Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 171-177 recibido 03-05-2011 aceptado 20-05-2011

Caracterización de Niveles de Informalidad en Textos de la Web 2.0*

Informality Level Caracterization in Web 2.0 Texts

Alejandro Mosquera
DLSI-Universidad de Alicante
Alicante
amosquera@dlsi.ua.es

Paloma Moreda
DLSI-Universidad de Alicante
Alicante
moreda@dlsi.ua.es

Resumen: El análisis de textos de la Web 2.0 es un tema de investigación relevante hoy en día. Sin embargo, son muchos los problemas que se plantean a la hora de utilizar las herramientas actuales en este tipo de textos. Para ser capaces de medir estas dificultades primero necesitamos conocer los diferentes registros o grados de informalidad que podemos encontrar. Por ello, en este trabajo intentaremos caracterizar niveles de informalidad para textos en inglés en la Web 2.0 mediante técnicas de aprendizaje automático no supervisado, obteniendo resultados del 68 % en F1.
Palabras clave: Clustering, Registros del Lenguaje, Web 2.0

Abstract: Analysis of Web 2.0 texts is a relevant investigation topic nowadays. However, many problems arise when using state of the art tools in this kind of texts. For being able to measure these difficulties first we need to identify the different registers or informality levels that we can find. Therefore, in this paper we will attempt to characterize the informality levels of english texts in Web 2.0 by using non-supervised machine learning techniques, obtaining results of 68 % in F1.
Keywords: Clustering, Language Registers, Web 2.0

1. Introducción

A día de hoy la popularidad de la Web 2.0 ha supuesto un incremento del uso de aplicaciones colaborativas en las cuales los usuarios se expresan, comunican y comparten información. El lenguaje utilizado en estas aplicaciones da lugar a la aparición de nuevos registros, presentando ciertas características que suponen un reto para el procesamiento del lenguaje natural.

Halliday en (Halliday y Ghadessy, 1988) define los registros del lenguaje como: "...un clúster de características asociadas teniendo una tendencia a ocurrir mayor que el azar...". En periódicos online, webs oficiales o artículos académicos se emplea el registro formal, mientras que en los blogs, chats, foros y redes sociales el registro del lenguaje que suele imperar es el conversacional, altamente contextualizado y con su propio vocabulario o "slang" (Squires, 2010).

Teniendo en cuenta estas características se considera necesario probar y medir el comportamiento de las herramientas de procesamiento de lenguaje natural (PLN) actuales en este tipo de textos. Para ello lo primero es identificar y conocer qué registros son los que se pueden presentar. Los estudios realizados hasta el momento se centran en el análisis multidimensional de la variación de registros y tipos de texto (Biber y Kurjian., 2007).

Sin embargo, en la web colaborativa todos los textos podrían considerarse informales. En este trabajo se plantea medir el grado o nivel de informalidad de este tipo de textos mediante la agrupación o "clustering" de características relevantes de los mismos. Por esta razón, en este artículo describimos nuestra propuesta empleando técnicas de aprendizaje automático no supervisado.

En el apartado 2 revisamos el estado del

* Este artículo ha sido cofinanciado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria d'Educación de la Generalitat Valenciana (proyectos PROMETEO/2009/119, ACOMP/2010/286 y ACOMP/2011/001)

arte del análisis de registros desde diferentes aproximaciones. En el apartado 3 explicamos como detectar grados de informalidad en textos. Y finalmente en el apartado 4 ofrecemos las conclusiones y los trabajos futuros.

2. Estado del Arte

Se podrían identificar diferentes líneas a la hora de afrontar este tipo de trabajos, tradicionalmente los estudios se han basado en reglas o análisis de patrones lingüísticos construídos de forma más o menos manual, mientras que los más actuales emplean técnicas de aprendizaje automático.

