ISSN 1870-4069

Extracción automática de eventos indicadores a partir de noticias en español

Ariatna Quinto[1], Belém Priego[1], David Pinto[2], José A. Reyes-Ortiz[1]

[1]Universidad Autónoma Metropolitana unidad Azcapotzalco,
Departamento de Sistemas, México

[2]Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, México

{al2113034676,abps,jaro}@azc.uam.mx,dpinto@cs.buap.mx

Resumen. Actualmente, se ha notado el incremento exorbitante de la información electrónica, claro ejemplo es la Web, la cual se ha convertido de fácil acceso. Ésta es posible estudiarla para saber los fenómenos que suceden a nivel de la lengua y a partir de ella se puede extraer meta información. En este artículo se han seleccionado artículos periodísticos en formato electrónico, como corpus, para llevar a cabo la extracción automática de eventos indicadores que representan la proporcionalidad de un evento con respecto a un total, es decir, porcentajes relacionados a algún suceso. El principal objetivo de esta investigación es mostrar estadísticas que ocurren alrededor del mundo mediante la búsqueda de patrones lingüísticos en un sistema de recuperación de información. Este artículo presenta los avances obtenidos hasta el momento, en la extracción automática de eventos indicadores a partir de noticias en español.

Palabras clave: Extracción automática, recuperación de información, eventos indicadores.

Automatic Extraction of Indicative Events from Spanish News Stories

Abstract. Nowadays there exist an increasing raising of information on Internet which is easy available for human beings. This huge volume of data can be analyzed in order to discover and model linguistic phenomena and extract information. In this paper we have selected the news stories genre for extracting indicative events that represent likelihood of some event to occur. The aim of this research is to bring to light statistical events occurring all around the world by employing techniques of natural language processing based on linguistic patterns in an information retrieval system. This paper presents the outcomes obtained up to now in the particular topic of automatic extraction of indicative events from Spanish news stories.

Keywords. Automatic extraction, information retrieval, indicative events.

1. Introducción

Procesamiento de Lenguaje Natural (denotado por PLN), es una disciplina de la Inteligencia Artificial que trata la formulación e investigación de mecanismos de computación para la comunicación entre personas y máquinas, mediante el uso de Lenguajes Naturales. Dichos lenguajes son utilizados para la comunicación ya sea de forma escrita, hablada o en forma de signos [4]. Entre las tareas que realiza el PLN se encuentra la extracción automática de eventos, cuyo objetivo es capturar ciertas partes relevantes de un texto.

En el análisis del lenguaje se estudia la estructura del lenguaje a cuatro niveles [4]:

- Análisis morfológico: El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos.
- Análisis sintáctico. El análisis de la estructura sintáctica de la frase mediante una gramática de la lengua en cuestión.
- Análisis semántico. La extracción del significado (o posibles significados) de la frase.
- Análisis pragmático. El análisis de los significados más allá de los límites de la frase, por ejemplo, para determinar los antecedentes referenciales de los pronombres.

Este artículo, pretende abordar la tarea de la extracción automática de eventos. En particular, los eventos serán los indicadores que representan la proporcionalidad de un evento con respecto a un total que aparecen en una colección de información (noticias) en español, es decir, éstos son los porcentajes relacionados con algún suceso. Para el desarrollo total de esta investigación se hará uso únicamente de tres niveles del lenguaje, el morfológico (etiquetado POS), el sintáctico (segmentación del párrafo en oraciones) y el semántico (patrones lingüísticos). Además, se ha seleccionado el género periodístico debido a que es un tipo de escritura estándar y homogénea, en otras palabras, cualquier hablante nativo que lea el contenido de una nota periodística, lo entiende. Inclusive, la mayoría de las personas tiene acceso a un periódico, en formato papel o digital. Sin embargo, dado que es una investigación que está iniciando en este artículo, se presentarán los avances obtenidos al tratar de extraer de manera automática eventos indicadores a partir de noticias en español.

2. Motivación

Debido a la basta cantidad de información que actualmente se encuentra en la Web, ésta se puede utilizar y procesar de modo que se pueda emplear para ciertas tareas del PLN, una de ellas es la extracción de información relevante de un texto. Además de que es posible extraer conocimiento de toda esta información. Uno de los medios de comunicación que proporciona información es el Periódico, que especialmente en los últimos años su acceso ha sido en formato digital; esto

debido a los avances tecnológicos, como es internet. Razón por la cual, en esta investigación se ha decidido trabajar con este género textual.

