Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 45-52 recibido 23-11-14 revisado 27-01-15 aceptado 16-02-15

Esquema de anotación para categorización de citas en bibliografía científica

Annotation scheme for citation classification in scientific literature

Myriam Hernández Álvarez
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Quito, Ecuador
myriam.hernandez@epn.edu.ec

José Gómez Soriano
Universidad de Alicante
Dpto. de Lenguajes y Sistemas Informáticos
Alicante, España
jmgomez@ua.es

Resumen: El análisis de citas bibliográficas que usa variaciones de métodos de conteo provoca deformaciones en la evaluación del impacto. Para enriquecer el cálculo de los factores de impacto se necesita entender el tipo de influencia de los aportes de un investigador sobre el autor que los menciona. Para ello, se requiere realizar análisis de contenido del contexto de las citas que permita obtener su función, polaridad e influencia. El presente artículo trata sobre la definición de un esquema de anotación tendiente a la creación de un corpus de acceso público que sea la base de trabajo colaborativo en este campo, con miras al desarrollo de sistemas que permitan llevar adelante tareas de análisis de contenido con el objetivo planteado.
Palabras clave: Análisis de citas bibliográficas, análisis de contenido, esquemas de clasificación de citas, anotación de corpus, función, polaridad, influencia.

Abstract: Citation analysis that uses counting methods causes deformations in impact factor assessment. To enrich impact factor calculation is necessary to understand the kind of influence that the contributions of an author have over another´s work. For this purpose, it is required to perform citation content analysis to obtain its function, polarity and influence in a context within an article that mentioned it. In this paper, we focus in the definition of an annotation scheme aimed at creating a public access corpus that be the basis of collaborative work in this field, in order to develop citation content analysis to obtain criteria for impact evaluation.
Keywords: Citation analysis, content analysis, citation classification schemes, corpus annotation, function, polarity, influence.

1 Introducción

El análisis de citas bibliográficas en artículos científicos sirve para evaluar el impacto de un autor, de su obra o, incluso, de la revista en la que publica. El método actual de medición utiliza, básicamente, métodos cuantitativos relacionados con el conteo de las citas (Garfield, 1972), aunque también se utilizan ciertas variaciones como el PageRank, que es un algoritmo que tiene en cuenta la propia relevancia del que cita, (Page, et al., 1999) o la cocitación (Small, 1973), que añade como medida de similitud entre dos trabajos el número de documentos comunes que los citan.

Está bien documentado el hecho de que los métodos que usan criterios puramente cuantitativos provocan deformaciones en la percepción del impacto y la relevancia de los artículos citados. Un artículo con un alto número de citas no necesariamente es un artículo correcto y sus resultados tampoco tienen por qué estar confirmados por los investigadores que lo citan, puesto que muchas de las citas pueden ser críticas o manifestar algún tipo de rechazo. Radicchi (2012) mostró

ISSN 1135-5948

que artículos incompletos, erróneos o controvertidos tienden a tener un número de citas mayor. Los investigadores pueden estar tentados a publicar este tipo de obras para recibir un mayor número de referencias (Marder, Kettenmann y Grillner, 2010) o utilizar prácticas poco éticas para adquirir relevancia. Por ejemplo, Van Noorden (2013) explicó el caso de cinco revistas brasileñas que usaron auto citas y citas cruzadas para sesgar el índice del Journal Citation Reports. El premio Nobel Randy Schekman, en una publicación en The Guardian[1] (Sample, 2013), denunció estas prácticas por parte de prestigiosas revistas científicas que prefieren novedad y polémica antes que trabajo científico serio. En la misma publicación[1], el editor en jefe de Nature declaró que muchas veces a lo largo del tiempo ha manifestado su preocupación respecto a los peligros que conlleva un exceso de confianza en los factores de impacto basados en conteos de citas.

