Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo

Automatic categorization of Spanish texts into linguistic registers: a contrastive analysis

John A. Roberto
Universidad de Barcelona
roberto.john@ub.edu

Maria Salamó
Universidad de Barcelona
maria.salamo@ub.edu

M. Antònia Martí
Universidad de Barcelona
amarti@ub.edu

Resumen

Las aplicaciones colaborativas como los Sistemas de Recomendación se pueden beneficiar de la clasificación de textos en registros lingüísticos. En primer lugar, el registro lingüístico proporciona información sobre el perfil de los usuarios y sobre el contexto de la recomendación. En segundo lugar, considerar las características de cada tipo de texto puede ayudar a mejorar los métodos actuales de procesamiento de lenguaje natural. En este trabajo contrastamos dos enfoques, uno morfosintáctico y el otro léxico, para categorizar textos por registro en español. Para su evaluación aplicamos 38 algoritmos de aprendizaje automático con los que obtuvimos niveles de precisión superiores al 89 %.

Palabras clave

Procesamiento del lenguaje natural, aprendizaje automático, registro lingüístico.

Abstract

Collaborative software such as Recommender Systems can benefit from the automatic classification of texts into linguistic registers. First, the linguistic register provides information about the users' profiles and the context of the recommendation. Second, considering the characteristics of each type of text can help to improve existing natural language processing methods. In this paper we contrast two approaches to register categorization for Spanish. The first approach is focused on morphosintactic patterns and the second one on lexical patterns. For the experimental evaluation we tested 38 machine learning algorithms with a precision higher than 89 %.

Keywords

Natural language processing, machine learning, linguistic register.

1 Introducción

El uso de aplicaciones colaborativas permite a los usuarios crear un tipo de contenidos que por su marcado carácter subjetivo y libre es difícil de tratar computacionalmente. Por esa razón, para que estas aplicaciones funcionen de manera adecuada, deberían garantizar que el tipo de texto que están tratando cumple ciertas condiciones básicas. Por ejemplo, los Sistemas de Recomendación que usan texto libre como mecanismo de reatroalimentación deben garantizar que están procesando opiniones de usuarios reales (*reviews*) y no resúmenes, comentarios o anécdotas sobre determinados productos.

Un sistema recomendador deberá, por tanto, ser capaz de identificar que un texto como el del Ejemplo 1 (tomado de una conocida web de cine) es por su lenguaje, estructura y estilo el comentario de un experto o el resumen del film y no un *review*—entendido este último como un tipo de texto subjetivo que describe la experiencia, el conocimiento y la opinión de un usuario con respecto a un producto (Ricci y Wietsma, 2006).

> Ejemplo **1** *Algo ha pasado, el mundo se va al garete. Una pareja forzosa (José Coronado y Quim Gutiérrez) deberá formar equipo para encontrar a sus seres queridos. Tras las buenas sensaciones dejadas con la aventura americana de "Infectados" (2009), los hermanos Álex y David Pastor regresan a España con "Los últimos días", estupendo drama de ciencia-ficción apocalíptica que les confirma como dos talentos a no dejar escapar. Ahora depende del espectador, claro, y de cómo responda esta crepuscular odisea en nuestro circuito de salas. Sea como fuere, estamos ante uno de los mejores títulos de género nacional de los últimos años. Ni más ni menos.*

Una forma de "validar" los *reviews* es atendiendo a su registro lingüístico. Junto con el género, la temática o el estilo, el registro es un indicador del tipo de texto. En este artículo contrastamos dos aproximaciones para la detección automática del registro lingüístico en español: una aproximación basada en patrones morfosintácticos y otra basada en patrones léxicos. Ambas

aproximaciones tienen en común la simplicidad y reducido coste computacional, dos características indispensables en el diseño de sistemas colaborativos.

En el apartado 2 revisamos las principales aproximaciones empleadas para la clasificación de textos por género o registro. En el apartado 3 describimos los corpus AnCora-ES y Hopinion usados en ésta investigación. En el apartado 4 presentamos las características usadas en cada una de las aproximaciones (morfosintáctica y léxica) y, en el apartado 5, analizamos su rendimiento. Finalmente, en la sección 7 ofrecemos las conclusiones.

2 Estado del arte

Los trabajos relacionados con la clasificación automática de textos por registro se pueden agrupar considerando tres factores: la unidad de análisis que hacen servir, la aproximación empleada para la clasificación de los textos y el tipo de rasgos que emplean.