A través de los periódicos se ha podido llegar incluso a más gente y mantener un ritmo de actualización de los datos mucho más intenso que antes, siendo hoy imposible esperar de un día para otro para conocer noticias. Lo interesante de los periódicos es que cuando se habla de una sociedad, más o menos compleja, se pueden encontrar distintos tipos de periódicos que dan con el perfil de grupos sociales particulares, de grupos de edad, de regiones geográficas, de actividades laborales, de intereses específicos como deportes, internacionales, espectáculos o política.

Debido a que la sociedad cada vez vive de una manera más acelerada, se ha decidido extraer información de las noticias y a través de esta extracción dar a conocer datos estadísticos que la búsqueda de eventos indicadores proporcionará como resultado. Lo que servirá para informar de manera más concisa y directa cierto suceso, esta información representará la proporcionalidad de un evento, sin la necesidad de que las personas lean toda la nota periodística, o más notas, para conocer datos importantes y relevantes. Al finalizar la investigación, se pretende crear un sistema web capaz de mostrar los resultados de la extracción de eventos indicadores. Dicho sistema servirá como medio de información resumida, lo cual beneficiará a la sociedad que desea estar informada.

3. Descripción del problema

En el Diccionario de la Real Academia [9], una de las acepciones para "evento" dice que es un suceso importante. De esta manera, este artículo retoma esa idea para mostrar un suceso importante que surge alrededor del mundo a través de los relatos periodísticos que se encuentran en la web. Sin embargo, se sabe que dar una definición concisa de evento es difícil. Por lo tanto, en esta investigación se tratará a un suceso como cualquier tipo de situación o acontecimiento que ocurre, restringiendo a eventos relacionados con la proporcionalidad de uno con respecto a un total, es decir porcentajes relacionados a un suceso. En (1) y (2) se pueden observar enunciados que muestran eventos indicadores.

(1) En México, cerca del 88 % de la energía primaria que se consume proviene del petróleo.

(2) La mitad de la población mundial está concentrada en tan solo seis países.

En (1) el evento indicador de porcentaje está explícitamente determinado por su signo ortográfico. Mientras que en (2) se nota un evento indicador, la mitad de la población mundial, pero no está explícitamente denotado por un signo ortográfico. Por lo que la tarea de extracción de eventos indicadores se convierte más complicada a medida que el lenguaje se desarrolla con diferentes factores de pensamiento, es decir, a medida que el lenguaje es abundante, enriquecido y creativo, la tarea se complica.

Típicamente, los eventos pueden ser expresados por verbos conjugados o en infinitivo, predicados en general y frases preposicionales. Sin embargo, como se ha notado, en esta investigación no se cumple con esos acuerdos, ya que se pueden

encontrar eventos expresados mediante un signo de puntuación (%), partes que dividen un todo (mitad, tercio, etc.), entre otras expresiones. Sobre un texto plano, se pretende identificar eventos indicadores mediante diferentes patrones lingüísticos que serán identificados y son los que ayudaran a la extracción de diferentes sucesos acontecidos alrededor del mundo. Se puede delimitar entre las etiquetas <*indicador*> y <*/indicador*>, el inicio y fin del indicador porcentual con el fin de poder extraer toda la idea que contiene al evento, en (3) se puede observar un ejemplo. Esto mediante un sistema que segmente a los párrafos en oraciones.

(3) La <*indicador*>mitad<*/indicador*> de la población mundial está concentrada en tan solo seis países.

En el contexto de este trabajo se contará con un esquema que establece reglas claras con guías de cómo se deben identificar los eventos indicadores a partir de diferentes patrones lingüísticos, con el objetivo de reducir las ambigëdades al mínimo.

4. Estado del arte

En esta sección, se describen los trabajos reportados en la literatura relacionados con la extracción automática de eventos. Cabe mencionar, que se incluyen trabajos que extraen eventos, sin embargo, corresponden a otro tipo que difiere al presentado en este trabajo, pero se incluyen debido a la temática presentada.