Estos análisis basados en contar el número de citas o variaciones de esta técnica, no toman en cuenta el tipo de influencia de un autor sobre otro (Zhang, Ding, y Milojević, 2013). Para entender esta influencia se requiere conocer la disposición del autor hacia el artículo citado y la función de la cita en el artículo que la menciona. No todas las citas tienen el mismo efecto en el artículo que las cita. El impacto de un artículo citado puede variar considerablemente si se toma en cuenta que la referencia contiene una crítica, es el punto de partida de un trabajo o si, simplemente, reconoce el trabajo de otros autores. Por esta razón, se vuelve importante identificar métricas más completas que tomen en cuenta el contenido de lo que se dice sobre la obra citada para evaluar su impacto y relevancia. Para ello, se hace necesario realizar un análisis de contenido del texto que contiene las citas para obtener ciertas características que puedan ser aplicadas a las actuales métricas para mejorar el cálculo bibliométrico de los investigadores y revistas. Se requiere la construcción de un índice más complejo que tenga en cuenta la intención del autor y su disposición hacia el trabajo que cita para determinar el impacto de éste de forma más precisa y analizando más factores que, únicamente, el número de citas recibidas. Intención y disposición son criterios que se relacionan con la función y la polaridad de la cita que forman parte del análisis de contenido.

Athar (2014) demostró que, para determinar tanto la función como la polaridad de una cita bibliográfica, se tienen mejores resultados cuando se analiza no solo la oración que contiene la cita sino también oraciones anteriores y/o posteriores que forman parte de un contexto. Este contexto debe ser definido de forma dinámica detectando las oraciones adyacentes a la cita que tienen algún argumento sobre ella. Sin embargo, el reconocimiento automático de argumentos en textos es todavía una tarea que presenta grandes retos lo que obstaculiza la detección automática del contexto de una cita. Por otra parte, la estructura del artículo y el sitio en el que se encuentra la cita también pueden servir para definir su función. Un artículo referenciado en la introducción probablemente sea una cita superficial, mientras que un artículo nombrado en la sección en la que se describe la metodología o los resultados, tiene una mayor probabilidad de cumplir con una función clara dentro del artículo que cita (White, 2004).

Dentro de este marco se ha definido que, para poder desarrollar la investigación en este campo, se requiere la generación de un corpus que clasifique las citas en forma estándar, que esté públicamente disponible y que permita el trabajo colaborativo en el área de Análisis de Contenido de referencias bibliográficas (Hernández y Gómez, 2014). Esta información permitirá obtener nuevos factores que podrán ser incorporados en el cálculo de un índice de impacto más completo, preciso y justo que permita evaluar de mejor forma la influencia de los artículos citados y que evite incentivos perniciosos. Para llevar a cabo esta ambiciosa tarea, es necesario que existan ciertos recursos para que los investigadores puedan avanzar y

[1]http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

poner a prueba sus sistemas. Uno de los más cruciales es la construcción de un corpus etiquetado que sirva de gold standard y que sea lo suficientemente grande para que las evaluaciones puedan dar resultados estadísticamente significativos.

Con este objetivo en mente, se ha definido un esquema de anotación para este corpus con las siguientes consideraciones: que contemple factores que se requiere incluir en un sistema de evaluación de citas bibliográficas: la clasificación de la influencia y la función de la cita y el análisis de la polaridad; que sirva como estándar de anotación en esta área de investigación; que pueda ser etiquetado por personas que no sean especialistas en el tema del artículo que se anota ni relacionadas con las Tecnologías del Lenguaje Humano; que clasifique de la manera más clara posible todas las citas del artículo, tomando en cuenta su función, influencia y polaridad; que elimine la máxima ambigüedad posible evitando el solapamiento de las categorías; que permita una construcción ágil de un corpus de citas bibliográficas anotadas que facilite una población lo suficientemente grande para que sea representativo y que sirva como gold standard; que se encuentre en el punto óptimo de granularidad y utilidad; que permita un etiquetado semi- supervisado y que realice un marcado lingüístico del texto para obtener reglas que faciliten trabajos posteriores.