Respecto a las aproximaciones basadas en análisis de patrones destaca la metodología del análisis multidimensional (MDA) (Biber, 1988), siendo una de las contribuciones más relevantes al estudio de la variación de los registros del lenguaje. A lo largo de su estudio son identificados 23 registros correspondientes a variedades del Inglés oral y escrito, obteniendo la frecuencia de un conjunto finito de características lingüísticas relevantes para el lenguaje. Mediante técnicas estadísticas de análisis de factores identifica las dimensiones o factores, que coincidirán con las características que aparecen en textos con mayor frecuencia. Las características lingüísticas variarán en su distribución a través de los registros, posibilitando la identificación de grupos de textos. Esta metodología ha servido de base para diversos trabajos posteriores enfocados en el análisis del género, tipos de texto u otros registros más especializados para diversos lenguajes (Biber, 2003).

En otra aproximación mediante patrones (Tribble, 1999), se aplica el uso de palabras clave para obtener resultados similares a los obtenidos con MDA pero de una forma menos compleja. Tribble emplea el software de exploración de corpus WordSmith (Scott, 1999) para generar una lista de palabras y extraer las palabras clave de forma automática. La frecuencia de aparición de dichas palabras clave se compara con la de un corpus de referencia. De esta forma obtenemos las palabras más relevantes que se repiten dentro de un tipo de texto o registro, permitiendo caracterizar grupos de textos.

Respecto a las aproximaciones basadas en técnicas de aprendizaje automático hay que destacar que son pocos los trabajos que se han realizado hasta la fecha, si bien se pueden encontrar tanto aproximaciones supervisadas como no supervisadas.

Dentro de las supervisadas destaca la de (Sharoff, Wu, y Markert, 2010). Aunque no se centra estrictamente en la clasificación de registros sino de géneros, ambos términos están estrechamente relacionados, ya que hasta (Biber, 1995) se empleaban los términos registro y género de forma indistinta. Este trabajo es relevante por emplear un modelo basado en trigramas de etiquetas POS, desarrollando un clasificador para la detección del género de textos de Internet con diversos grados de precisión. Para ello utiliza el algoritmo de aprendizaje supervisado basado en máquinas de vectores de soporte ("support vector machines", SVM) (Cortes y Vapnik, 1995) y las anotaciones del corpus Brown (Francis y Kucera, 1979) para el idioma inglés.

Sobre los trabajos basados en aprendizaje automático no supervisado encontrados, en (Gries, Newman, y Shaoul, 2009) se utiliza un algoritmo de clústering aglomerativo jerárquico para diferenciar registros del lenguaje empleando n-gramas. Los resultados se obtienen en base a los corpus BNC-Baby e ICE-GB, analizando posteriormente la longitud óptima de los n-gramas aplicados a subregistros del lenguaje.

La propuesta de MDA es la más avanzada hasta la fecha, pero requiere laboriosas anotaciones manuales. La aproximación por palabras clave es significativamente mas sencilla aunque no permite clasificar con un granulado tan fino y necesita además un corpus de referencia. Por otra parte, el trabajo de Gries, al estar basado en n-gramas, es muy sensible a cambios en los datos. Por otra parte, los registros obtenidos no nos proporcionan información gradual sobre niveles de informalidad.

Nuestra propuesta es novedosa en el aspecto de detectar grados de informalidad en textos de la Web 2.0, ya que los textos de la web colaborativa, por su naturaleza hacen difícil una distinción binaria clara entre registro formal o informal, predominando la segunda respecto a la primera.

3. Detección de Grados de Informalidad

A la hora de analizar textos de la Web 2.0 nos encontramos con la problemática de estar trabajando con un lenguaje escrito con características muy heterogéneas, extremadamente variables y sin contar con información so-

bre la distribución de sus registros lingüísticos.

El objetivo principal de este estudio es clasificar textos atendiendo a grados de informalidad mediante técnicas de aprendizaje automático no supervisado, con el fin de que se disponga de grupos de registros sobre los cuales poder evaluar la efectividad de las herramientas de PLN actuales en los diferentes tipos de textos que ofrece la Web 2.0.

