Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 143-150 recibido 27-03-2015 revisado 27-04-2015 aceptado 20-05-2015

CRiSOL: Base de conocimiento de opiniones para el español*

CRiSOL: Opinion Knowledge-base for Spanish

M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia
Departamento de Informática
Universidad de Jaén, E-23071 - Jaén, España
{mdmolina, emcamara, maite}@ujaen.es

Resumen: El presente trabajo se centra en la clasificación de polaridad de comentarios de hoteles en español (COAH) y presenta un nuevo recurso léxico, CRiSOL. Este nuevo recurso toma como base la lista de palabras de opinión iSOL, a la cual incluye los valores de polaridad de los *synsets* de SentiWordNet. Debido a que SentiWordNet no es un recurso para español, se ha tenido que usar como pivote la versión española de WordNet incluida en el Repositorio Central Multilingüe (MCR). Se ha desarrollado un clasificador de la polaridad no supervisada para evaluar la validez de CRiSOL. Los resultados obtenidos con CRiSOL superan los obtenidos por los lexicones base iSOL y SentiWordNet por separado, lo cual nos anima a seguir trabajando en esta línea.
Palabras clave: Análisis de Opiniones, combinación de recursos de opinión, clasificación de la polaridad

Abstract: In this paper we focus on Spanish polarity classification in a corpus of hotel reviews (COAH) and we introduce a new lexical resource called CRiSOL. This new resource is built on the list of Spanish opinion words iSOL. CRiSOL appends to each word of iSOL the polarity value of the related synset of SentiWordNet. Due to the fact that SentiWordNet is not a Spanish linguistic resource, a Spanish version of WordNet had to be used. The Spanish version of WordNet chosen was Multilingual Central Repository (MCR). An unsupervised classifier has been developed with the aim of assessing the validity of CRiSOL. The results reached by CRiSOL are higher than the ones reached by iSOL and SentiWordNet, so that encourage us to continue this research line.
Keywords: Sentiment Analysis, opinión resources combination, polarity classification

1 Introducción

Con el paso de los años, el estudio computacional de la opinión se ha ido convirtiendo en una aplicación del Procesamiento del Lenguaje Natural (PLN) que no cesa de atraer el interés de nuevos investigadores. El persistente interés está motivado principalmente por la continuada progresión de la necesidad de conocer la orientación de las opiniones que se publican en Internet.

El Análisis de Opiniones (AO) es la tarea encargada del estudio de la opinión en el ámbito del PLN. Según Cambria y Hussain (2012), el AO se define como el conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios. En efecto, esta actualización de la definición clásica de AO de Pang y Lee (2008) sintetiza las distintas operaciones que implica el procesamiento de la opinión.

Una de las operaciones que se mencionan en la definición de Cambria y Hussain (2012) es la de clasificación. La clasificación de la polaridad tiene como fin la determinación de la categoría de opinión que se le puede asignar a un mensaje. La categoría puede ser binaria, positiva y negativa, o estar conformada por diversos niveles de intensidad de opinión. En la experimentación que aquí se presenta, se evalúa un sistema de clasificación de la polaridad binaria.

* Esta investigación ha sido parcialmente financiada por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España y el proyecto AORESCU (P11-TIC-7684 MO) del gobierno autonómico de la Junta de Andalucía. Por último, el proyecto CEATIC (CEATIC-2013-01) de la Universidad de Jaén también ha financiado parcialmente este artículo.

ISSN 1135-5948

La clasificación de la polaridad se puede realizar mediante un sistema basado en aprendizaje automático (Pang, Lee, y Vaithyanathan, 2002), no supervisado (Turney, 2002) o híbrido (Prabowo y Thelwall, 2009). Independientemente de la estrategia que se emprenda, en muchos casos los sistemas requieren de la información que aportan los recursos lingüísticos de opinión. La disponibilidad de recursos de opinión no es muy abundante, y menos aún en español, de manera que la generación de nuevos recursos, o la mejora de los existentes, son aportaciones valiosas a la tarea de AO.

Las listas de palabras de opinión son recursos léxicos constituidos por palabras categorizadas normalmente en dos clases de opinión, positiva y negativa. Por otro lado, las bases de conocimiento de opinión son recursos que toman como sustento bases de conocimiento léxicas, como puede ser WordNet, y asignan a los conceptos que las conforman unos valores de polaridad.

