Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 235-243 recibido 1-05-2009; aceptado 5-06-2009

Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán*

Features and affective traits of humor: A study of automatic humour recognition in Catalan texts

Antonio Reyes y **Paolo Rosso**
Universidad Politécnica de Valencia
Camino de Vera s/n. 46022
Valencia, España
{areyes,prosso}@dsic.upv.es

Antònia Martí y **Mariona Taulé**
Universidad de Barcelona
Gran Via, 585 - 08007
Barcelona, España
{amarti,mtaule}@ub.edu

Resumen: Las nuevas tendencias de investigación en Procesamiento del Lenguaje Natural (PLN) cada vez dan mayor importancia al análisis de fenómenos relacionados con los procesos cognitivos que se proyectan a través del lenguaje. El estudio de los sentimientos, las emociones o el humor son un reflejo de esta tendencia. En esta investigación se muestran los resultados relativos a un estudio acerca del Reconocimiento Automático del Humor (RAH) realizado sobre un corpus de textos humorísticos de escolares en catalán. Los resultados señalan que la identificación de características semánticas y afectivas permite la clasificación de los textos con un porcentaje considerable de acierto.
Palabras clave: Humor, Afectividad, Sentimientos, Reconocimiento, Clasificación

Abstract: The analysis of phenomena related to cognitive processes is a very important trend in Natural Language Processing (NLP) research. The study of sentiments, emotions or humour, through language, are a sample about how this tendency acquires a greater importance in the area. In this paper, we present the results obtained on a study of Automatic Humour Recognition (AHR) performed on a corpus of children's texts. The results indicate that through the identification of semantic and affective features the text classification can be achieved with success.
Keywords: Humour, Affective, Sentiments, Recognition, Classification

1. Introducción

La masificación de la tecnología requiere la creación de recursos y herramientas más afines a la forma de pensar de los usuarios. Esto implica la incorporación de nuevos paradigmas de investigación así como la adaptación de los actuales a las nuevas exigencias. Tal es el caso de las recientes tendencias de investigación en PLN, donde el estudio de fenómenos más vinculados con los procesos de cognición adquieren mayor importancia. Análisis de sentimientos, minado de opinión, generación y reconocimiento automático del humor (GAH, RAH) son ejemplos de cómo la investigación se dirige hacia la exploración de esferas más abstractas del comportamiento humano que adquieren representación en el plano lingüístico.

En este marco, el presente trabajo se centra en el estudio de un conjunto de textos en catalán producidos por escolares de entre 6 y 16 años con el fin de identificar las características que lo definen como humorístico. En este sentido, se trata de dar argumentos que permitan reconocer, mediante técnicas de aprendizaje automático, un fenómeno cognitivo de carácter subjetivo, pero que a la vez es social, cultural y lingüístico: el humor.

El artículo está organizado de la siguiente forma. En la Sección 2 se describen algunos trabajos relacionados con la GAH y el RAH. En la Sección 3 se plantea nuestra hipótesis

* Queremos dar las gracias al proyecto CesCa (2008ARIE-00053) por permitirnos experimentar con su corpus, así como a los subproyectos MiDEs y Lang2World del proyecto TEXT-MESS (TIN2006-15265-C06) y TeLMoSis (UPV PAID08-3294) por financiar parcialmente este trabajo.

 © 2009 Sociedad Española para el Procesamiento del Lenguaje Natural

en el ámbito de RAH. El desarrollo de los experimentos y los resultados se presentan en la Sección 4. La evaluación del conjunto de características así como la discusión de los resultados obtenidos se tratan en la Sección 5. Finalmente, en la Sección 6 se presentan las conclusiones y se esboza el trabajo futuro.

2. Estado del arte

En años recientes la investigación relativa al humor computacional se ha realizado básicamente desde dos perspectivas, la generación y el reconocimiento. En relación con la generación (GAH) se pueden citar los trabajos de Binsted (1996) y Binsted y Ritchie (1997), cuyo objetivo era generar de forma automática textos humorísticos basados en características lingüísticas de base fonética y semántica (véase ejemplo 1). También se destaca el proyecto realizado por Stock y Strapparava (2005) donde, basándose en la incongruencia y los conceptos opuestos, se pretendía generar nuevos sentidos humorísticos para una serie de acrónimos (véase ejemplo 2).