En primer lugar, si consideramos la unidad de análisis tenemos trabajos que se centran en la palabra (Brooke, Wang, y Hirst, 2010), la oración (Lahiri, Mitra, y Lu, 2011) o el documento (Sheikha y Inkpen, 2010). Así, (Brooke, Wang, y Hirst, 2010) exploran diferentes métodos para determinar el nivel de formalidad de ítems léxicos. Su objetivo es clasificar palabras que comparten significado pero no registro, por ejemplo "acquire" *adquirir* (formal) y "snag" *agarrar* (informal). Para conseguir su objetivo, los autores emplean la longitud de las palabras mediante *FS score* (Simple Formality Measure), LSA (Latent Semantic Analysis) y un método híbrido que combina los dos anteriores. Por su parte, (Lahiri, Mitra, y Lu, 2011) analizan el grado de formalidad a nivel de la oración mediante el cálculo del F-score (Formality Score) y evaluando el grado de acuerdo entre anotadores a partir de los coeficientes Kappa y Jaccard. Por último, los trabajos que vienen a continuación tratan el problema del registro a nivel de documento.

En segundo lugar, las aproximaciones empleadas para la clasificación de textos por registro o género se basan en el análisis de patrones lingüísticos y en el uso de técnicas de aprendizaje automático. La propuesta más representativa basada en el análisis de patrones lingüísticos es la metodología del análisis multidimensional (MDA) de Biber (1988). En su trabajo, Biber aplica 67 rasgos lingüísticos para identificar 21 géneros del inglés hablado y escrito. Biber determina los rasgos que concurren en un mismo género mediante técnicas estadísticas de análisis de factores. En la misma línea, (Tribble, 1999) propone un método menos complejo que el MDA consistente en caracterizar los textos a partir de la detección de palabras clave. Tribble extrae listas de palabras de forma automática y las compara con un corpus de referencia para obtener las palabras más relevantes dentro de un género o registro. El método de Tribble fue contrastado con el de Biber por (Xiao y McEnery, 2005) obteniendo resultados similares.

Más recientes pero también más escasas, son las aproximaciones basadas en técnicas de aprendizaje automático (especialmente para el español). En cuanto al aprendizaje supervisado, (Sharoff, Zhili, y Katja, 2010) emplean el algoritmo SVM (*Support Vector Machines*), un modelo basado en trigramas de etiquetas POS y las anotaciones del corpus Brown (Francis y Kucera, 1979) para la detección del género en textos de la Web. Según los autores, el uso de rasgos léxicos es más efectivo para detectar el género de un documento que la información basada en Part-Of-Speech. En (Gries, Newman, y Shaoul, 2009) se presenta un algoritmo de clústering aglomerativo jerárquico basado en n-gramas para detectar registros en textos provenientes de dos corpus diferentes: el BNC-Baby (*British National Corpus Baby*) y el ICE-GB (*British Component of the International Corpus of English*).

En tercer lugar, como comentamos, las investigaciones en clasificación automática de textos por registro se pueden agrupar considerando el tipo de rasgos que hacen servir. Son características las bolsas de palabras (zu Eissen y Stein, 2004), n-gramas de caracteres (Mason, Shepherd, y Duffy, 2009; Kanaris y Stamatatos, 2007) y Part-Of-Speech (Santini, 2007), así como el uso de etiquetas HTML para analizar el metacontenido de las páginas (Boese y Howe, 2005). Más reciente es el uso de rasgos léxico-gramaticales (Sheikha y Inkpen, 2010) como pueden ser la frecuencia de interjecciones, palabras formales e informales, uso de la forma pasiva, etc. En estas aproximaciones se suele evaluar el rendimiento y la utilidad de tales rasgos mediante la aplicación de técnicas de aprendizaje automático como las descritas.

En español tenemos el trabajo de (Mosquera y Moreda, 2011). Los autores de esa investigación identifican grados de informalidad en textos de la Web 2.0 usando un algoritmo de "hard-clustering" (K-Means) y un conjunto de 19 características léxico-gramaticales. Las conclusiones a las que llegan son positivas aunque señalan la necesidad de añadir nuevas características.

La principal novedad de nuestra propuesta en relación con las anteriores investigaciones consiste en hacer un análisis exhaustivo, con 38 algoritmos de aprendizaje automático y diversas técnicas de selección de atributos, de dos aproximaciones diferentes para la clasificación de textos según su registro. A diferencia de Mosquera, nosotros usamos aprendizaje supervisado para categorizar los textos en dos grandes grupos (formal e informal) y esto lo hacemos distinguiendo entre características morfosintácticas y léxicas. En cuanto a estas últimas, otro aporte a destacar es la incorporación de diferentes métricas de la riqueza léxica para la detección del registro lingüístico.