La orientación semántica que asigna una lista de opinión a una palabra es totalmente rígida, ya que, o la palabra pertenece a una clase u a otra. Por lo tanto, cabe preguntarse ¿se mejoraría la capacidad de clasificación de una lista de palabras de opinión si se incluyen a sus palabras los valores de polaridad de una base de conocimiento de opinión?

Por otra parte, en ocasiones las bases de conocimiento de opinión insertan ruido en los sistemas de clasificación de la polaridad. Sin embargo, si se filtra la información de una base conocimiento con una lista de palabras de opinión ¿sería posible una mejora de la clasificación?

Estas dos preguntas son las que tratamos de responder en el presente artículo. Para ello, se estudiará la clasificación de la polaridad mediante la construcción de un nuevo recurso que combina la base de conocimiento de opinión SentiWordNet (SWN) (Baccianella, Esuli, y Sebastiani, 2010) y la lista de opinión iSOL (Molina-González et al., 2013). El nuevo recurso CRiSOL (*Combined Resources in iSOL*) toma como base el lexicón iSOL y para cada uno de los términos de dicho lexicón, se intenta asociar su *synset* en SWN extrayendo e integrando la tripleta de valores que mantiene dicho recurso (positivo, neutro, y negativo). El sistema resultante es evaluado sobre un corpus de opiniones en el dominio de hoteles, COAH (Molina-González et al., 2014).

El artículo se estructura de la siguiente manera: la siguiente sección resumirá algunos trabajos relacionados. La Sección 3 se circunscribirá a la descripción de los recursos lingüísticos que se han empleado. Posteriormente se detallará el sistema construido. En la Sección 5 se encontrará el análisis de los resultados obtenidos. La última sección detallará las conclusiones alcanzadas, así como las actuales líneas de trabajo.

2 Trabajos relacionados

El presente trabajo se encuadra en la investigación relacionada con la clasificación de la polaridad en un idioma distinto al inglés, basado en el desarrollo de un sistema fundamentado en el uso de una lista de palabras y en la combinación de recursos lingüísticos de opinión, con el fin de emplear la información oportuna que propicie la mejor clasificación posible.

Se remarca el hecho de que la investigación que se expone no es sobre textos en inglés, porque la mayor parte de la investigación en AO se centra en dicha lengua, como se puede comprobar en (Pang y Lee, 2008; Liu, 2012; Tsytsarau y Palpanas, 2012).

Un ejemplo de la relevancia de emplear recursos lingüísticos en AO, tanto para inglés como para español, se puede encontrar en (Brooke, Tofiloski, y Taboada, 2009). En dicho trabajo, los autores concluyen que la inclusión de la información que aportan los recursos de opinión es beneficiosa para un sistema de clasificación de la polaridad, ya sea éste supervisado o no supervisado.

En el contexto de la generación de recursos lingüísticos para AO, destacan aquellos en los que se presentan nuevos corpus de opiniones, como por ejemplo el corpus *Spanish Movie Review* (Cruz et al., 2008), la versión española de SFU corpus (Brooke, Tofiloski, y Taboada, 2009), el corpus EmotiBlog (Boldrini et al., 2009) o el corpus que se va a emplear en esta evaluación, el corpus COAH (Molina-González et al., 2014). En el ámbito de la generación de listas de palabras de opinión deben ser resaltados algunos estudios. Rangel, Sidorov, y Suárez-Guerra (2014) presentan un léxico de emociones en español compuesto por 2.036 vocablos acompañados de un valor, que representa la probabilidad de uso afectivo (PFA) del término con respecto a una de las siguientes emociones: alegría,

enfado, miedo, tristeza, sorpresa y repulsión. ML-SentiCon (Cruz et al., 2014) es un recurso que integra listas de palabras de opinión en español, inglés, gallego, vasco y catalán.