(1) What's the difference between leaves and a car? One you b**rush** and **rake**, the other you **rush** and b**rake**.

(2) **F**ederal **B**ureau of **I**nvestigation **F**antastic **B**ureau of **I**ntimidation

En cuanto al reconocimiento (RAH) destacan los trabajos de Mihalcea y Strapparava (2006a) y (2006b), cuyo objetivo es la identificación de las características desde una perspectiva que abarca la fonética, la sintaxis y la semántica, así como la ambigüedad y la ironía, que permiten reconocer un determinado texto como humorístico. Asimismo, la investigación de Sjobergh y Araki (2007) señala que rasgos como la ambigüedad, el solapamiento de palabras o los modismos, son características que definen el humor. Por otro lado, Buscaldi y Rosso (2007) señalan que por medio de bolsa de palabras, cálculo de n-gramas o la longitud de una oración, es posible discriminar entre datos humorísticos y no humorísticos. Si bien Mihalcea y Pulman (2007) trabajan con artículos humorísticos, la clase de estructuras que analizan los trabajos arriba mencionados para obtener dichas características son los denominados one-liners (OLs), es decir textos cortos con una estructura sintáctica simple y que producen el efecto humorístico de manera instantánea. En 3 y 4 se ejemplifican esta clase de textos.

(3) Jesus saves, and at today's prices, that's a miracle!

(4) Tu credi che ti voglia sposare per i tuoi otto milioni di dote? Come ti sbagli! Se tu ne avessi nove milioni ti sposerei lo stesso[1].

3. Reconocimiento Automático del Humor

El humor es un objeto de estudio multidimensional y, como tal, no es posible definirlo como una suma de variables ni desde una sola perspectiva. Además de constituir un complejo proceso cognitivo, en el humor también interviene información de tipo cultural, social, de competencia lingüística, etc. Es por ello que hasta la fecha no hay aún una teoría que satisfaga completamente la pregunta de cómo funciona el humor ni que establezca cuáles son los mecanismos que subyacen a todo acto humorístico (Ritchie, 2003).

El reconocimiento automático del humor exige un planteamiento que integre toda esta problemática, lo que supone serios dilemas: i) ¿cómo identificar algo que en su naturaleza es abstracto y subjetivo?, ii) ¿qué rasgos se deben tener en cuenta para discriminar algo chistoso de algo serio?, iii) ¿con qué parámetros se determina que algo es humorístico? A este respecto, los trabajos en RAH han demostrado que es viable responder en parte a estas preguntas mediante las técnicas y herramientas disponibles[2]. No obstante, la tarea de describir el fenómeno dista mucho de ser 100 % satisfactoria. En este sentido, en este trabajo se prueban algunas de las técnicas que han servido en la tarea de RAH sobre un tipo de estructuras humorísticas distintas: chistes de escolares.

El objetivo consiste en identificar las características que definen este conjunto de

[1] En el Anexo 1 se incluyen las traducciones o claves para la interpretación de los ejemplos que no aparecen en inglés.

[2] Si bien es cierto que con el tipo de datos con los que se ha trabajado es posible identificar determinadas características y variables más o menos constantes, sería arriesgado señalar que éstas son generalizables a cualquier tipo de humor. Es por ello que, para los efectos de la investigación, sólo se hace referencia al humor verbal, es decir, aquél que se expresa lingüísticamente (Attardo, 2001) y (Ritchie, 2003).

datos y reconocerlas automáticamente mediante el empleo de un clasificador. Para ello se parte de un subconjunto de chistes y narraciones extraídas del corpus del proyecto CesCa[3].

A continuación se describen los experimentos realizados y los resultados obtenidos.

4. *Experimentos*

Se realizaron 6 tipos de experimentos para conseguir nuestro objetivo. En el primero se buscó diferenciar los datos, es decir el contraste entre chistes y narraciones, a través de la medida de perplejidad (PPL) de los conjuntos. En el segundo se extrajeron las palabras clave (Keywords, KW) con el fin de encontrar elementos contrastantes. En el tercero se midió la información mutua (IM) del léxico para encontrar plantillas subyacentes en las estructuras de los textos. En el cuarto experimento se realizó un etiquetado manual cuyo propósito era distinguir una taxonomía primaria del humor. En el quinto, a partir de la frecuencia del léxico, se hizo un etiquetado semántico para distinguir información que complementara la taxonomía previa. Por último, basados en la información semántica, en el sexto se llevó a cabo una clasificación relacionada con la orientación de los datos. Como medida de evaluación de los resultados obtenidos se utilizaron los clasificadores de Weka de Naïve Bayes y del modelo de regresión lógica multinomial (Witten y Frank, 2005) para un conjunto de 250 textos, de los cuales 220 procedían de chistes y 30 de narraciones.