3 Datos

En este artículo se ha utilizado un subconjunto de 3270 textos tomados de los corpus AnCora-ES y Hopinion. Usamos estos dos corpus ya que representan registros opuestos del español actual:

- AnCora-ES es un corpus del español formal constituido principalmente por artículos periodísticos. AnCora-ES está anotado con diferentes tipos de información lingüística, por ejemplo, *Part of Speech*, estructura argumental, papeles temáticos, correferencia, entre otros.

- Hopinion es un corpus del español coloquial constituido por opiniones de hoteles descargadas de la web de TripAdvisor. Hopinion está anotado con información morfosintáctica que ha sido revisada manualmente por un grupo de lingüistas.

El Cuadro 1 describe las características principales del subconjunto de datos utilizado. Por simplicidad, en adelante nos referiremos a ambos subconjuntos como corpus AnCora y corpus Hopinion.

Característica	Hopinion	AnCora-Es
Número de textos	1635	1635
Total palabras	206.812	443.380
Promedio de palabras por texto	126.49	271.18
Fuente de datos	TripAdvisor	El Periódico Agencia EFE
Registro asociado	Colloquial	Formal

Cuadro 1: Características de los corpus.

4 Aproximaciones

En esta sección se describen las dos aproximaciones contrastadas en nuestro estudio.

4.1 Aproximación basada en patrones morfosintácticos

Esta aproximación utiliza una serie de características morfosintácticas para la clasificación automática de los textos según su registro lingüístico. Las características que hemos seleccionado, once en total (ver Cuadro 2), representan cinco de las principales manifestaciones lingüísticas descritas en los estudios sobre el español coloquial:

- Sintaxis concatenada (SXC): consiste en la acumulación de enunciados producto de la ausencia de planificación en la producción del mensaje (Narbona, 1989).

- Elipsis (ELP): es la omisión de elementos lingüísticos que se presuponen a partir de entidades que se hallan presentes en el contexto discursivo. Una forma básica pero muy frecuente de representar tales omisiones en el lenguaje coloquial, es mediante el uso de los puntos suspensivos.

- Redundancia (RED): consiste en duplicar, de manera exacta o aproximada, algunas partes del discurso (Tannen, 1989).

- Deixis (DXS): es un recurso para la cohesión textual que el hablante utiliza para introducir las entidades o referentes del contexto situacional en el discurso.

- Riqueza léxica (RL): son diferentes métricas que se utilizan para conocer la competencia léxica de un hablante (Read, 2005).

N°	Característica	Manifest.
1	Densidad léxica	RL
2	Signos de puntuación	SXC
3	Co-occurencia de palabras	RED
4	Conjunciones coordinantes	SXC
5	Conjunciones subordinantes	SXC
6	Pronombres personales y demostrativos	DXS
7	Puntos suspensivos	ELP
8	Interjecciones	PAR
9	Repetición de vocales y consonantes	PAR
10	Oraciones consecutivas[1]	INT
11	Variación léxica (TTR)	RL

Cuadro 2: Características morfosintácticas evaluadas en los textos.

Para calcular las frecuencias de estas once características, los textos se han etiquetado con Part-Of-Speech y lema. La riqueza léxica se ha obtenido mediante el cálculo de la densidad (ver Ecuación 1) y la variación[2] (ver Ecuación 2) léxicas.

$$DL = \frac{palabras\ léxicas}{total\ palabras} \times 100 \quad (1)$$

$$VL = \frac{palabras\ diferentes}{total\ palabras} \times 100 \quad (2)$$

4.2 Aproximación basada en patrones léxicos

Para obtener el registro lingüístico de los textos, esta aproximación se basa en la detección de términos informales y de emoción conjuntamente con el cálculo de la riqueza léxica.

Por una parte, los términos con un uso informal se obtuvieron consultando las versiones en línea de Wikcionario[3] y TheFreeDictionary[4]. Las marcas de uso que se hicieron servir para reconocer dichos términos fueron: "Coloquial", "Despectivo", "Malsonante", "Familiar", "Informal", "Peyorativo" y "Vulgar". De otro lado, la detección de los términos de emoción se efectuó mediante el *Spanish Emotion Lexicon* (Sidorov et al., 2012), un recurso léxico creado de forma totalmente manual por investigadores del Instituto Politécnico Nacional de México.