En cuanto a la combinación de métodos de clasificación y de recursos también se pueden encontrar ejemplos en la literatura relacionada con AO. Aunque no se trata de una experimentación sobre textos en español, en (Kennedy y Inkpen, 2006) se muestra como la combinación de un método supervisado con un clasificador de polaridad basado en una lista de palabras de opinión mejora los resultados de los dos clasificadores base por separado. Centrándonos exclusivamente en español, Martínez-Cámara et al. (2014), siguiendo una metodología de *stacking* (Wolpert, 1992), combinan con éxito los dos recursos que se emplean en el presente trabajo, iSOL y SentiWordNet. En dicho trabajo, se evaluaron dos sistemas no supervisados para la clasificación de las opiniones recogidas en el corpus *Spanish Movie Review*, demostrándose que la combinación de dos clasificadores basados en dos recursos de opinión mejora los resultados que obtienen por separado.

3 Recursos

En esta sección se describe, en primer lugar, el corpus de opiniones sobre hoteles. Este corpus se llama COAH y está disponible libremente. En segundo lugar se comentarán los lexicones usados para la experimentación. Se parte del lexicón independiente del dominio iSOL, al que le seguirá la descripción de SentiWordNet, recurso léxico ampliamente usado en documentos escritos en inglés, y el recurso lingüístico Repositorio Central Multilingüe (MCR). Para terminar se detallará la combinación usada de estos recursos léxicos para generar el nuevo recurso CRiSOL.

3.1 Corpus COAH

COAH[1] (*Corpus of Opinion about Andalusian Hotels*) es un corpus que contiene comentarios sobre 10 hoteles de cada una de las ocho provincias andaluzas, obteniendo un total de 1.816 opiniones escritas en español en los últimos años sobre los 80 hoteles elegidos en total. En (Molina-González et al., 2014) se detalla la generación del corpus.

Este corpus se compone de dos tipos de información. Una sobre el hotel (nombre, dirección) y otra sobre la opinión del huésped del hotel (valoración global, la identificación del usuario, la valoración de relación calidad/precio, la limpieza, etc.).

La valoración global del hotel está en una escala de 1 a 5. El valor 1 significa que el autor manifiesta una opinión muy negativa sobre el hotel, mientras que una puntuación de 5 representa una valoración positiva. Los hoteles con valor 3 se pueden catalogar como hoteles neutros, ni buenos ni malos, y por tanto, difíciles de clasificar. Para los experimentos se descartan aquellas opiniones neutras, es decir, con valoración 3. El resto de opiniones son catalogadas como positivas si su valoración es 4 ó 5, y negativas si su valoración es 1 ó 2. Por tanto, la clasificación binaria de las opiniones sobre hoteles del corpus COAH es la que se muestra en la Tabla 1.

Clases	**Opiniones**
Positiva	1.020
Negativa	511
Total	**1.531**

Tabla 1: Clasificación binaria del corpus COAH

3.2 iSOL

Este lexicón fue generado a partir de una traducción automática del inglés al español del lexicón de Bing Liu generando el recurso SOL (*Spanish Opinion Lexicon*) (Martínez-Cámara et al., 2013).

La corrección manual de SOL dio lugar a iSOL. Por un lado, debido a la inflexión morfológica española, se tiene que mientras un adjetivo inglés, por lo general, no posee ni género ni número, y es representado por un solo término, al adjetivo español le corresponde hasta cuatro posibles palabras traducidas del inglés, dos para el género (masculino o femenino) y dos para el número (singular o plural). Por otra parte, siguiendo la filosofía de Bing Liu se introdujo en las listas algunas palabras mal escritas o inexistentes en el Diccionario de la Real Academia Española (DRAE) ya que aparecen con mucha frecuencia en el contenido de los medios de comunicación social, como por ejemplo "kaput", "pillín" o "coñacete". Finalmente iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas. Por ende, iSOL contiene 8.135 palabras de opinión.

[1] `http://sinai.ujaen.es/coah`

3.3 SentiWordNet

SentiWordNet (SWN) es uno de los lexicones más usados en AO y está construido sobre la base de datos léxica WordNet. Asigna a cada *synset* en WordNet tres propiedades (positivo, neutro y negativo), e indica la probabilidad de que el concepto sea positivo, neutro o negativo. Al tratarse de valores de polaridad, la suma de la tripleta debe ser 1. SWN cubre la totalidad de los *synsets* de WordNet, en concreto 117.000.

En SWN cada entrada contiene la categoría morfológica y un índice, que identifican unívocamente al *synset* en WordNet, junto a las tres propiedades que indican la probabilidad de que el *synset* sea positivo, neutro o negativo.