4.1. Corpus

El corpus CesCa (Català escolar escrit a Catalunya) está constituido por diferentes tipos de textos (vocabularios, definiciones, narraciones y chistes) que se han recogido a partir de 2.460 encuestas realizadas a niños de diferentes centros escolares de Cataluña cuya edad va de los 6 a los 16 años. El número total de palabras de este corpus es de 229.217. Para este estudio se han seleccionado sólo las particiones de chistes y de narraciones. La primera partición conforma el conjunto de datos positivos y contiene un total de 1.867 chistes, mientras que la segunda constituye el conjunto de datos negativos y está compuesta por 2.172 narraciones. La longitud media, en términos de unidades léxicas por chiste y narración, es de 27.10 y 23.22 ítems respectivamente.

Cabe destacar que para los experimentos de PPL, KW e IM se utilizó como modelo de referencia la versión catalana del Leipzig Corpus (LzC) (Quasthoff, Richter, y Biemann, 2006) que tiene un total de 300.000 oraciones.

Las oraciones de este corpus no fueron preprocesadas de ninguna forma, aunque en los datos hay muchos errores ortográficos, falta de separación entre palabras y, en ocasiones, una alternancia de lenguas entre el catalán y el español[4], con el fin de que posteriormente se pudieran analizar los usos lingüísticos de los escolares.

4.2. Perplejidad

Como primera estrategia para saber cuan diferente era la estructura de ambos conjuntos de datos (chistes y narraciones), se calculó su grado de PPL (Jurafsky y Martin, 2007) en relación con un modelo de lenguaje basado en el LzC. Los conjuntos de chistes y narraciones del corpus CesCa se utilizaron como test y el LzC como modelo de referencia. En la Figura 1 se muestran los resultados.

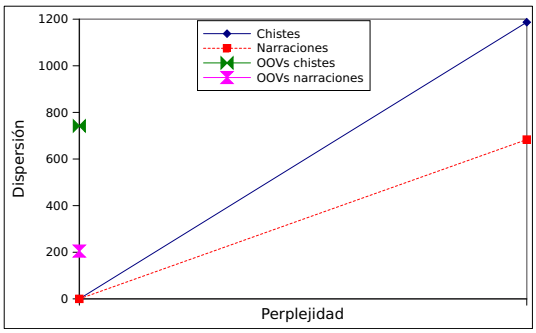

Figura 1: Dispersión de PPL por conjunto

Tal y como se observa en la Figura 1, el grado de dispersión que tienen los chistes es casi el doble que el de las narraciones. Según este modelo, es menos predecible saber qué palabra sigue a otra cuando aparece en los chistes que cuando aparece en las narraciones. De ahí que la perplejidad de los chistes

[3]Este proyecto busca proporcionar a la comunidad educativa catalana una herramienta para conocer los usos lingüísticos de su alumnado. Más información en http://clic.ub.edu/cesca/.

[4]Se cuenta con una versión del corpus segmentado por unidades léxicas para facilitar un procesamiento morfosintáctico posterior. Sin embargo, esta versión no fue utilizada puesto que no presenta un valor añadido para este estudio.

sea mayor. Esto significa que, por medio de un cómputo simple de n-gramas, se pueden diferenciar completamente ambos conjuntos. Otro punto interesante es que, dado el modelo de referencia, la cantidad de palabras no vistas (out of vocabulary, OOVs) es muy superior en los datos positivos (es decir, en los chistes). Esta situación hace pensar que: i) o hay un número importante de neologismos y unidades nuevas en estos datos o bien, ii) los errores que anteriormente se mencionaban se producen más en este conjunto. Este tema queda abierto para análisis posteriores.