2 Esquema de anotación

Para delimitar un esquema de anotación se hace necesario considerar que hay, al menos, dos enfoques a la categorización de funciones de las citas. El primero define que un esquema es útil en la medida en que es exhaustivo en su profundidad y granularidad. Un ejemplo de este enfoque es el esquema de 35 categorías de Garzone (1996). El segundo enfoque es el tomado por Teufel y Moens (1999) que cuestiona esta categorización de grano fino pues asevera que la mayoría de instancias son difíciles de detectar porque no se encuentran pistas lingüísticas evidentes que puedan servir para clasificarlas y que aún si están presentes claves explícitas, el problema de detectarlas es formidable. Teufel, y Moens (1999) también expresan que juzgar la naturaleza de la cita conlleva un alto nivel de subjetividad y que hay una ausencia de medios para mapear esa naturaleza a los propósitos de la cita.

En el punto medio de granularidad, nuestro esquema está diseñado para especificar características que pueden definir impacto de una cita tomando en cuenta su función, polaridad e influencia dentro del texto que la referencia. Con los resultados preliminares de la anotación veremos la relación entre la clasificación de acuerdo a los tres criterios. Tratamos que cada categoría sea fácilmente diferenciable de las demás para que los anotadores humanos puedan distinguirlas y para facilitar la generación de un corpus anotado que permita que se lo siga aumentando en forma manual o automática.

Para aplicar este diseño es necesario definir el contexto de la cita que debe incluir lo más posible lo que se dice sobre la cita. Debido a la complejidad que presenta la búsqueda de argumentos para definir un contexto, se resolvió establecer el contexto con la longitud de un párrafo. Para tomar esta decisión partimos del hecho de que, por definición, un párrafo es el conjunto de oraciones que expresan una idea o argumento completo, por lo tanto, un párrafo tiene una buena posibilidad de incluir la argumentación relevante sobre una cita. Esto funciona bien, y mantiene el criterio de que el contexto varíe de acuerdo a la presencia de argumentos en torno a la referencia.

Para construir un corpus gold-standard con información extra que facilite la obtención de reglas que sirvan de guía en un proceso de anotación manual o automático, proponemos una nueva metodología de anotación. El codificador detecta estructuras sintácticas fijas (palabras clave) y variables (patrones en forma de etiquetas XML) en las oraciones y forma patrones que le ayudan a reconocer las categorías del esquema. Se ha comprobado experimentalmente que este tipo de anotación aclara dudas en los codificadores y permite una mayor coherencia en las anotaciones y un porcentaje de acuerdo más alto. Los datos de patrones y palabras clave (skip-grams) se guardan y ofrecen información adicional para anotaciones sucesivas. Esta información

facilita la labor de los codificadores, puesto que les sirve como ejemplos adicionales que aclaran los casos que se van presentando y permiten una definición objetiva de las diferencias entre funciones. Esta información de patrones y palabras clave también será la base para un sistema de aprendizaje que automáticamente genere un corpus más extenso, a través de su uso como reglas en un modelo de aprendizaje semi-supervisado.

Con el objetivo mencionado, se delinean dos pasos para la anotación. En el primero se pide al anotador que lea el párrafo y establezca las partes variables y fijas que detecta; marcando solamente lo que le parezca indispensable para definir la estructura de la oración y lo que se dice sobre la cita. Estos patrones deben ser lo más simple posible, por lo que se le pide al anotador que solamente marque las partes básicas. Cuando se han establecido estos patrones, se le sugiere al anotador que revise la información disponible respecto a palabras clave y partes variables. Con estos datos, resulta mucho más fácil, tomar la decisión respecto a la clase a la que pertenece la cita.

El esquema de clasificación usado, se presenta en las Tablas 1, 2, 3 y en la Figura 1.