3.4 MCR

MCR (*Multilingual Central Repository*) (Atserias et al., 2004; Gonzalez-Agirre, Laparra, y Rigau, 2012) es un recurso lingüístico a gran escala que puede ser usado en procesos semánticos que necesitan gran cantidad de conocimiento lingüístico.

MCR integra en el mismo marco de trabajo de EuroWordNet, diversas versiones de WordNet para diferentes lenguas, inglés, español, vasco, catalán y gallego. Los *synsets* han sido construidos siguiendo el modelo propuesto por EuroWordNet, en los cuales los WordNet se enlazan mediante un índice entre lenguas (*InterLingual Index-ILI*). Por medio de este *ILI* los lenguajes están conectados, haciendo posible ir desde una palabra de un idioma a otras palabras similares traducidas a otros idiomas. Este hecho es el que nos permite enlazar SentiWordNet para el idioma español utilizando el *ILI* de MCR en el SWN en inglés.

La versión final de MCR contiene alrededor de 1,6 millones de relaciones semánticas entre los synsets, siendo la mayoría de ellos adquiridos mediante métodos automáticos. Este recurso está en continuo crecimiento siendo la última versión disponible MCR 3.0.

3.5 CRiSOL

Como se ha comentado en la Sección 3.2, iSOL es un lexicón de palabras de opinión en español compuesto por 2.509 palabras positivas y 5.626 palabras negativas.

En el presente artículo se pretende generar un nuevo recurso que combine la información de opinión de iSOL y de SentiWordNet. Para ello se intenta añadir a iSOL las puntuaciones de polaridad de los conceptos de SentiWordNet. iSOL es un recurso formado por palabras, o mejor dicho, por formas, ya que, tanto lemas como algunas de sus derivaciones constituyen iSOL. Por otro lado, SentiWordNet es un recurso conformado por conceptos en inglés, de manera que se hace obligatorio el uso de un recurso auxiliar para enlazar las formas de iSOL y SentiWordNet, el cual será el ya descrito MCR.

El proceder habitual en el uso de una base de conocimieto léxica basada en la estructura de WordNet, como es el caso de MCR y de SentiWordNet, se corresponde con el uso del identificador de los conceptos (*ILI*) para recuperar la información asociada al concepto. En este caso no se cuenta con *ILIs*, sino con formas lingüísticas de una lista de palabras de opinión. MCR asocia a cada lema un *ILI*, lo cual identifica inequivocamente uno de los posibles conceptos del lema. Tomando ese *ILI*, ya sí es posible acudir a SentiWordNet y obtener las puntuaciones de polaridad asociadas a dicho concepto. Por tanto, el proceso de generación de CRiSOL comenzó con la obtención de los lemas de las palabras de iSOL. Una vez obtenidos los lemas, el siguiente paso fue encontrar el *ILI* asociado al lema en MCR. Como es sabido, un lema puede tener asociados varios identificadores, dado que es común que un lema represente a varios conceptos. Para la primera versión de CRiSOL, se siguió como heurística el tomar como *ILI* el primero de los asociados al lema. El último paso fue el de recuperar de SentiWordNet los valores de polaridad asociados al *ILI*.

La Figura 1 representa el proceso de generación de CRiSOL, la cual se trata de una base de conocimiento compuesta por los mismos términos de iSOL, de los cuales 4.434, además de contar con la etiqueta de polaridad de iSOL, están complementados por la categoría morfológica y las puntuaciones de polaridad de SentiWordNet.

4 Experimentos y resultados

Antes de llevar a cabo los experimentos, las opiniones de hoteles del corpus COAH han sido preprocesadas con el fin de tener en cuenta los mismos criterios que se utilizaron en la generación del lexicón iSOL. Por ejemplo, las letras mayúsculas se han cambiado a minúsculas y se han eliminado las tildes.