4.3. Palabras clave

Partiendo de los resultados obtenidos en el experimento de PPL que indican que tanto los chistes como las narraciones son dos estructuras diferenciables, es lógico pensar que debe haber unidades descriptoras que lo corroboren. Para verificar esta premisa se extrajeron las 100 unidades cuyo valor de *keyness* (Scott, 1997) fuera lo suficientemente elevado como para ser considerada como una palabra clave (*Keyword*, KW). Este cálculo de *keyness* se hace con base al Log Likelihood test (Dunning, 1993) por medio de la comparación de la frecuencia de todas las unidades de cada conjunto con la frecuencia de esas mismas unidades en el corpus de referencia (LzC). Algunas de las unidades significativas más importantes de cada grupo son: *diu, cau, riure, jaimito, tonto, amiga, nen, mare, dos, vaig* para los datos positivos, y *va, novio, raptar, maquineta, noia, agradar, novia* para los negativos. De acuerdo con este cálculo se puede considerar que estas potenciales KWs son un reflejo del tipo de descriptores que usan los niños en sus textos. Es decir, una unidad como *jaimito* podría considerarse un elemento prototípico que utilizarían para generar un determinado tipo de chiste.

4.4. Información mutua

Otro punto importante a determinar era si con la información disponible es posible detectar patrones recurrentes en la construcción de chistes. Para ello se hizo una medición de la Información Mutua (IM) (Oakes, 1998) de las unidades de cada conjunto. Con esta medida se buscó evaluar la probabilidad de que dos unidades formasen un patrón recurrente y no fuesen producto de la casualidad o el estilo. Para el cómputo de la IM se determinaron los siguientes umbrales de búsqueda: rango mínimo de IM = 5 en una ventana no mayor a 3 unidades. En el Cuadro 1 se muestran algunos de los agrupamientos más importantes de los datos positivos y el único agrupamiento de los negativos.

Frec.	W_1	**IM**	W_2	**Set**
16	mis	8.42	tetas	+
28	caca	6.28	pipi	+
45	cotxe	6.08	negres	+
103	fa	5.88	tonterias	+
54	gos	5.33	pere	+
5	casa	5.43	lucas	-

Cuadro 1: Ejemplo de agrupamientos por IM

Cabe destacar que estos resultados dan cierta información que permite pensar que en las narraciones no hay nada que evidencie que se siga un patrón, mientras que en los chistes es notorio que los niños están repitiendo una plantilla que les da resultado como generadora de humor.

Otro rasgo interesante es que aunque el rango de edad es bastante amplio, la frecuencia de uso del léxico no es tan dispersa, tal y como se puede apreciar en la Figura 2, donde se representan con líneas discontinuas el total de unidades léxicas por edad (datos sin normalizar), mientras que con líneas continuas se representan las unidades de los primeros 100 chistes para cada rango de edad (datos normalizados).

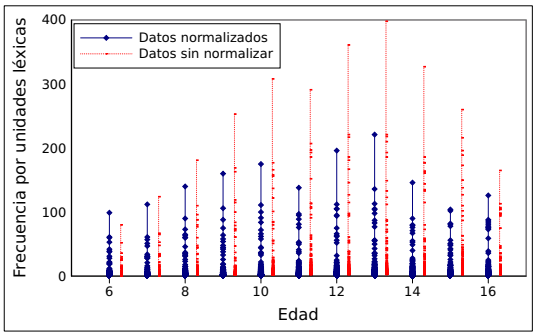

Figura 2: Dispersión léxica por edad

Como se ve en esta figura, la concentración de información por tokens (unidades léxicas) para cada rango de edad está en no más de 100 unidades y se corresponde con el comportamiento esperado, a excepción de los mayores que son quienes menos léxico utilizan, debido a que la proporción de participantes de esta edad es menor que en la de otras edades.

4.5. Etiquetado taxonómico

Con el fin de enriquecer el conjunto de características que definen las propiedades de estos chistes, se llevó a cabo un experimento cuya motivación subyacente era la de obtener elementos que en un futuro pudiesen constituir una taxonomía primaria de esta clase de humor. Para ello, se partió de la premisa que el humor de estos textos se produce invocando dos tipos de referentes: uno interno (I), donde no es necesario recurrir a información externa al chiste para poder entenderlo, por ejemplo, en los juegos de palabras donde el humor se da por factores fonéticos (véase ejemplo 5); y otro externo (E), que requiere de información contextual o metalingüística, no expresada verbalmente, para poder interpretar el chiste (véase ejemplo 6).