El el apartado 1 detallamos las características empleadas en la clasificación de los textos. En el apartado 2 explicamos el proceso de análisis. En el apartado 3 revisamos los corpus utilizados para el estudio y sus principales orígenes de datos. Finalmente en el apartado 4 comentamos los resultados obtenidos.

3.1. Información Utilizada para la Clasificación

Se ha definido un conjunto de 19 características léxico-gramaticales (ver Cuadro 1) las cuales nos van a permitir caracterizar de forma significativa rasgos de los textos para extraer información sobre su nivel de informalidad. Estas características son:

- (1) frecuencia media de palabras por oración.
- (2) tamaño medio en caracteres de las palabras del texto.
- (3) tamaño medio en caracteres de la oración.
- (4) número de caracteres no imprimibles respecto al texto.
- (5) frecuencia de preposiciones respecto al texto.
- (6) frecuencia de adverbios respecto al texto.
- (7) frecuencia de verbos respecto al texto.
- (8) frecuencia de verbos en pasivo respecto al texto.
- (9) frecuencia de adjetivos respecto al texto.
- (10) frecuencia de determinantes respecto al texto.
- (11) frecuencia de sustantivos respecto al texto.
- (12) frecuencia de números respecto al texto.
- (13) frecuencia de emoticonos respecto al texto.
- (14) frecuencia de pronombres respecto al texto.
- (15) frecuencia de interjecciones respecto al texto.
- (16) frecuencia de palabras en mayúscula respecto al texto.
- (17) frecuencia de palabras informales respecto al texto.
- (18) frecuencia de palabras desconocidas por el diccionario respecto al texto.
- (19) frecuencia de palabras neutrales respecto al texto.

Nº	Característica
1	WordsPerSentence
2	WordLength
3	SentenceLength
4	NonPrintable
5	Prepositions
6	Adverbs
7	Verbs
8	PasiveVerbs
9	Adjetives
10	Determinants
11	Nouns
12	Numbers
13	Emoticons
14	Pronouns
15	Interjections
16	UppercaseWords
17	InformalWords
18	UnknowWords
19	NeutralWords

Cuadro 1: Conjunto de Características Evaluadas en los Textos

Para calcular las frecuencias de las características gramaticales se emplean etiquetadores (POS) como TreeTagger (Schmid, 1994) y Freeling (Atserias et al., 2006). El criterio de neutralidad/informalidad de las palabras se ha establecido mediante diccionarios que proporcionan dicha información, buscando etiquetas clave con connotación informal: "Informal", "Colloquial", "Offensive", "Vulgar", "Slang" y "Onomatopoeia", mediante la consulta a las versiones en línea de Wiktionary (Foundation, 2011) y TheFreeDictionary (Farlex, 2011).

3.2. Clasificación

Con el fin de llevar a cabo el proceso de aprendizaje no supervisado se emplea un algoritmo de "hard-clustering" como K-Means (Hartigan y Wong, 1979). Se trata de un algoritmo clasificado como Método de Particionado y Recolocación, su denominación procede

de la representación de los clústeres por la media de sus puntos (centroide), pudiéndose aplicar únicamente en atributos numéricos. Para calcular la distancia de un elemento a su centroide mas próximo se ha utilizado la función de distancia euclídea :

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2} \qquad (1)$$

La representación mediante centroides permite una interpretación gráfica y estadística inmediata, siendo necesario proporcionar con antelación el número de k-particiones o clústeres al algoritmo. Para esta primera aproximación se ha tomado como el número inicial de clústeres K el valor óptimo obtenido mediante el algoritmo de X-Means: K=2 (Pelleg y Moore, 2000).

3.3. Datos Utilizados

Las pruebas se han realizado sobre un subconjunto de los corpus de la Fundación Barcelona Media (FBM, 2009) con la distribución mostrada en el Cuadro 2, constando en total de 7 diferentes orígenes de datos de la Web 2.0: Ciao, Kongregate, Slashdot, Digg, MySpace, Twitter y Engadget. Mientras que todos tienen en común su naturaleza colaborativa, cada cual muestra caracteristicas propias.