Suponiendo que C es un comenta-

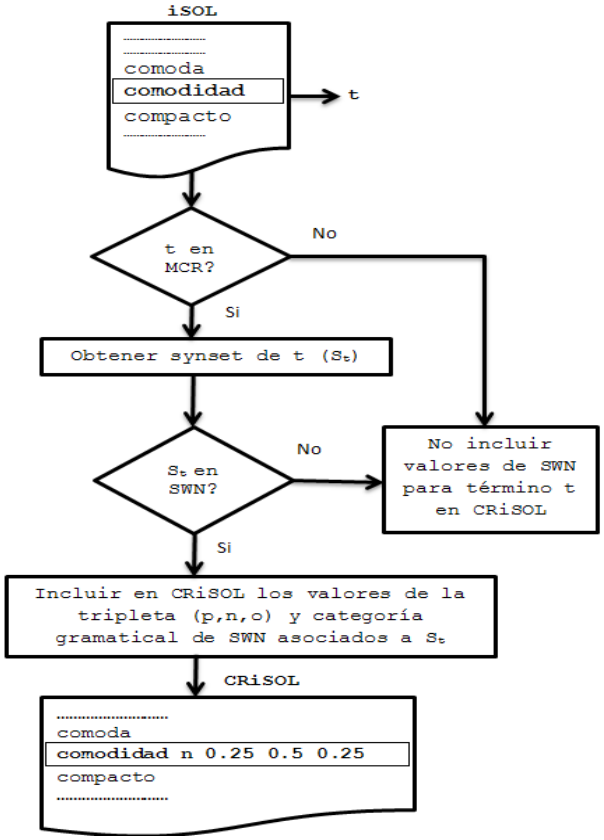

Figura 1: Proceso de generación de CRiSOL

rio de COAH, que t es un término de C, que $SWN[synset][positivo]$ y $SWN[synset][negativo]$ representan el valor de polaridad positiva y negativa de SWN, y que $iSOL^+$ e $iSOL^-$ es la lista de términos positivos y la lista de términos negativos de iSOL, se implementan cinco experimentos con el fin de responder a las preguntas hechas anteriormente.

El primer experimento (Algoritmo 1), sigue una metodología simple basada en la cuenta del número de palabras en cada opinión del corpus COAH incluidas en las listas iSOL. Así, nuestro método clasifica el comentario como positivo si el número de palabras positivas encontradas es igual o mayor que el número de palabras negativas encontradas, o como negativo en el resto de casos.

El segundo experimento (Algoritmo 2) hace uso de SentiWordNet. Como se ha comentado en la Sección 3.4, a través del *ILI* del MCR es posible ir desde una palabra de un idioma a otras palabras similares traducidas a otros idiomas. Este hecho es el que nos permite conseguir una versión de SentiWordNet para el español utilizando el *ILI* de MCR en el SWN en inglés. Usando este método,

```
Entrada: COAH, iSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ iSOL+ entonces
                Positivas ⟵ Positivas + 1
            fin
            si t ∈ iSOL− entonces
                Negativas ⟵ Negativas + 1
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 1: iSOL: Clasificación de la polaridad basada en el uso de la iSOL.

los valores de la positividad y negatividad de las palabras encontradas en SWN para cada comentario del corpus se sumarán respectivamente. Así para cada comentario se obtendrán dos resultados, uno de positividad y otro de negatividad, y si el primer valor es mayor o igual que el segundo se considerará el comentario como positivo, siendo catalogado como comentario negativo en caso contrario.

```
Entrada: COAH, MCR, SWN
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ MCR entonces
                synset ⟵ MCR[t]
                si synset ∈ SWN entonces
                    Positivas ⟵ Positivas +
                    SWN[synset][positivo]
                    Negativas ⟵ Negativas +
                    SWN[synset][negativo]
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 2: SWN: Clasificación de la polaridad basada en el uso de SentiWordNet.

El tercer experimento (Algoritmo 3) utiliza los resultados del experimento 2 para aquellas palabras que existan en SWN. El resto de palabras se buscarán en iSOL y si están contenidas la polaridad positiva y negativa se hallará contando las palabras encontradas en cada lista iSOL. Halladas las polaridades positivas y negativas usando SWN e iSOL, se sumarán las positivas por un lado y las negativas por el otro. Se considera un comentario positivo si el valor de polaridad positiva es