(5) Papa, Papa. Què Com sescriu campana. Tal com sona. Doncs llavors escric Tan, Tan, Tan[5].

(6) - Manolo, les set. - Que passin.

La asignación de etiquetas se hizo de forma manual con una muestra de 250 textos, de los cuales 220 eran chistes y 30 narraciones. Las narraciones así como las oraciones incompletas o sin sentido de los chistes se han anotado con la etiqueta *narración*(N). El Cuadro 2 muestra los resultados de esta clasificación.

I	E	N
92	99	59

Cuadro 2: Asignación de referentes

Además de este etiquetado, se buscó en un segundo nivel especificar el tipo de tópicos que pudieran definir clases de chistes. Las categorías que se emplearon son: sentido común (SC), estereotipos (ST), información cultural (IC), narraciones (NA) y otros (O). Las oraciones de 7 a 11 ejemplifican cada una de estas categorías respectivamente.

(7) un soldat li pregunta a un altre. Quina ora es Les tres de la matinada Tan tard. Si mo aguesis preguntat abans (SC).

(8) Perquè el tontos no entren a la cuina Perquè ha un pot que diu sal (ST).

(9) Van dos i cau el del mig (IC).

(10) Quan tota la meva familia estaven a Calì colombia Estaven tots a casa del meu pare que era molt gran. La meva avia estava en el jardì prenen el sol i el meu avi estava a la cuina i li diu a la meva avia donde esta la escalera i li respon en la cocina i li respon qui no esta i li diu mira en el jardin i diu vale i el meu avi li diu que has entendido yo la açucarera no hombre es la escalera (NA).

(11) Quan l'Alex fa tonteries (O).

En el Cuadro 3 se dan los resultados obtenidos.

SC	ST	IC	NA	O
36	23	67	68	58

Cuadro 3: Asignación de categorías

A partir de esta clasificación, el objetivo era saber en qué medida esta información permite discriminar un chiste de una narración. Este punto se trata con más detalle en la Sección 5.

4.6. Patrones semánticos

Un aspecto interesante del experimento previo es que se observó que algunos chistes aparecían recurrentemente variando apenas la estructura narrativa o las unidades empleadas. Para comprobar si existía algún patrón conceptual subyacente se realizó un etiquetado semántico manual del mismo conjunto de prueba. Con este objetivo se seleccionaron las 100 palabras más frecuentes sin tomar en cuenta las *stopwords* y se agruparon en las siguientes categorías: agente (AG), tema (TM), acción (AC), lugar (LG), partes del cuerpo (PC), entes animados (EA)[6] y otros (O). La idea es que estas categorías, que también fueron seleccionadas *a priori*, estén relacionadas con información semántica que dé indicios de qué es lo que proyecta el sentido humorístico de los datos y, en consecuencia, ser capaces de identificarlo y extraerlo automáticamente. El proceso de

[5] En los textos en catalán se ha mantenido la escritura con la ortografía original.

[6] La diferencia entre AG y EA está en que el AG se concibe como un humano que realiza una acción volitivamante, por ejemplo, *Jaimito*, mientras que un EA es un participante no humano, por ejemplo, *gos* (perro).

asignación se hizo en dos fases: en la primera se etiquetó cada palabra con la categoría que reflejaba de manera prioritaria su papel (o función) en el chiste y, en la segunda, la categoría que le seguía en importancia, si es que la había. La Figura 3 muestra la distribución por categorías para cada una de las fases de etiquetado.

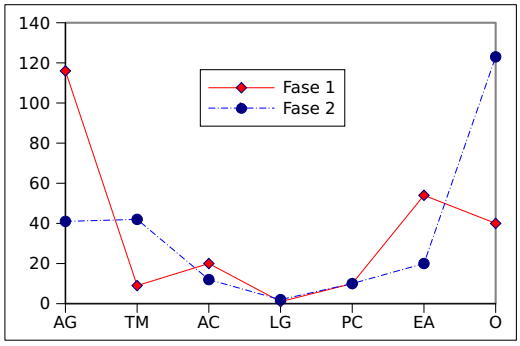

Figura 3: Distribución de categorías

Según la información que se desprende de la Figura 3 se puede inferir que, dadas las categorías más importantes que se focalizan en los chistes, el patrón semántico que subyace a estos datos estaría definido por la secuencia de categorías: "AG/EA - AC - TM/AG"[7]. Es decir, algún humano o animal, realiza una acción sobre él mismo, sobre un objeto u otro humano.