En la Figura 1 tenemos dos ejemplos del formato (XML) y el tipo de información con que se ha anotado cada una de las palabras en los corpus Ancora y Hopinion a partir de los recursos señalados.

Conjuntamente con los atributos "registro" (*register*) y "emoción" (*emotion*) hemos usado nueve métricas de la riqueza léxica[5] (Roberto, Martí, y Salamó, 2012): densidad léxica (ver Ecuación 1), type token ratio (ver Ecuación 2), sofisticación léxica (ver Ecuación 3), perfíl de frecuencia léxica (ver Ecuación 4), a^2 (ver Ecuación 5), índice de Uber (ver Ecuación 6), Z de Zipf (ver Ecuación 7), variación de palabras léxicas (ver Ecuación 8) y variación modal (ver Ecuación 9).

[1]El patrón que usamos para identificar los esquemas oracionales consecutivos es: *[intensificador: tanto, tan, tal, etc.] + [nombre OR adjetivo OR adverbio] + [que]*

[2]Concretamente *type-token ratio*.

[3]http://es.wiktionary.org/

[4]http://es.thefreedictionary.com/

[5]Incluidas las dos de la aproximación morfosintáctica.

```
<wd lemma="cabrear"
    register="coloquial"
    emotion="enojo"
    source="ancora" />

<wd lemma="acojonante"
    register="coloquial"
    emotion="sorpresa"
    source="hopinion" />
```

Figura 1: Ejemplos de palabras anotadas con registro y emoción.

$$SL = \frac{N_{slex}}{N_{lex}} \quad (3)$$

$$PFL = \frac{T_s}{T} \quad (4)$$

$$a^2 = \frac{\log N - \log T}{\log^2 N} \quad (5)$$

$$IU = \frac{(\log N)^2}{logN - logT} \quad (6)$$

$$ZIPF = \frac{Z \times N \times \log(N/Z)}{(N - Z)\log(p \times Z)} \quad (7)$$

$$VPL = \frac{T_{lex}}{N_{lex}} \quad (8)$$

$$VM = \frac{T_a + T_r}{N_{lex}} \quad (9)$$

Note: N (tokens), T (types), lex (unidades léxicas), s (unidades sofisticadas), p (token más frecuente dividido por la longitud del texto) y Z (una medida de la riqueza léxica, en este caso $Z = TTR$).

Para obtener estos valores hemos usado la herramienta para el Análisis de Textos de Opinión en lenguaje natural (ATOp) (Queral, 2013). ATOp es una plataforma en Java que estamos desarrollando como parte de nuestro trabajo en minería de opiniones (ver Figura 2).

5 Evaluación y resultados

En esta sección presentamos los resultados obtenidos al aplicar las dos aproximaciones para la detección automática del registro lingüístico descritas en la sección anterior.

En los experimentos empleamos técnicas de aprendizaje supervisado el cual nos permite predecir la clase a la que pertenece un determinado objeto a partir de una serie de ejemplos de entrenamiento. En nuestro caso, el objetivo es predecir si un texto x pertenece a la clase formal (AnCora) o coloquial (Hopinion) basándonos en

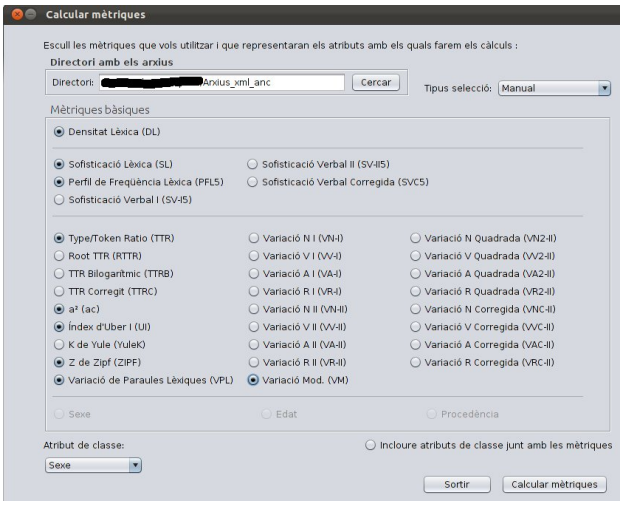

Figura 2: Selección de las métricas de riqueza léxica en ATOp.

las carecterísticas o atributos enumerados en la sección 4.