mayor o igual que el valor de la polaridad negativa, siendo el comentario negativo para el resto de casos.

```
Entrada: COAH, MCR, SWN, iSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ MCR entonces
                synset ⟵ MCR[t]
                si synset ∈ SWN entonces
                    Positivas ⟵ Positivas +
                    SWN[synset][positivo]
                    Negativas ⟵ Negativas +
                    SWN[synset][negativo]
                si no
                    si t ∈ iSOL+ entonces
                        Positivas ⟵
                        Positivas + 1
                    fin
                    si t ∈ iSOL− entonces
                        Negativas ⟵
                        Negativas + 1
                    fin
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 3: SWN_iSOL: Clasificación basada en la ampliación de SWN con iSOL.

El cuarto y quinto experimento utilizan el lexicón enriquecido CRiSOL, sumando los atributos de polaridad (positividad y negatividad) de aquellas palabras que aparecen en cada comentario del corpus de hoteles.

El cuarto experimento (Algoritmo 4) clasifica el comentario como positivo si la polaridad positiva total hallada mediante la suma de las polaridades procedentes de CRiSOL en el comentario es mayor o igual que la polaridad negativa total hallada. La opinión será negativa en el resto de los casos.

El quinto experimento (Algoritmo 5) hace uso de los resultados obtenidos en el experimento cuarto. Así, a estos resultados se les añadirá el número de palabras positivas o negativas existentes en CRiSOL pero que no se han encontrado en SWN mediante la metodología explicada en la Sección 3.4 y por tanto no tienen valor en los atributos de positividad y negatividad. Las sumas resultantes se compararán y al igual que en casos anteriores, este experimento clasificará el comentario como positivo si el valor resultante positivo total es mayor o igual que el valor resultante negativo total, o como comentario negativo en el resto de casos.

En la Tabla 2 se muestran los resultados obtenidos por los cinco algoritmos expuestos anteriormente.

```
Entrada: COAH, CRiSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ CRiSOL[SWN] entonces
                Positivas ⟵ Positivas +
                CRiSOL[SWN][t][positivo]
                Negativas ⟵ Negativas +
                CRiSOL[SWN][t][negativo]
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 4: CRiSOL[SWN]: Clasificación de la polaridad basada en el uso de los valores de SWN que se encuentran en CRiSOL

```
Entrada: COAH, CRiSOL
inicio
    para cada C en COAH hacer
        Positivas ⟵ 0
        Negativas ⟵ 0
        para cada t en C hacer
            si t ∈ CRiSOL Y CRiSOL[fuente] =
            SWN entonces
                Positivas ⟵ Positivas +
                CRiSOL[SWN][t][positivo]
                Negativas ⟵ Negativas +
                CRiSOL[SWN][t][negativo]
            fin
            sinó, si t ∈ CRiSOL[fuente] = iSOL
            entonces
                si t ∈ CRiSOL[iSOL+]
                entonces
                    Positivas ⟵ Positivas + 1
                fin
                si no
                    Negativas ⟵
                    Negativas + 1
                fin
            fin
        fin
        si Positivas ≥ Negativas entonces
            C_Polaridad ⟵ positivo
        si no
            C_Polaridad ⟵ negativo
        fin
    fin
fin
```

Algoritmo 5: CRiSOL: Clasificación de la polaridad basada en el uso de CRiSOL.

5 Análisis de resultados

De los resultados recogidos en la Tabla 2 se deben destacar varios aspectos. Primeramente resaltar el buen comportamiento por separado de iSOL, y de SWN. Es remarcable también el hecho de que una lista de palabras de opinión se comporta mejor que una

Algoritmo	Macro-P	Macro-R	Macro-F1	Accuracy
iSOL	91,64 %	83,21 %	87,22 %	88,50 %
SWN_MCR	88,85 %	82,27 %	85,71 %	87,46 %
SWN_iSOL	90,52 %	85,17 %	87,76 %	89,22 %
CRiSOL[SWN]	88,19 %	83,36 %	85,70 %	87,52 %
CRiSOL	90,26 %	87,13 %	88,66 %	90,07 %

Tabla 2: Resultados obtenidos en la clasificación binaria de corpus COAH usando diferentes lexicones

base de conocimiento de opinión, aunque este comportamiento ya se intuía, porque como se indica en la Introducción, las bases de conocimiento de opinión en ocasiones insertan ruido al proceso de clasificación. Ambos recursos adolecen del mismo problema, un bajo *recall*, porque su capacidad de clasificación está limitada al vocabulario que cubren.