La utilización de características sintácticas, semánticas y contextuales, minería de textos, es uno de los varios enfoques que se utilizan para la extracción de eventos. El trabajo de Hernández [3], aborda su investigación sobre dicho enfoque, diseñó e implementó un sistema web para la anotación semántica de actores y eventos a partir de un corpus de textos periodísticos mexicanos. Este trabajo tiene la particularidad principal, que la extracción la realiza en textos periodísticos mexicanos, sin embargo, son relatos periodísticos completamente diferentes a los que se abordan en este trabajo. El trabajo de Gil-Vallejo et al. [2], incluido en este enfoque, extrae eventos mediante las entidades y relaciones que posee el texto usando información lingística, relaciones sintácticas y semánticas, obteniendo un patrón de extracción de información relevante para un cierto evento. Los eventos que considera son los que contienen verbos dentro de su estructura de frase.

Las técnicas de aprendizaje automático, también, son consideradas para el reconocimiento de eventos. Un trabajo presentado por Moncecchi y Rosá [5], que entra en el marco de este enfoque, reconoce eventos en textos españoles utilizando como base el esquema de anotación SIBILA [14] y dos algoritmos de aprendizaje automático, campo aleatorio condicional (CRF, por sus siglas en inglés Conditional Random Fields) [11] y máquinas de soporte vectorial (SVM, por sus siglas en inglés Support Vector Machines) [10], logrando mejores resultados con SVM.

Además de investigaciones, existe el desarrollo de herramientas capaces de extraer eventos en un texto. JASPER [1] y TES [12], son un claro ejemplo de ello.

En la primera, JASPER: Journalist's Assistant for Preparing Earnings Reports [1], se extraen ciertas piezas clave de información de un rango limitado de texto. El sistema está basado en el uso de plantillas, técnicas de comprensión parcial y procedimientos heurísticos para extracción de información. Esta información, puede ser utilizada de varias maneras, como rellenar valores en una base de datos, generar resúmenes del texto de entrada, entre otras. Los eventos que principalmente extrae JASPER son los comunicados de prensa para generar historias de noticias. Con respecto a la segunda, TES: Terminology Extraction Suite [13], es una herramienta desarrollada para la extracción automática de terminología, que permite obtener términos y buscar automáticamente equivalentes de traducción. La herramienta está escrita en Perl, con interfaces gráficas implementadas en Tk. En este caso, los términos son eventos o sucesos que un documento, texto, contiene.

5. Metodología propuesta

En la extracción automática de eventos indicadores a partir de noticias en español, se han identificado dos etapas principales (ver Figura 1), las cuales serán alimentadas por un corpus de notas periodísticas en texto plano escritas en español,. A partir de estas notas se desarrollará la extracción automática de los eventos indicadores. La extracción se llevará a cabo mediante el diseño de patrones lingüísticos que cubrirán tres niveles de la lengua, morfológico, sintáctico y semántico. Finalmente, se pretende visualizar los resultados en un sistema web.

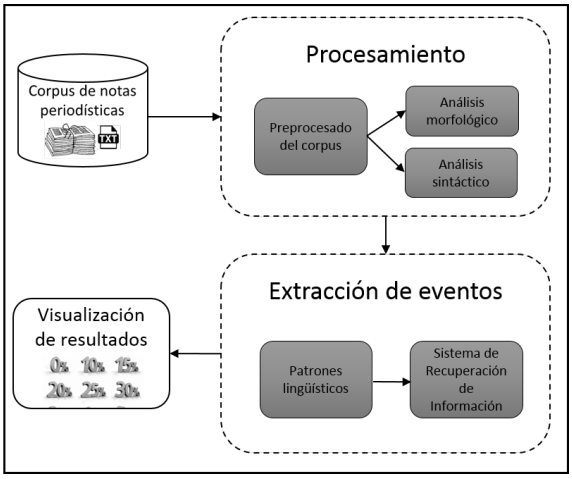

Fig. 1. Metodología propuesta para la extracción automática de eventos indicadores.

5.1. Conjunto de datos

En esta sección se describe el conjunto de datos, corpus periodístico, en español de notas periodísticas que será utilizado para la extracción automática de eventos indicadores. Cabe mencionar que la descripción realizada es general debido a que el corpus a utilizar es el realizado en [9].

El corpus ha sido extraído del sitio de intenet de la Organización Editorial Mexicana[1], OEM, que contiene relatos periodísticos escritos en español mexicano. A pesar de ser un sitio mexicano, no excluye las notas periodísticas internacionales pero de igual manera escritas en español mexicano. Los relatos periodísticos corresponden al período de tiempo del año 2007 al 2013.