4.7. Orientación

El último experimento toma como base la premisa de que el humor se alimenta de factores emocionales, incluso negativos (Ruch, 2001), para producirse. En este sentido, es necesario definir parámetros que indiquen que esas emociones, que encuentran una salida por medio de la expresión verbal del humor, pueden servir como rasgos discriminadores entre una expresión chistosa y una seria. Para ello, en función de los resultados de Mihalcea y Pulman (2007), donde se destaca que el humor tiene una tendencia hacia las connotaciones negativas, se busca evaluar si con la presencia de determinados elementos, tales como verbos o adjetivos que denoten un carácter negativo, es posible establecer una característica útil para discriminar el conjunto positivo del negativo.

Para la realización de este experimento se establecieron dos criterios: i) orientación negativa, que está basada en el léxico más frecuente, por ejemplo, verbos como *caure* ('caer'), o unidades que aparecen en los chistes de la categoría estereotipo (ST): *lepe*, *xines*, *espanyol*, y ii) orientación neutra, donde el rasgo más importante es que se cuente un suceso, por ejemplo, los chistes etiquetados como narraciones (NA). En el Cuadro 4 se presentan los resultados obtenidos.

Negativa	Neutra
131	119

Cuadro 4: Orientación del conjunto de prueba

Considerando que de los 119 textos etiquetados como neutros, 30 de ellos pertenecen al conjunto negativo y de los 59 que fueron asignados al referente N, 29 son oraciones sin sentido o inconclusas, queda un porcentaje interesante de chistes que caen dentro de la orientación negativa. Esto significa que, dada esta diferencia, la característica de orientación podría traducirse en un elemento interesante que sirva como diferenciador de los conjuntos.

5. Evaluación y discusión

Con el fin de saber cuan efectivas resultan las características identificadas en la sección previa, se realizó un proceso de evaluación por medio de dos clasificadores: Bayes[8] y el modelo de regresión lógica multinomial. Cada clasificador fue evaluado utilizando las características en el orden en que fueron descritas y aplicando el método de validación cruzada (Witten y Frank, 2005). Las únicas características que no fueron evaluadas son las de la PPL, la cual sólo se utilizó como elemento para mostrar que había dos conjuntos distinguibles uno de otro, y la de IM, cuyo objetivo era encontrar patrones prototípicos en los datos. En la Figura 4 se observa el porcentaje de acierto en la clasificación.

[7] En este caso no se toma en cuenta O aunque su frecuencia es considerable puesto que, como se dijo en la Sección 4.1.2, hay muchos errores en el corpus y, por tanto, hay oraciones que ni siquiera están completas, motivo por el cual no se entendía el chiste y no se podía asignar una categoría más específica.

[8] Es pertinente señalar que los resultados que se obtienen con este tipo de clasificadores dependen de lo balanceado que esté el corpus, por tanto, aclaramos que, dada nuestra colección, el resultado puede verse afectado por el tamaño de las clases consideradas.

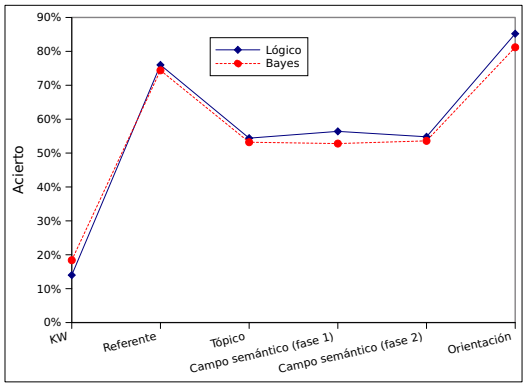

Figura 4: Clasificación por característica

Como se muestra en la Figura 4, el porcentaje de clasificación alcanzado con ambos clasificadores es aceptable, al menos para un par de características (Referente y Orientación). Mientras que con el rasgo de KW el porcentaje de acierto es mínimo, con los rasgos de tópico y de etiquetado semántico se ve una mejoría que eleva el acierto hasta un porcentaje superior al 50 %. Sin embargo, es con los rasgos restantes con los que este porcentaje llega al 85.2 % de precisión con el clasificador lógico. Esto significa que los elementos detectados tienen una pauta que permite discriminar ambos conjuntos de datos, aunque deben ser mejorados para aumentar la eficacia del reconocimiento. A continuación se analizan los resultados.