Ciao es una web de opiniones y comparación de precios donde los consumidores comentan las características de los productos que han adquirido. Kongregate es una web de juegos en línea que permite la comunicación entre jugadores mediante un chat en tiempo real. Slashdot es un portal de discusión de noticias relacionadas con la tecnología. Digg es un sitio web de noticias de ciencia y tecnología principalmente, donde los usuarios pueden realizar comentarios. MySpace es una popular red social que contiene foros de discusión sobre diferentes temáticas. Twiter es una red social de micro-blogging, donde los usuarios se expresan mediante mensajes cortos. Engadget es un blog y podcast de productos de electrónica y tecnología, donde los usuarios pueden comentar los productos analizados.

Los orígenes de datos expuestos anteriormente en las diferentes comunidades muestran bastante heterogeneidad a 2 niveles cuantitativos, ya que si bien los textos de Kongregate o Twitter siguen el estilo informal propio de los chat con palabras cortas, pocas oraciones y uso de slang, expresiones coloquiales y frecuentes errores o desviaciones gramaticales, en comparación Slashdot o Ciao presentan textos de mayor tamaño, narrativa más elaborada y mayor grado de formalidad.

Origen	**Núm. documentos**
Slashdot	50
Ciao	50
Kongregate	50
Digg	50
MySpace	50
Twitter	50
Engadget	50
Total	**350**

Cuadro 2: Distribución del subconjunto de textos del corpus Caw de FBM

3.4. Evaluación y Resultados

Analizando los 2 clústeres generados por el algoritmo K-Means (ver Cuadro 3) se puede observar su correspondencia con 2 grados de informalidad:

- El Cluster0, cuyas características más significativas son palabras de menor longitud, poca densidad de palabras por oración, gran número de interjecciones, presencia de emoticonos, un número significativo de palabras en mayúsculas y más de un 17 % de palabras informales, tiene las características propias de un registro muy informal.
- El Cluster1, cuyas características más significativas serían frases y palabras de mayor longitud, mayor número de verbos en forma pasiva, así como una mayor frecuencia en general de elementos gramaticales detectados, corresponderían a un registro menos informal del lenguaje.

La evaluación de los clústeres obtenidos mediante K-Means se ha realizado analizando y anotando de forma manual todas las muestras de texto con un valor binario "informal" en caso de encontrar características del registro informal o "neutro" en caso de no encontrar ningún rasgo del registro informal. Las etiquetas de la clasificación corresponden a cada una de las K particiones definidas inicialmente por X-Means. Para el resto de los casos de K sería necesario definir etiquetas

Característica	Cluster0	Cluster1
WordsPerSentence	6.4739	15.8169
WordLength	3.9714	4.325
NonPrintable	0.2064	0.1066
Prepositions	0.0349	0.0966
Adverbs	0.0347	0.0845
Verbs	0.0672	0.1664
PasiveVerbs	0.0002	0.0014
Adjetives	0.0669	0.0843
Determinants	0.0258	0.0856
Nouns	0.4240	0.2232
Emoticons	0.0183	0.0003
Pronouns	0.0277	0.0878
Interjections	0.0737	0.0014
UppercaseWords	0.0281	0.0107
InformalWords	0.1769	0.0364
UnknowWords	0.0231	0.0059
Total instancias	**70**	**280**

Cuadro 3: Características agrupadas en clústeres

adicionales para los correspondientes grados de informalidad.

Tomando como G el conjunto de los documentos anotados manualmente, T el número total de documentos y k el número de clústeres generados por K-Means, se ha calculado la precisión y la cobertura media ponderada por cada clúster por una medida representativa en este tipo de algoritmos no supervisados (Andritsos et al., 2003):

$$P = \sum_{i=1}^{k} \frac{|Gi|}{|T|} Pi \qquad R = \sum_{i=1}^{k} \frac{|Gi|}{|T|} Ri \qquad (2)$$

Como objeto de tener un único valor que permita comparar los experimentos se ha calculado la medida de F1, que combina la precisión y la cobertura.

$$F1 = 2\frac{PR}{P+R} \qquad (3)$$

Evaluando nuestra clasificación realizada con K-Means nuestras medidas indican buenos resultados para ser una primera aproximación (ver Cuadro 4).