Función de la cita	Descripción
Based on, supply	El artículo que cita se construye sobre material del artículo citado que puede ser un concepto o herramienta. El artículo citado es usado en el artículo que cita.
Useful, Standard	El material del artículo citado (concepto o herramienta) se aplica en algún otro trabajo, no en el propio. El trabajo citado se relaciona con una idea usada como medida, norma o modelo.
Acknowledge, Corroboration, Positive contrast	El artículo citado se menciona para reconocer algún trabajo previo. El artículo que cita confirma o soporta algún aspecto de la cita. Se hace contraste positivo.
Weakness, correct, negative contrast	Se nota o corrige un error o debilidad del trabajo citado. Se establece un contraste negativo.
Weakness, Hedges	Uso de lenguaje cuidadoso para ocultar una disposición negativa hacia el trabajo citado.

Tabla 1: Esquema de anotación para funciones

Influencia	Descripción
Significant	La cita es importante para el trabajo que la referencia. El trabajo citado, por alguna razón merece atención.
Perfunctory	La cita no cumple un rol importante en el artículo que la menciona. La cita se hace para reconocer trabajo previo.

Tabla 2: Esquema de anotación para influencia

Polaridad	Descripción
Positive	El autor tiene una disposición favorable hacia el trabajo citado.
Negative	El autor tiene una disposición no favorable hacia el trabajo citado.

Tabla 3: Esquema de anotación para polaridad

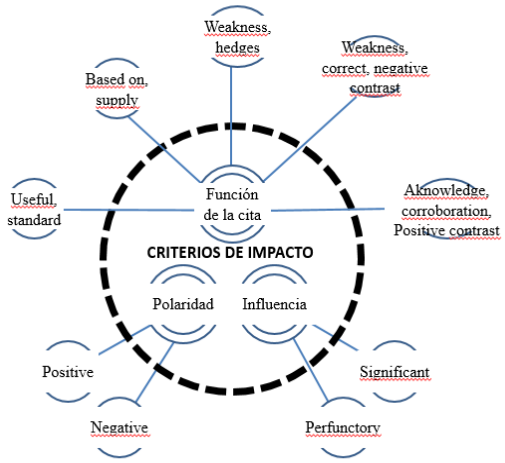

Figura 1: Criterios para evaluar el impacto de una cita de acuerdo a su función, polaridad e influencia

La lógica de la clasificación tiene que ver con funciones con definición de clases

claramente separadas que cubren las posibilidades existentes y que además se van a relacionar con medidas de impacto. En orden de evaluación de impacto, considerando la profundidad de la relación entre el trabajo que cita y el trabajo citado, tendremos los grupos de funciones: "Based on, supply", "Useful, Standard", "Acknowledge, Corroboration, Positive Contrast"; valores negativos de impacto podrían tener las funciones "Weakness" y "Weakness, Hedges". Juntamos categorías que reflejan similar importancia de la cita para el artículo que la referencia; así disminuimos la complejidad del esquema y logramos que continúe sirviendo a nuestro objetivo que es el análisis de funciones en relación con la evaluación del impacto.

"Weakness, Hedges" es una categoría interesante. Esta clase es un caso especial de "Weakness" en el que se presenta una forma de lenguaje cauteloso para ocultar una disposición negativa respecto al artículo citado, con el fin de evitar reacciones no deseadas por parte de los afectados. Detectar la presencia de "hedges" permitirá descubrir citas que implican posiciones y polaridades negativas encubiertas de diversas maneras. Para ello nos basamos en un análisis realizado por Hyland (1996).

2.1 Ejemplos

Usando los patrones para clasificación de citas, se pueden obtener los ejemplos para cada función. Al momento hemos considerado solamente referencias bibliográficas en idioma inglés, puesto que es el lenguaje común de la ciencia; sin embargo, opinamos que los patrones, las etiquetas y la clasificación, podrían ser aplicados también en otros idiomas; por supuesto sería necesario tomar en cuenta características especiales de los distintos idiomas y multilingüismo. Veamos un ejemplo de patrones en inglés:

Texto: "Argumentative Zoning (Teufel, 1999; Teufel and Moens, 2002) attempts to solve this problem by representing the structure of a text using a rhetorically-based schema. We used another technique". Patrón: METHOD (CITE) attempts to solve * METHOD. AUTHOR used another technique. Clasificación de la cita, función: Weakness, Hedges; polaridad: Negativa. En este último caso, al usar la secuencia de palabras "attempts to solve" se podría indicar que intentaron resolver un problema, pero no lo consiguieron y por ello se prefirió usar otro método. Muestra velada negatividad hacia la cita.