De acuerdo con los resultados de los experimentos y de la evaluación, se pueden inferir las siguientes consideraciones:

i) La PPL, además de demostrar que los dos conjuntos son diferentes entre sí, apunta a que las asociaciones del léxico de los chistes son más heterogéneas.

ii) En cuanto a las KWs, aunque con ellas se refleja que hay ciertas unidades más prototípicas para producir el humor, es claro que en la evaluación como característica para discriminar los datos positivos de los negativos no dan buenos resultados. El motivo presumiblemente radica en el procedimiento de elección de las 5 KWs con mayor valor de keyness de cada conjunto, que no necesariamente aparecieron en el subconjunto de prueba, castigando así el resultado de la clasificación y mostrándolo como el rasgo menos relevante.

iii) En relación con los resultados del cálculo de IM, se puede suponer que los niños emplean mayor riqueza comunicativa, léxica y cognitiva en las narraciones, mientras que en los chistes explotan elementos constantes dentro de patrones estables que les dan el resultado esperado sin tener que recurrir a más estrategias como en las narraciones. Todo ello sugiere que hay elementos que se pueden generalizar, siempre dentro de esta clase de humor, para generar patrones básicos de chistes. Asimismo, las unidades más frecuentes por edad podrían justificar que hay un conjunto de tópicos comunes que varían de acuerdo con el patrón en el que se insertan.

Los resultados de los experimentos posteriores podrían ser refutados como subjetivos, sin embargo, mientras no exista un esquema que indique cómo sería una clasificación objetiva del humor (Ritchie, 2003), se justifican los mismos bajo un argumento operativo, siempre con la precaución de no generalizarlos hasta que no se verifiquen en más investigaciones.

iv) Respecto a la pretendida *taxonomía*, es relevante empezar a construir un esquema, con base en datos empíricos, que dé cuenta de cómo se establecen las relaciones en el interior de la estructura de los chistes. En este sentido, la clasificación a través de los dos niveles propuestos sirvió como primer modelo para tratar de jerarquizar y dar un orden más manejable a los datos. Los resultados que se lograron con esta clasificación reflejan que, aunque subjetiva, para los fines perseguidos resultó funcional. No obstante, se requiere una metodología menos subjetiva que pueda reflejar una taxonomía del humor más general.

v) Sobre los *patrones semánticos*, los resultados que se obtuvieron sugieren que las categorías semánticas pueden relacionarse con la forma en la que los niños expresan sus sentimientos. Los rasgos que proyectan en cada categoría son distintos entre sí, mientras que en el humor se proyecta información más orientada hacia la adjetivación y partes del cuerpo, en las narraciones se orienta más hacia descripciones y localizaciones.

vi) El último experimento demostró que el rasgo más importante de los chistes es el de tener una *orientación* negativa. De manera que el humor surja por el hecho de hacer mofa o burla de alguien por pertenecer a un determinado colectivo o por un acto ridículo. Resulta interesante comprobar que la clasificación da mejores resultados si se toma en cuenta este rasgo.

Finalmente, los resultados de la evalu-

ación demuestran que hay características interesantes que sirvieron para identificar los chistes. Por ejemplo, si se toman en cuenta los rasgos de referente y orientación, el porcentaje de clasificación mejora notablemente, llegando hasta un 85.2 % para este último.

6. Conclusiones y trabajo futuro

En esta investigación se ha realizado un trabajo de identificación y extracción de características sobre un conjunto de textos infantiles con el fin de utilizarlas como elementos discriminadores en la tarea de RAH. Los resultados muestran que éstas pueden llegar a ser útiles, sobre todo en el caso de las características PPL, referente y orientación. Asimismo, es destacable señalar que, dado el tipo de datos, a pesar de los errores, se pueda discriminar un chiste de una narración con un porcentaje alto de acierto.