Como herramienta de análisis se ha usado Weka (Witten y Frank, 2000). De esta herramienta se han seleccionado 38 conocidos algoritmos de aprendizaje automático supervisado, los cuales se pueden dividir en 5 grandes categorías, ver Cuadro 3. En dicho cuadro, la primera columna describe la categoría y la segunda los algoritmos usados para la evaluación de esa categoría.

Todos los experimentos tienen la misma configuración. Se ha recopilado un fichero de datos donde se describen todos los textos en función de un conjunto de atributos, éstos se corresponden con las características descritas anteriormente (en la Sección 4) y como atributo de clase se ha definido el registro "coloquial" o "formal". La experimentación se ha realizado con validación cruzada[6] (*ten-fold cross-validation*). El resultado de los clasificadores se da en términos de su precisión (*Prediction Accuracy*), es decir, el porcentaje de instancias que fueron correctamente clasificadas.

Con el fin de determinar los atributos que tienen más peso, hemos aplicado varios métodos de selección de atributos. Los métodos de evaluación y de búsqueda usados en la selección supervisada se enumeran en el Cuadro 4. Los métodos de selección de atributos reducen el número de variables, seleccionando el mejor subconjunto de características del conjunto de inicial.

[6]La validación cruzada es una técnica para la evaluación experimental en la que los datos disponibles se dividen aleatoriamente en un conjunto de entrenamiento y un conjunto de test.

Categoría	Algoritmo de Aprendizaje
Bayes	BayesNet, BayesianLogisticRegression, ComplementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, 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, lmt.LogisticBase

Cuadro 3: Listado de los algoritmos utilizados. Como se puede observar Weka permite elegir entre múltiples algoritmos de clasificación, distribuidos en cinco categorías.

Métodos de evaluación
CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, LatentSemanticAnalysis, OneRAttributeEval, PrincipalComponents, ReliefFAttributeEval, SVMAttributeEval, WrapperSubsetEval
Métodos de búsqueda
BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, RandomSearch, RankSearch, Ranker, ScatterSearchV1, SubsetSizeForwardSelection

Cuadro 4: Métodos de selección de atributos usados por los clasificadores en la experimentación y divididos en dos características: el método de evaluación utilizado y el método de búsqueda.

5.1 Clasificación mediante patrones morfosintácticos

En este apartado describimos los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus Hopinion y Ancora, usando las once características morfosintácticas enumeradas en la sección 4.1.

Aplicando la configuración detallada al principio de esta sección entrenamos un total de 14.400 clasificadores[7]. En promedio, los clasificadores que usan algún método de selección de atributos funcionan mejor que los que no los usan (ver Figura 3).

[7]La combinación de algunos métodos de evaluación y de búsqueda no son posibles en Weka.

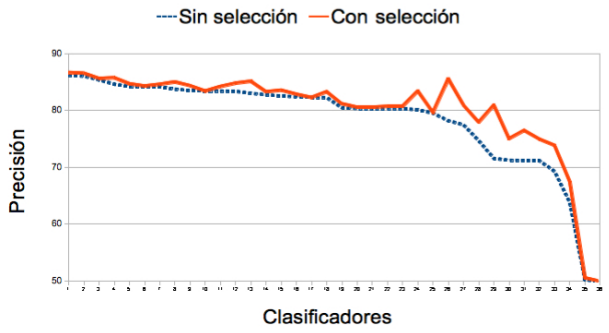

Figura 3: Precisión con y sin selección de atributos. Usando la selección de atributos se consigue reducir el número de características, mientras que la precisión mejora o se mantiene.

El clasificador que obtuvo un nivel de precisión más elevado (89 %) fue *trees.RandomForest* con *CfsSubsetEval* como método de evaluación y de búsqueda *Ranker*. Random Forests es una técnica de agregación que incorpora aleatoriedad en la construcción de cada clasificador para mejorar la precisión. El método *CfsSubsetEval* considera el valor predictivo individual de cada atributo seleccionando aquellos que estén altamente correlacionados con la clase y tengan entre ellos baja intercorrelación. *Ranker*, por su parte, devuelve una lista ordenada de los atributos según su calidad.

Para identificar los atributos con mayor valor predictivo, en el Cuadro 5 hemos contrastado el rendimiento promedio de los clasificadores cuando usan un determinado atributo (A_x) y cuando dejan de usarlo ($\neg A_x$). De esta manera, el valor *Diferencia* determinará el impacto que tiene la omisión del atributo x a nivel de la precisión.