Mediante la aplicación del algoritmo de selección de características por correlación CFS (Hall, 1998) se han extraído las características más relevantes (ver Cuadro 5). Este algoritmo evalúa un subconjunto de atributos

Precisión	Cobertura	F1
0.678	0.657	0.667

Cuadro 4: Resultados empleando Treetagger y Farlex

considerando la habilidad individual de predicción de cada variable, así como el grado de redundancia entre ellas, prefiriéndose los subconjuntos de características que estén altamente correlacionados con la clase y tengan baja intercorrelación.

Nº	Característica
3	SentenceLength
4	NonPrintable
5	Prepositions
11	Nouns
13	Emoticons
16	UppercaseWords
17	InformalWords

Cuadro 5: Subconjunto de características más relevantes obtenidas mediante CFS

Evaluando los resultados obtenidos empleando únicamente los atributos seleccionados por CFS podemos observar que nuestras medidas se han mejorado en un 2 % (ver Cuadro 6).

Precisión	Cobertura	F1
0.691	0.674	0.683

Cuadro 6: Resultados empleando Treetagger y Farlex con el subconjunto definido por CFS

Existen diferencias significativas entre los resultados de los etiquetadores POS empleados, TreeTagger detectó y clasificó un 32 % más determinantes, un 32 % más sustantivos y 93 % más pronombres (ver Cuadro 7), mientras que empleando Freeling obtenemos resultados ligeramente inferiores (ver Cuadro 8).

Respecto a las consultas en línea, el diccionario de Farlex ha sido capaz de catalogar un 43 % más palabras informales que Wiktionary (ver Cuadro 9), ya que internamente es capaz de distinguir entre mayúsculas y minúsculas, además de contar con un diccionario de acrónimos. Por lo que los resultados obtenidos con Wiktionary con ligeramente in-

Característica	TreeTagger	Freeling
Prepositions	0.0842	0.0835
Adverbs	0.0746	0.0648
Verbs	0.1466	0.1565
Adjetives	0.0808	0.0700
Determinants	0.0736	0.0500
Nouns	0.2634	0.1780
Pronouns	0.0758	0.0048
Interjections	0.0159	0.0164

Cuadro 7: Comparación entre frecuencias de características obtenidas mediante los etiquetadores POS

Precisión	Cobertura	F1
0.617	0.620	0.618

Cuadro 8: Resultados empleando Freeling

feriores (ver Cuadro 10).

Característica	Wiktionary	Farlex
InformalWords	0.0368	0.0645
UnknowWords	0.0094	0.0094
NeutralWords	0.8828	0.8551

Cuadro 9: Comparación de frecuencias de características obtenidas mediante diccionarios online

Por último, desde un punto de vista mas lingüístico, podemos tener en cuenta la relación expuesta entre elementos formales e informales de (Heylighen y Dewaele, 1999) por la cual los elementos que forman parte de la categoría formal, no deíctica de las palabras, disminuyen su frecuencia a la par que incrementa la informalidad del texto, incluyendo en esta categoría los nombres, adjetivos, preposiciones y artículos. Mientras que los que forman parte de la categoría deíctica, aquellas que se espera que su frecuencia aumente a la par que se incrementa la informalidad del texto, está formada por los pronombres, verbos, adverbios e interjecciones. Analizando nuestros resultados, vemos que dicha relación no se cumple con excepción de las interjecciones, cuyo número es significativamente superior en el Cluster0. Dicha circunstancia se debe a las características propias del lenguaje de los textos de la Web 2.0, demostrando la necesidad de estudiar los grados de

Precisión	Cobertura	F1
0.664	0.651	0.657

Cuadro 10: Resultados empleando Wiktionary

informalidad existentes y adaptar las herramientas de PLN para poder trabajar con dichos textos.