Si bien, el corpus presenta diferentes metadatos, para el caso de este artículo, sólo se considerará el texto plano de la nota periodística. El corpus utilizado para esta tarea consta de 378,890 noticias, un total de 4,579,284 oraciones y alrededor de 11,1595,71 palabras.

5.2. Etapa de procesamiento

La etapa de procesamiento comprende el preprocesado del corpus, el análisis morfológico y sintáctico de dicho conjunto de datos.

La primera actividad, el preprocesado del corpus, contempla la eliminación de signos de puntuación y de palabras cerradas, mediante la utilización de un lexicón de éstos [8]. Además, se eliminan los caracteres especiales, poniendo atención a los caracteres relacionados con los eventos indicadores (%); estos caracteres a eliminar, se consideran irrelevantes para la extracción de los e entos indicadores debido a que ocasionarán conflicto al momento de recuperar la información que se está extrayendo y podrían incrementar el tiempo de respuesta.

La segunda actividad, el análisis morfológico, realiza el etiquetado de las partes de la oración (PoS de sus siglas en inglés, Part of Speech) en las notas periodísticas; se hace uso de las herramientas FreeLing [6] y/o TreeTagger[13]. El etiquetado consiste en identificar la categoría gramatical de cada palabra y asignarle una etiqueta dependiendo de la categoría gramatical a la que corresponda.

La tercera actividad, el análisis sintáctico, busca segmentar los párrafos que componen al corpus periodístico en oraciones, debido a que será más accesible la manipulación de oraciones y éstas tendrán una longitud más regular, es decir, el tamaño de un párrafo tiene menos proporcionalidad con respecto a una oración.

5.3. Etapa de extracción de eventos

La etapa de extracción de eventos incluye dos actividades esenciales, el diseño de los patrones lingísticos, para la extracción automática de eventos indicadores, y la búsqueda de éstos en un sistema de recuperación de información (denotado por SRI).

[1] Para más información sobre la Organización Editorial Mexicana consultar: https://www.oem.com.mx/oem/

La primera actividad, diseño de patrones lingísticos, permite realizar el diseño de los patrones lingísticos que, en una etapa posterior, se implementan con el fin de extraer los eventos indicadores. Este diseño de patrones permite descubrir los elementos lingüísticos empleados con frecuencia en las notas periodísticas, para ello se utiliza uno o varios modelos que sirven como muestra para identificar y agrupar los eventos indicadores.

La segunda actividad, búsqueda de los patrones lingísticos en un SRI, posibilita, una vez que se han identificado los patrones lingüísticos, la implementación de éstos. Es decir, mediante un SRI, alimentado con el corpus de notas periodísticas y como consulta los patrones lingüísticos identificados, extraer la información relevante relacionada a los eventos indicadores.

6. Resultados obtenidos

El proceso de extracción de patrones lingüísticos se ha llevado a cabo mediante una técnica conocida como bootstrapping. Se considera un conjunto inicial de muestras etiquetadas manualmente, las cuales son posteriormente enriquecidas usando muestras similares obtenidas mediante un sistema de recuperación de información. De esta manera, es posible obtener un conjunto considerable de datos que comparten una estructura morfosintáctica que permite obtener patrones lingísticos que muestran la regularidad de estructuras para un tipo de expresión lingística en particular.

En la Tabla 1 se muestran ejemplos de los patrones lingüísticos más comunes encontrados mediante este proceso de generalización basado en un conjunto inicial moderadamente pequeño de muestras manualmente etiquetadas, pero que fue enriquecido mediante la técnica anteriormente mencionada.

Tabla 1. Patrones lingüísticos más frecuentes para eventos indicadores usando explícitamente el porcentaje.

Patrón lingüístico	**Semántica**
NUM por ciento	Indicador puntual
alrededor del NUM por ciento	Indicador aproximado
mayor que NUM por ciento	Indicador de punto base
entre NUM y NUM por ciento	Intervalo
de NUM a NUM por ciento	Incremento
hasta un NUM por ciento	Máximo temporal
NUM por ciento más que	Indicador comparativo incremental
un incremento de NUM por ciento	Indicador comparativo incremental
NUM por ciento menos que	Indicador comparativo decremental
un decremento de NUM por ciento	Indicador comparativo decremental

Es importante aclarar que la etiqueta **NUM** se refiere a la especificación de un número en el texto periodístico en cualquiera de sus expresiones. Ejemplos de **NUM** serían los siguientes: 90, 28.3, noventa y tres, etc.

