VarLex y DenLex estarían en un segundo plano, con niveles similares de rendimiento. Finalmente, la edad, el sexo y la procedencia de los usuarios está poco relacionada con los niveles de sofisticación o singularidad léxica.

	RiqLex		**VarLex**		**SofLex**		**DenLex**	
	PA	*R*	*PA*	*R*	*PA*	*R*	*PA*	*R*
Sx	55.62	(1)	54.80	(2)	54.49	(3)	53.72	(4)
Ed	54.92	(2)	53.62	(4)	53.93	(3)	56.42	(1)
Ed2	81.02	(1)	80.25	(2)	78.36	(3)	76.91	(4)
Pr	73.76	(3)	74.64	(2)	73.72	(4)	74.78	(1)
Rango		(7)		(10)		(13)		(10)

Cuadro 4: Precisión (PA) obtenida para cada atributo latente.

4.3. Experimento 3

En el tercer experimento recurrimos a la técnica de selección de rasgos con el objetivo de detectar las métricas más útiles en el caso de los atributos que han obtenido una precisión superior al 70 %, esto es, Ed2 y Pr.

La configuración de este experimento es la siguiente. Empleamos los 13 mejores algoritmos de clasificación de los experimentos anteriores, conjuntamente con 7 métodos evaluadores y 5 selectores (ver la nota al pie en la Figura 3). En el caso de Ranker, el selector se configuró para que recuperara solo los

te, el atributo a aprender.

25 atributos más discriminantes ya que por defecto solo pondera y se ha de establecer manualmente el nivel de selección.

En primer lugar, en la Figura 1 tenemos el uso promedio que todos los clasificadores hacen de cada una de las métricas según las tres dimensión de la RiqLex. Por ejemplo, se observa que 20 de las 32 métricas están por encima de un 30 % de aparición para Ed2, mientras solo 8 de las métricas lo están para Pr. Esto significa que hacen falta más rasgos para predecir Ed2 que Pr.

En segundo lugar, en la misma figura vemos que para predecir Ed2 destacan las métricas de la VarLex basadas en categorías gramaticales como VG-Ir, VG-IIa y VG2-IIa. Por su parte, las métricas más frecuentes para predecir Pr corresponden al TTR tradicional y sus variaciones (RTTR, TTRB y TTRC) así como a dos métricas de la SofLex (SL5 y PLF5).

De acuerdo con estos resultados podemos concluir que el análisis de las palabras con contenido léxico es una buena opción para clasificar usuarios por edad dada la alta frecuencia de aparición de las medidas que utilizan este tipo de palabras. Además, los resultados apuntan a que una clasificación basada en el análisis de la frecuencias de palabras léxicas conjuntamente con las gramaticales favorece la detección de usuarios por procedencia.

5. Conclusiones y trabajos futuros

En el presente artículo hemos analizado la utilidad que tiene el cálculo de la riqueza léxica para predecir atributos demográficos latentes en textos de opinión del español. Analizando un gran número de métricas estándar basadas en transformaciones algebraicas hemos conseguido detectar las dimensiones y las métricas que mejor funcionan para cada uno de los atributos demográficos a predecir.

Los principales beneficios de esta aproximación están en su bajo coste computacional, su independencia del dominio y que por su simplicidad, facilita enormemente la extracción de rasgos. En consecuencia, las métricas de riqueza léxica pueden adaptarse como algoritmos de referencia (*baselines*) en la extracción de atributos demográficos latentes.

El trabajo futuro está enfocado a mejorar el rendimiento de los clasificadores explorando el uso de nuevas métricas y a extender el análisis a otros dominios.

Bibliografía

Breeder, P., G. Extra, y R. van Hout. 1986. Measuring lexical richness and diversity in second language research. *Polyglot*, 8:1–16.

Carroll, J. 1964. *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Chaudron, C. y K. Parker. 1990. Discourse markedness and structural markedness: The acquisition of english noun phrases. *Studies in Second Language Acquisition*, 12:43–64.