4. Conclusiones y Trabajos Futuros

En este trabajo se han identificado niveles o grados de informalidad de textos de la Web 2.0 de forma que se pueda medir la utilidad de las actuales herramientas de PLN en los diferentes grupos de textos que nos podemos encontrar.

Los resultados obtenidos (68 % en F1) muestran que las técnicas de aprendizaje automático no supervisado pueden emplearse como una forma de determinar los niveles de informalidad. Esta primera aproximación es necesaria para poder aplicar posteriormente otras técnicas de análisis o extraer las características que definen los grados o niveles de informalidad en los textos de la web colaborativa.

En el análisis de los resultados se han detectado necesidades que vamos a intentar solucionar en trabajos futuros. Entre ellas se podría destacar la debilidad de los algoritmos de clustering contra el ruido y la variabilidad de los datos, sobre todo en corpus muy grandes, afectando a la calidad y número de particiones obtenidas. La exploración de otros algoritmos de clasificación no supervisada, identificar grados de informalidad adicionales, así como añadir nuevas características tales como el uso de "bundles" (Biber y Cortes., 2004) o construcciones no detectables mediante diccionarios, podría añadir información relevante y mejorar los resultados obtenidos.

Bibliografía

Andritsos, Periklis, Panayiotis Tsaparas, Renee J. Miller, y Kenneth C. Sevcik. 2003. Limbo: A scalable algorithm to cluster categorical data. Informe técnico, University of Toronto, Department of Computer Science.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluis Padró, y Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, páginas 48–55.

Biber, D. 1988. *Linguistic features: algorithms and functions in Variation across speech and writing*. Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. New York: Cambridge University Pres Linguistics.

Biber, D. 2003. Variation among university spoken and written registers: A new multidimensional analysis. *Language and Computers*, 46:47–70.

Biber, D. y J. Kurjian. 2007. Towards a taxonomy of web registers and text types: A multi-dimensional analysis. En N. Nesselhauf In M. Hundt y C. Biewer, editores, *Corpus linguistics and the web*. Amsterdam, Rodopi, páginas 109–132.

Biber, Douglas, Susan Conrad y Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25:371–405.

Cortes, Corinna y Vladimir Vapnik. 1995. Support-vector networks. En *Machine Learning*, volumen 20, páginas 273–297.

Farlex, Inc. 2011. The free dictionary: http://www.thefreedictionary.com/.

FBM, Fundacion Barcelona Media. 2009. Caw 2.0 training datasets.

Foundation, Wikimedia. 2011. Wiktionary: The free dictionary. http://en.wiktionary.org/.

Francis, W. N. y H. Kucera. 1979. Brown corpus. Informe técnico, Department of Linguistics, Brown University, Providence, Rhode Island, US.

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

Hall, M A. 1998. Correlation-based feature selection for machine learning. *PhD dissertation Hamilton NZ Waikato University Department of Computer Science*.

Halliday, M.A.K. y Mohsen Ghadessy. 1988. *On the language of physical science. In Mohsen Ghadessy (ed.), Registers of Written English: situational factors and linguistic features*. London and New York: Pinter Publishers. 162-178.

Hartigan, J. A. y M. A. Wong. 1979. A K-means clustering algorithm. *Applied Statistics*, 28:100–108.

Heylighen, Francis y Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Informe técnico, Free University of Brussels.

Pelleg, Dan y Andrew W. Moore. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. En *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, páginas 727–734, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of the International Conference on New Methods in Language Processing*, páginas 44–49.

Scott, M. 1999. Wordsmith tools version 3.

Sharoff, Serge, Zhili Wu, y Katja Markert. 2010. The web library of babel: evaluating genre collections. En *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 3063–3070.

Squires, L. 2010. Enregistering internet language. *Language in Society*, 39(04):457–492.

Tribble, Christopher. 1999. Writing difficult texts. *Ph.D. dissertation. Lancaster University*.