2.2 Validez del esquema

Un requerimiento indispensable para respaldar la calidad de un modelo como el discutido, requiere demostrar la confiabilidad de los datos obtenidos. La confiabilidad de los datos se relaciona con la confiabilidad del proceso de anotación. Para evaluar este parámetro con varios codificadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que ha sido revisado por los mismos anotadores. Para poder generalizar los resultados obtenidos en esta muestra a todo el proceso en el que probablemente van a intervenir nuevos anotadores y no solo los que codificaron la muestra, se necesita que el proceso sea confiable (Artstein y Poesio, 2008).

Según Krippendorff (2004) la confiabilidad o reproducibilidad de la anotación se asegura cumpliendo tres requerimientos: un esquema claro con instrucciones detalladas y criterios específicos para escoger codificadores. Para medir reproducibilidad se deben tener al menos tres anotadores que deben trabajar independientemente entre sí.

En nuestro experimento se cumplieron estos tres requerimientos. Se propuso un esquema claro, detallado y con suficientes ejemplos de aplicación; los anotadores son personas que lo han revisado cuidadosamente y tienen conocimientos del área de lingüística computacional; y, por último, para anotar la muestra se tuvieron tres codificadores que trabajaron en forma separada. Se pidió a los anotadores que sigan en forma consistente un procedimiento claramente establecido, en el cual se realiza primeramente una pre anotación con los patrones y palabras claves, de la forma como se ha explicado.

3 Organización de los experimentos y datos

Se usaron artículos del archivo de la Asociación de Lingüística Computacional (ACL por sus siglas en inglés), escogidos de forma aleatoria. Los textos se pre procesaron para marcar párrafos para detectar contexto. Para validar el modelo y calcular el acuerdo entre anotadores se utilizaron citas en artículos que fueron etiquetados de forma independiente por tres personas con conocimientos en el campo de lingüística computacional. Los datos usados para el cálculo del acuerdo entre anotadores contienen 101 citas, una variable para función, una variable para polaridad y una variable para influencia con 303 decisiones cada una.

El proceso de anotación se realizó en varias etapas. La primera consistió en un proceso de pre anotación para reconocer y numerar las citas en el texto. Para ello, se utilizó un programa desarrollado por el grupo de investigación de PNL de la Facultad de Ingeniería de Sistemas de la Escuela Politécnica Nacional de Quito. Este programa reconoce expresiones regulares asociadas a las referencias bibliográficas en el formato oficial de la Antología de la ACL[2]. La segunda etapa consistió en reconocer patrones en el texto. La información de las etiquetas de los patrones en el texto, guían al anotador en la clasificación de la función y la polaridad. Por último, se realizó un procesamiento de cada uno de los artículos, para definir el número de veces que la referencia fue nombrada y el sitio en la cual se la mencionó: introducción, método, resultados, discusión o sus equivalentes. Estos datos se usan para definir el tipo de influencia de la cita. Para estos últimos pasos se usó el editor de NetBeans IDE 8.02 para formato XML. Las anotaciones de cada codificador se guardaron en archivos de texto separados, cuya primera línea correspondía a los nombres de las variables, es decir la función, la polaridad y la influencia, separadas por una tabulación; y, las siguientes líneas fueron los resultados de las anotaciones para cada cita. Estos archivos fueron cargados en el programa desarrollado por Geertzen, J., 2012, para obtener el cálculo del nivel de acuerdo entre anotadores.