	Promedios			
A	A_x	$\neg A_x$	*Diferencia*	
1	78.3 %	65.5 %	12.8	
2	78.5 %	76.1 %	2.4	
3	78.1 %	72.7 %	5.4	
4	78.6 %	62.2 %	16.4	•
5	78.6 %	76.3 %	2.3	
6	78.6 %	62.2 %	16.4	•
7	78.6 %	65.7 %	12.9	
8	78.5 %	71.3 %	7.2	
9	78.6 %	76.2 %	2.4	
10	78.6 %	76.1 %	2.5	
11	78.6 %	65.9 %	12.7	

Cuadro 5: Atributos morfosintácticos con mayor valor predictivo.

Los atributos más informativos son el 4 y el 6, es decir, conjunciones coordinantes y pronombres personales y demostrativos (DXS). Los menos informativos son el 2 y el 9 (signos de puntuación y repetición de vocales y consonantes).

Finalmente, para conocer el rendimiento de los clasificadores, en la Figura 4 relacionamos la precisión de cada algoritmo de clasificación con el número de atributos seleccionados. Como se observa, es posible obtener niveles de precisión adecuados con solo 6 atributos. De manera específica, la precisión media de los clasificadores que usan seis atributos es de 85.6 %, siendo *trees.LMT* (*Logistic Model Tree*) el clasificador que obtuvo el valor más alto: 85.9 % (también con seis atributos).

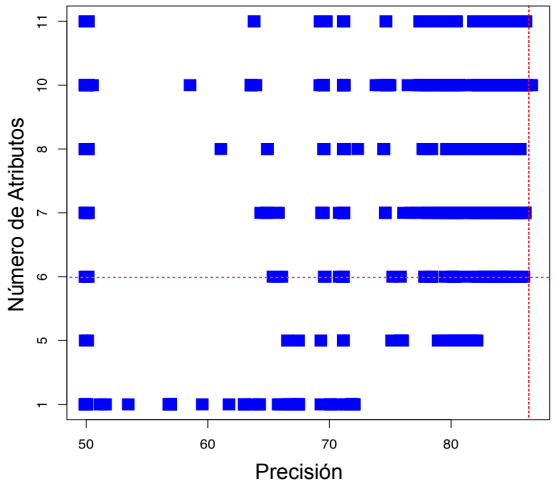

Figura 4: Rendimiento: número de atributos morfosintácticos versus precisión.

5.2 Clasificación mediante patrones léxicos

En este apartado describimos los niveles de precisión alcanzados al predecir el registro lingüístico de los textos en los corpus Hopinion y Ancora, usando las once características léxicas enumeradas en la sección 4.2.

En total entrenamos 13.768 clasificadores. A diferencia de lo que sucede con la clasificación mediante patrones morfositácticos, los clasificadores que usan algún método de selección de atributos y los que no los usan se comportan de manera similar. En ambos casos obtuvimos una precisión promedio del 84 %.

El clasificador que obtuvo un nivel de precisión más elevado (93.8 %) usa selección de atributos: *trees.RandomForest* con *ConsistencySubsetEval* como método de evaluación y de búsqueda *RankSearch*. El método *ConsistencySubsetEval* mide la consistencia de un subconjunto de atributos en términos de las clases. *RankSearch* ordena los atributos utilizando un evaluador individual o de conjuntos y crea un ranking de subconjuntos prometedores.

Debido a que nueve de los once atributos tiene que ver con la riqueza léxica y dos con el uso de términos coloquiales y de emoción, en la Figura 5 (ver Apéndice) presentamos la relación que hay entre estos tres tipos de atributos y la precisión. Como se puede observar, la mediana más alta la tienen los clasificadores que usan los tres tipos de atributos (RIQ+USO+EMO), luego están los que utilizan "riqueza" conjuntamente con "uso" (RIQ+USO), seguidos por los clasificadores que utilizan solo el atributo "emoción" (EMO) o "riqueza" (RIQ). Ningún clasificador emplea el atributo USO de forma aislada.

Adicionalmente, en la Figura 6 (ver Apéndice) presentamos los atributos que han seleccionado los 13.768 clasificadores como los más relevantes. En total tenemos seis atributos, el de emoción, el del uso coloquial y cuatro de riqueza léxica. En este último caso la densidad léxica (DL) ha dado mejores resultados.

Por último, en la Figura 7 (ver Apéndice) tenemos el número de atributos usados por los clasificadores para obtener los distintos niveles de precisión. Los clasificadores que tienen un mejor rendimiento usan siete atributos. Estos clasificadores presentan precisiones máximas superiores al 90 % y sus medianas están entre el 80 % y el 90 %.