Crystal, D. 1982. *Profiling Linguistic Disability*. London: Edward Arnold.

Dugast, D. 1979. *Vocabulaire et stylistique. I Théâtre et Dialogue. Travaux de linguistique quantitative*. Geneva: Slatkine-Champion.

Engber, C. 1995. The relationship of lexical proficiency to the quality of esl compositions. *Journal of Second Language Writing*, 4(4):139–155.

Estival, D., T. Gaustad, S. Pham, W. Radford, y B Hutchinson. 2007. Author profiling for english emails. *Proc. of the 10th Conference of the Pacific Association for Computational Linguistics*, páginas 263–272.

Granger, S. y M. Wynne. 2000. Optimising measures of lexical variation in efl learner corpora. En John M. Kirk, editor, *Corpora galoreh*. Amsterdam: Rodopi, páginas 249–258.

Guiraud, P. 1960. *Problemes et methodes de la statistique linguistique*. Dordrecht, The Netherlands: D. Reidel.

Harley, B. y M. King. 1989. Verb lexis in the written compositions of young l2 learners. *Studies in Second Language Acquisition*, 11:415–440.

Herdan, G. 1960. *Quantitative linguistics*. Butterworth, London.

Hyltenstam, K. 1988. Lexical characteristics of near-native second-language learners of swedish. *Journal of Multilingual and Multicultural Development*, 9:67–84.

Jarvis, S. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1):57–84.

Jojo, S. y M. Dras. 2009. Contrastive analysis and native language identification. *Proc. of the Australasian Language Technology Workshop*, páginas 53–61.

Kabbur, S., E. Han, y G. Karypis. 2010. Content-based methods for predicting web-site demographic attributes. *Proc. of ICDM'2010*, páginas 863–868.

Koppel, M., S. Argamon, y A. Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 14(4).

Koppel, M., J. Schler, y S. Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Koppel, M., J. Schler, y K. Zigdon. 2005. Automatically determining an anonymous author's native language. *IEEE International Conference on Intelligence and Security Informatics*, 3495/2005:41–76.

Laufer, B. y P. Nation. 1995. Vocabulary size and use: lexical richness in l2 written production. *Applied Linguistics*, 16:307–322.

Linnarud, M. 1986. *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund: CWK Gleerup.

Lu, X. 2011. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*.

Maas, H. 1972. Zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 8:73–79.

Malvern, D., B. Richards, N. Chipere, y P. Duran. 2004. *Lexical diversity and language development: Quantification and assessment*. Houndmills.

McClure, E. 1991. A comparison of lexical strategies in l1 and l2 written english narratives. *Pragmatics and Language Learning*, 2:141–154.

Miranda-García, A. y J. Calle. 2005. Yule's characteristic k revised. *Language Resources and Evaluation*, 39:287–294.

Read, J. 2005. *Assessing vocabulary*. Cambridge University Press, 5 edición.

Sarawgi, R, K Gajulapalli, y Y Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. *Proc. of the Fifteenth Conference on Computational Natural Language Learning*.

Schler, J., M. Koppel, S. Argamon, y J. Pennebaker. 2006. Effects of age and gender on blogging. En *Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, mar.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 31(1).

Smith, J. y C. Kelly. 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36:411–430.

Templin, M. 1957. *Certain language skills in children: Their development and interrelationships*. The University of Minnesota Press.

Thordardottir, E. y S. Weismer. 2001. High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language and Communication Disorders*, 36:221–244.

Tweedie, F. y H. Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

Ure, J. 1971. Lexical density and register differentiation. En G. E. Perren y J. L. M. Trim, editores, *Applications of linguistics*, páginas 443–452. Cambridge University Press.

Witten, I. y E. Frank. 2000. *DataMining*. Morgan Kaufmann Publishers.

Wolfe-Quintero, K., S. Inagaki, y H. Kim. 1998. Second language development in writing: Measures of fluency, accuracy, and complexity. Informe técnico, University of Hawai'i, Second Language Teaching and Curriculum Center.

Yule, G. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.