4 Resultados y discusión

Artstein y Poesio, 2008 dicen que los datos son confiables si se muestra que los anotadores entendieron las categorías asignadas y por tanto producen en forma consistente resultados similares. De este modo, la confiabilidad es un requisito para demostrar la validez de un esquema. Si los codificadores no muestran acuerdo entre sí, puede deberse a que algunos de ellos están equivocados o a que el esquema de anotación no es apropiado para los datos. Adicionalmente, la confiabilidad implica la habilidad de distinguir entre las clases, la que se posibilita si el esquema es claro. El acuerdo entre anotadores puede evaluarse utilizando coeficientes que toman en cuenta la corrección por la probabilidad de que los codificadores estén de acuerdo en un ítem simplemente por azar. Artstein y Poesio (2008) sugieren que los coeficientes se escojan de acuerdo a la tarea. En el caso de que la anotación tenga un sesgo hacia algunas(s) categorías, de acuerdo a la recomendación de los autores mencionados, para la evaluación de confiabilidad se debe usar el coeficiente de Krippendorff. Obtenemos valores para varios coeficientes incluido el mencionado. Las Tablas 4, 5 y 6 se muestran los resultados.

Fleiss	Krippendorff	Pairwise avg.
A_obs=0.845 A_esp=0.365 Kappa=0.756	D_obs = 0.155 D_esp = 0.637 Alpha = 0.756	% agr = 84.5 Kappa=0.756

Tabla 4: Acuerdo entre anotadores con pre anotación correspondiente a la Función

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.96 A_exp = 0.72 Kappa=0.859	D_obs = 0.04 D_exp = 0.281 Alpha = 0.859	% agr = 96 Kappa=0.86

Tabla 5: Acuerdo entre anotadores con pre anotación correspondiente a la Influencia

[2] http://www.aclweb.org/anthology/

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.98 A_exp=0.776 Kappa=0.912	D_obs = 0.02 D_exp = 0.225 Alpha = 0.912	% agr = 98 Kappa=0.913

Tabla 6: Acuerdo entre anotadores con pre anotación correspondiente a la Polaridad

Se obtienen los coeficientes: Kappa de Fleiss (Fleiss, 1971), Alpha de Krippendorf (Krippendorff, 2004) y Kappa para el promedio por pares. Se utilizó el software de Geertzen, J. (2012) para el cálculo de los coeficientes. A_obs es el acuerdo observado, A_exp es el acuerdo esperado, D_obs es el desacuerdo observado y D_exp es el desacuerdo esperado. Los valores bajos en los coeficientes nos indicarían que los anotadores tuvieron problemas para distinguir entre categorías y lo contrario validaría la claridad del esquema, como sucede en nuestro caso.

De acuerdo a los resultados, los valores para la clasificación de la polaridad pueden ser mapeados directamente de las funciones. Son funciones positivas: "Based on, Supply", "Useful, Standard", "Acknowledge, Corroboration, Positive Contrast". Son negativas: "Weakness" y "Weakness Hedges".

Los resultados demostraron que la influencia de la cita tiene que ver con el número de veces que se la menciona en una sección del artículo diferente a la Introducción. Los artículos que se califican como superficiales aparecen principalmente en la Introducción, son fundamentales las citas que tienen una función "Based on, Supply" o se citan más de dos veces. Con estos criterios, una buena aproximación para la clasificación de la influencia, podría realizarse en forma automática con los datos de las funciones y de la ubicación de la cita y esta clasificación no requeriría ser anotada manualmente.

5 Conclusiones y trabajo futuro

Un requerimiento indispensable para respaldar y probar un modelo como el discutido, es la demostración de confiabilidad en los resultados obtenidos. La confiabilidad de los datos tiene que ver con la confiabilidad del proceso de anotación. Para medir la confiabilidad en la anotación del esquema con varios codificadores, es necesario medir el acuerdo obtenido en una pequeña muestra del corpus que ha sido revisado por las mismas personas. De esta manera, el modelo del esquema se valida a través del resultado que lo califica como reproducible.