Para mejorar la cobertura de ambos recursos, la solución inmediata que se piensa es la de su uso conjunto (SWN_iSOL), y como se puede apreciar en la tabla de resultados, dicha combinación proporciona unos mejores resultados. La mejora se produce por aminorar el problema anterior, es decir, por mejorar la cobertura de la clasificación.

El problema de la inserción de ruido en el proceso de clasificación por parte de una base de conocimiento de opinión se puede solucionar con el filtrado de la misma con una lista de palabras de opinión, y es eso precisamente lo que se hace en CRiSOL[SWN]. Los resultados muestran que se mejora la capacidad de clasificación de la base de conocimiento, debido principalmente a un incremento de la cobertura del clasificador. Por último, si se emplea CRiSOL, es decir, el uso combinado de una lista de palabras de opinión, y una base de conocimiento de opinión filtrada por la lista, los resultados que se obtienen son globalmente mejores que el uso por separado de ambos recursos.

6 Conclusiones y Trabajo Futuro

La evaluación llevada a cabo permite contestar a las dos preguntas planteadas en la Introducción. Por un lado, filtrar una base de conocimiento de opinión, como SWN, con una lista de palabras de opinión, como iSOL (Algoritmo 4), es beneficioso para la posterior clasificación de la polaridad. Por otro lado, y lo más relevante, que la combinación de dicho filtro, con las polaridades propias de la lista de palabras (Algoritmo 5) mejora tanto la clasificación que proporcionan por separado tanto la lista de opinión, como la base de conocimiento. Por consiguiente, se puede afirmar que los resultados han demostrado la validez de CRiSOL.

La afirmación anterior es el punto de partida de una serie de nuevos trabajos. Sin considerar la revisión manual de CRiSOL, la investigación que se está llevando en este momento se centra en el estudio de la manera óptima de insertar el conocimiento de SWN en CRiSOL. Asimismo, se está evaluando la posibilidad de incluir bases de conocimiento adicionales.

En AO es de vital importancia la consideración del dominio en el que se circunscriben las opiniones. En (Molina-González et al., 2014) se expone un método de adaptación de listas de palabras de opinión a un dominio concreto, el cual será tenido en cuenta para añadir a CRiSOL información relativa a un dominio determinado.

Bibliografía

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2004. The meaning multilingual central repository. En *GWC 2012 6th International Global Wordnet Conference*. Brno: Masaryk University.

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 2200–2204, Valletta, Malta.

Boldrini, E., A. Balahur, P. Martínez-Barco, y A. Montoyo. 2009. Emotiblog: a fine-grained model for emotion detection in non-traditional textual genres. En *WOMSA*, páginas 22–31.

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-linguistic sentiment analysis:

From english to spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54, Borovets, Bulgaria, September. ACL.

Cambria, E. y A. Hussain. 2012. *Sentic Computing*, volumen 2 de *SpringerBriefs in Cognitive Computation*. Springer Netherlands.

Cruz, F., J. A. Troyano, F. Enriquez, y J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, 41:73–80.

Cruz, Fermín L., Jose A. Troyano, Beatriz Pontes, y F. Javier Ortega. 2014. Ml-senticon: a multilingual, lemma-level sentiment lexicon. *Procesamiento del Lenguaje Natural*, 53:113–120.

Gonzalez-Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual central repository version 3.0. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Kennedy, A. y D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y J. M. Perea-Ortega. 2014. Integrating spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y L. A. Ureña López. 2013. Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 87–93, Atlanta, Georgia. ACL.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ureña-López. 2014. Cross-domain sentiment analysis using spanish opinionated words. En *Natural Language Processing and Information Systems*, volumen 8455. Springer International Publishing, páginas 214–219.

Molina-González, M. D., E. Martínez-Cámara, María T. Martín-Valdivia, y J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volumen 10 de *EMNLP '02*, páginas 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Prabowo, R. y M. Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143 – 157.

Rangel, I. D., G. Sidorov, y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein*, 1(29):31–46.

Tsytsarau, M. y T. Palpanas. 2012. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514.

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, páginas 417–424, Stroudsburg, PA, USA. ACL.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks*, 5(2):241 – 259.