6 Discusión

Nuestros experimentos indican que es posible usar el registro lingüístico para clasificar textos del español de manera fiable. El empleo de patrones léxicos ha sido más efectivo que la aproximación morfosintáctica, tanto a nivel de precisión como al número de atributos empleados. A la misma conclusión llegan (Sharoff, Zhili, y Katja, 2010) en su estudio sobre clasificación de textos por género para el inglés. Si bien, tal como se comenta en el mismo estudio, la eficiencia de los patrones léxicos puede estar relacionada con su capacidad para predecir el dominio antes que el género o el registro de los textos, éste no es nuestro caso ya que los rasgos empleados (riqueza léxica, términos coloquiales y de emoción) son independientes del dominio.

7 Conclusiones

En este artículo hemos contrastado dos aproximaciones para la clasificación del registro lingüístico en textos del español. La precisión de base obtenida con la aproximación morfosintáctica fue del 89 % (con diez atributos) y con la léxica del 93.8 % (con siete atributos).

Los atributos más informativos son, en la aproximación morfosintáctica, las conjunciones coordinantes, los pronombres personales y los demostrativos. En la aproximación léxica son la densidad léxica, los términos de emoción y los de uso coloquial. Este último siempre aparece correlacionado con otros atributos.

Agradecimientos

Este trabajo ha sido posible gracias los proyectos DIANA (DIscourse ANAlysis for knowledge understanding, TIN2012-38603) y TIN2009-14404-CO2 del Ministerio de Ciencia e Innovación, así como a la beca FI 2010FI_B 00521 de la Generalitat de Catalunya.

Bibliografía

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK.

Boese, Elizabeth S. y Adele E. Howe. 2005. Effects of web document evolution on genre classification. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, páginas 632–639, New York, NY, USA. ACM.

Brooke, Julian, Tong Wang, y Graeme Hirst. 2010. Automatic acquisition of lexical formality. En *In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING*.

Francis, W. y H. Kucera. 1979. Brown corpus. Text, Department of Linguistics, Brown University.

Gries, Stefan, John Newman, y Cyrus Shaoul. 2009. N-grams and the clustering of genres. *ELR Journal*, 5:1–13.

Kanaris, Ioannis y Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. En *In Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence, v.2*, páginas 3–10.

Lahiri, Shibamouli, Prasenjit Mitra, y Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. En *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, páginas 446–457, Berlin, Heidelberg. Springer-Verlag.

Mason, J., M. Shepherd, y J. Duffy. 2009. An n-gram based approach to automatically identifying web page genre. En *Proceedings of the 42nd Hawaii International Conference on System Sciences*, HICSS '09, páginas 1–10, Washington, DC, USA. IEEE Computer Society.

Mosquera, A. y P. Moreda. 2011. Caracterización de niveles de informalidad en textos de la web 2.0. *Procesamiento del Lenguaje Natural*, 47:171–177.

Narbona, Antonio. 1989. *Sintaxis española: nuevos y viejos enfoques*. Ariel.

zu Eissen, Sven Meyer y Benno Stein. 2004. Genre classification of web pages: User study and feasibility analysis. En *IN: BIUNDO S., FRUHWIRTH T., PALM G. (EDS.): ADVANCES IN ARTIFICIAL INTELLIGENCE*, páginas 256–269. Springer.

Queral, Bakary Singateh. 2013. Plataforma en java per l'anàlisi de textos d'opinió en llenguatge natural (atop). Master's thesis, Universitat de Barcelona. Facultat de Matemàtiques.

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

Ricci y Wietsma. 2006. Product reviews in travel decision making. *Proceeding of Information and Communication Technologies in Tourism (ENTER)*, páginas 296–307.

Roberto, J., M. Martí, y M. Salamó. 2012. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, 1(48):97–104.

Santini, Marina. 2007. *Automatic Identification of Genre in Web Pages*. Ph.D. tesis, University of Brighton.

Sharoff, Serge, Wu Zhili, y Markert Katja. 2010. The web library of babel: evaluating genre collections. *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC10)*, páginas 3063–3070.

Sheikha, Abu y Diana Inkpen. 2010. Automatic classification of documents by formality. En *International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, páginas 1–5.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, y J. Gordon. 2012. Empirical study of opinion mining in spanish tweets. *LNAI*, páginas 7629–7630.