En la clasificación de funciones de las citas, el porcentaje de acuerdo, sin tomar en cuenta una corrección por acuerdos al azar, es de 84.5%. Con la respectiva corrección incluida en el cálculo de Kappa, se tiene un K = 0,756. De acuerdo a Landis y Koch (1977), se puede concluir que esta magnitud de Kappa corresponde a un substancial acuerdo entre anotadores, nivel que se considera para Kappa entre 0,6 y 0,8. El acuerdo entre anotadores es incluso mayor para la clasificación de polaridad e influencia. La polaridad tiene un K = 0,912 y la influencia un K = 0,859. Los mismos autores (Landis y Koch, 1977) califican a los resultados para Kappa, que van entre 0,8 y 1,0, como perfectos. Se esperaban valores mayores de Kappa para clasificaciones binarias, como las realizadas para Influencia y Polaridad, porque puede relacionarse directamente la precisión de los anotadores con el número de clases entre las que tienen que decidir. A partir de estos resultados, el esquema de clasificación propuesto y la metodología de anotación fueron validados a satisfacción.

Una de las contribuciones de este trabajo, es la de presentar un esquema y metodología de anotación que permitirán el desarrollo de un corpus base, que podrá ser extendido a través de aprendizaje automático o incluso de métodos manuales.

Como trabajo futuro nos planteamos usar este esquema de clasificación para construir un corpus anotado de acceso público, que pueda servir a la comunidad científica para el desarrollo de sistemas enfocados a evaluar el factor de impacto en bibliografía científica utilizando criterios adicionales obtenidos a partir del análisis de contenido del contexto de las citas.

La experimentación que confirma que el

esquema planteado es relevante para tareas de aprendizaje automático está en etapa de desarrollo y los resultados serán presentados en un artículo posterior. Se están usando dos técnicas: aprendizaje supervisado con reglas desarrolladas como expresiones regulares formadas por los patrones en las anotaciones; y, SVM utilizando como características las etiquetas de los patrones.

Bibliografía

Artstein, R., y Poesio, M. 2008. Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555-596.

Athar, A. 2014. Sentiment analysis of scientific citations. Technical Report, University of Cambridge.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378–382.

Garfield, E. 1972. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science 178: 471–9.

Garzone, M. A. 1996. Automated classification of citations using linguistic semantic grammars. Master's thesis, The University of Western Ontario.

Geertzen, J. 2012. Inter-Rater Agreement with multiple raters and variables. Retrieved October 8, 2014, from https://mlnl.net/jg/software/ira/.

Hernández, M., y Gómez, J. M. 2014. Survey in sentiment, polarity and function analysis of citation. In Proceedings of the First Workshop on Argumentation Mining ACL 2014, Baltimore, MD, pp. 102–3.

Hyland, K. 1996. Writing without conviction? Hedging in science research articles. Applied linguistics, 17(4), 433-454.

Krippendorff, Klaus. 2004. Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3):411–433.

Landis, J. R., y Koch, G. G. 1977. The measurement of observer agreement for categorical data. Biometrics, 159-174.

Marder, E., Kettenmann, H., y Grillner, S. 2010. Impacting our young. Proceedings of the National Academy of Sciences of the United States of America 107: 21233.

Page, L., Brin, S., Motwani, R., y Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab.

Radicchi, F. 2012. In science "there is no bad publicity": Papers criticized in comments have high scientific impact. Scientific Reports 2: 815.

Sample, I. 2013. Nobel winner declares boycott of top science journals. The Guardian. Available at http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals

Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science 24: 265–9.

Teufel, S., y Moens, M. 1999. Discourse-level argumentation in scientific articles: Human and automatic annotation. In M. Walker (ed.), Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop, pp. 84–93. Somerset, NJ: Association for Computational Linguistics.

Van Noorden, R. 2013. Brazilian citation scheme outed. Nature, 500(7464), 510–1.

White, H. D. 2004. Citation analysis and discourse analysis revisited. Applied linguistics, 25(1), 89-116.

Zhang, G., Ding, Y., y Milojević, S. 2013. Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology, 64(7), 1490-1503.