Como trabajo futuro se busca automatizar y, sobre todo, estandarizar los criterios empleados para lograr un modelo que describa de forma más completa el humor y pueda ser aplicado a otras clases de textos.

Bibliografía

Attardo, S. 2001. *Humorous Texts: A semantic and pragmatic analysis*. Mouton de Gruyter.

Binsted, K. 1996. *Machine humour: An implemented model of puns*. Ph.D. tesis, University of Edinburgh, Edinburgh, Scotland.

Binsted, K. y G. Ritchie. 1997. Computational rules for punning riddles. *Humour*, 10:25–75.

Buscaldi, D. y P. Rosso. 2007. Some experiments in humour recognition using the italian wikiquote collection. En Springer-Verlag, editor, *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, Lecture Notes in Computer Science, páginas 464–468.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Jurafsky, D. y D. Martin. 2007. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Mihalcea, R. y S. Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. En Springer-Verlag, editor, *8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2007*, Lecture Notes in Computer Science, páginas 337–347.

Mihalcea, R. y C. Strapparava. 2006a. Learning to Laugh (Automatically): Computational Models for Humor Recognition. *Journal of Computational Intelligence*, 22(2):126–142.

Mihalcea, R. y C. Strapparava. 2006b. Technologies that make you smile: Adding humour to text-based applications. *IEEE Intelligent Systems*, 21(5):33–39.

Oakes, M. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Quasthoff, U., M. Richter, y C. Biemann. 2006. Corpus portal for search in monolingual corpora. En *Proceedings of the 5th International Conference on Language Resources and Evaluation*, páginas 1799–1802.

Ritchie, G. 2003. *The Linguistic Analysis of Jokes*. Routledge.

Ruch, W. 2001. The perception of humor. En World Scientific, editor, *Emotions, Qualia, and Consciousness. Proceedings of the International School of Biocybernetics*, páginas 410–425.

Scott, M. 1997. Pc analysis of key words - and key key words. *System*, 25(1):1–13.

Sjobergh, J. y K. Araki. 2007. Recognizing humor without recognizing meaning. En Springer-Verlag, editor, *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, Lecture Notes in Computer Science, páginas 469–476.

Stock, O. y C. Strapparava. 2005. Hahacronym: A computational humor system. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, páginas 113–116.

Witten, I. y E. Frank. 2005. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers. Elsevier.

A. Anexo 1: Traducciones

En este anexo se dan las traducciones y claves para la interpretación de los ejemplos que están en italiano y catalán.

Ejemplo 4. ¿Acaso piensas que me caso contigo por tus ocho millones de dote? Como te equivocas. Igual me casaría contigo si tuvieses nueve.

Ejemplo 5. Llega el niño con su padre y le pregunta: Papá, papá, ¿cómo se escribe campana? El padre le responde: así como suena. Y el niño en su tarea escribe tan, tan, tan.

Ejemplo 6. La escena de esta oración es la de Manolo en la cama, el reloj marcando las 7 y la mujer ya levantada, avisándole de la hora. El efecto humorístico está en la ambigüedad que produce la frase *les set* (las siete), la cual Manolo completa con un sustantivo, por ejemplo chicas, que da sentido a su respuesta: Las siete...[chicas], ¡qué pasen!

Ejemplo 7. Un soldado le pregunta a otro: ¿qué hora es? Las tres de la mañana responde. ¿Tan tarde? Si me lo hubieras preguntado antes.

Ejemplo 8. ¿Por qué los tontos no entran a la cocina? Porque hay un bote que dice sal.

Ejemplo 9. Van dos y se cae el de en medio. La clave de este chiste está en que se está focalizando lo absurdo y el sin sentido que llega a tener cualquier situación.

Ejemplo 10. Cuando toda mi familia estaba en Calí, Colombia. Todos estaban en la casa de mi padre, la cual era muy grande. Mi abuela estaba tomando el sol en el jardín y mi abuelo estaba en la cocina. Entonces mi abuelo le pregunta a mi abuela dónde está la escalera, ella le responde: en la cocina. Él le dice que no está ahí y le dice que busque en el jardín. Mi abuela sorprendida le pregunta, ¿qué me pediste? ¿La azucarera? Y mi abuelo le responde, ¡No, la escalera!

Ejemplo 11. Cuando Alex hace tonterías.