Tannen, D. 1989. Repetition in conversation: Towards a poetics of talk. En D. Tannen, editor, *Talking Voices. Repetition, dialogue and imagery in conversational discourse*. Cambridge, CUP.

Tribble, Christopher. 1999. *Writing Difficult Texts*. Ph.D. tesis, Lancaster University.

Witten, I. y E. Frank. 2000. *DataMining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers.

Xiao, Zhonghua y Anthony McEnery. 2005. Two approaches to genre analysis. three genres in modern american english. *Journal of English Linguistics*, 33(3):62–82.

Apéndice

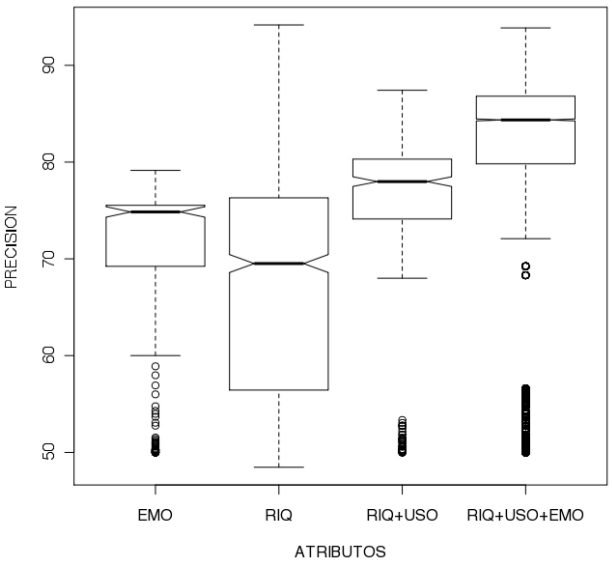

Figura 5: Precisión en relación con los tipos de atributos léxicos "emoción", "riqueza" y "uso".

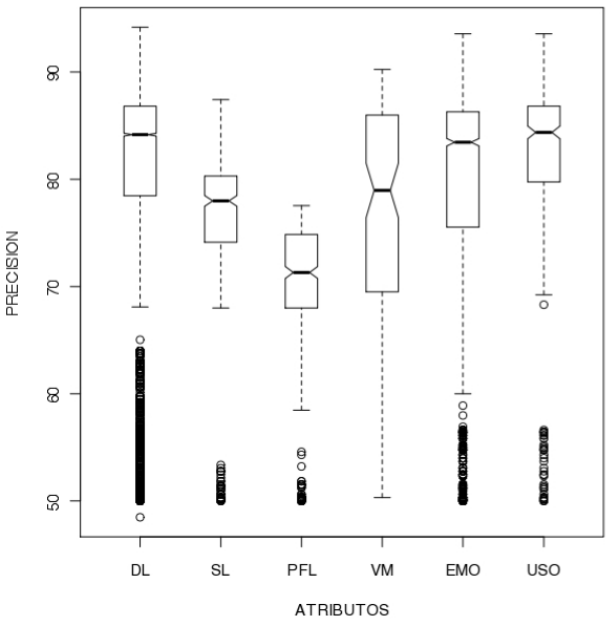

Figura 6: Mejores atributos léxicos seleccionados por los clasificadores.

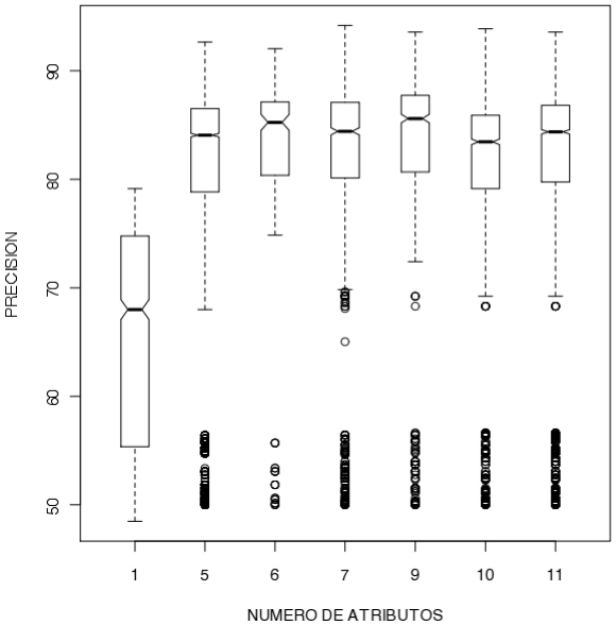

Figura 7: Precisión en relación con el número de atributos léxicos utilizados por los clasificadores.