Por otro lado, en la Tabla 1, se ha usado el texto "por ciento", el cual puede ser encontrado también sustituido por el símbolo ortográfico "%" por lo que el lector debe considerar que los patrones anteriormente mencionados pueden ocurrir con cualquiera de estas dos expresiones textuales.

En la Tabla 2, se muestran otros patrones morfosintácticos que expresan el uso de números fraccionarios porcentuales, pero expresados en lenguaje natural.

Tabla 2. Patrones lingüísticos más frecuentes para eventos indicadores que usan otro tipo de expresiones del lenguaje natural relacionadas implícitamente con un porcentaje.

Patrón lingüístico	**Semántica**
un total de NUM de las/los NUM	Indicador porcentual basado en cociente
la mitad	Indicador 50 %
un tercio	Indicador 33 %
una cuarta parte	Indicador 25 %

Ejemplos de oraciones que contienen a algunos de los patrones presentados se muestran en la Tabla 3.

Tabla 3. Ejemplos de eventos indicadores.

Oraciones del género periodístico con un evento indicador
Una cuarta parte de hogares poblanos apenas tiene acceso a Internet
Sólo un tercio -31.8 por ciento- de los poblanos consideró al gobierno municipal
Juntos canjearon 39 planillas, la mitad con nombre de Eugenio, la otra ...
Ventas de comercio al por menor crecerán hasta 7 por ciento en 2016
La economía informal contribuyó con el 24.8 por ciento del Producto Interno Bruto
Esto supone entre un 25 y un 30 por ciento de los ingresos
Nuevo Sistema de Justicia Penal, al 90 % en Michoacán
El 75 % de los trabajadores en México está sometido a algún grado de estrés laboral, y eso a la larga es la causa del 25 % de los 75 mil infartos ...

7. Conclusiones y perspectivas

En este trabajo se han presentado experimentos relacionados con la extracción de eventos indicadores que utilizan un número fraccionario (tomando como base el 100) para expresar una unidad de valor.

El trabajo aporta una serie de patrones morfosintácticos útiles en la tarea de identificación de eventos indicadores. Se han extraído y presentado aquellos patrones lingüísticos que han mostrado una mayor regularidad de ocurrencia en los eventos indicadores.

Como trabajo a futuro se considera incrementar sustancialmente el número de noticias sobre el cual se llevarán a cabo los experimentos y llevando a cabo una

evaluación manual de todos y cada uno de los eventos extraídos manualmente, lo cual será por supuesto una tarea costosa desde el punto de vista del tiempo y esfuerzo humano.

Es importante analizar el conjunto inicial de eventos indicadores, a fin de poder enriquecer los patrones morfosintácticos y encontrar otros que aunque poco frecuentes, sean de interés en la extracción de la información basada en eventos.

Referencias

1. Andersen M., Hayes J., Huettner A., Schmandt L., Nirenburg I.: Automatic Extraction of Facts from Press Releases to Generate News Stories, Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics, ANLC ̀92, pp. 170–177 (1992)
2. Gil-Vallejo L., Castellón I., Coll-Florit M.: Hacia una definición de la similitud verbal para la extracción de eventos, Centro Virtual Cervantes (2015)
3. Hernández L. D.: Sistema web para identificar eventos y actores en textos periodísticos, royecto terminal, Departamento de Sistemas, Universidad Autónoma Metropolitana Unidad Azcapotzalco, México (2015)
4. Martín F. J., Ruiz J. L.: Procesamiento del lenguaje natural, España: Universidad de Sevilla. Disponible en: https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf (2013)
5. Moncecchi G., Rosá A.: Reconocimiento automático de eventos en textos, Proyecto de grado, Facultad de Ingeniería, Universidad de la República, Uruguay (2010)
6. Padró L., Stanilovsky L.: FreeLing 3.0: Towards Wider Multilinguality Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.Istanbul, Turkey (2012)
7. Priego-Sánchez B., Pinto D.: Identification of Verbal Phraseological Units in Mexican News Stories. Computación y Sistemas, Vol 19(4), pp. 713–720 (2015)
8. Ramos O., Pinto D., Priego-Sánchez B., Olmos I., Beltrán B.: Análisis empírico de la dispersión del español mexicano. Research in Computing Science 74, pp. 9–19 (2014)
9. Real Academia Española.: Diccionario de la lengua española [Dictionary of the Spanish Language] (22nd ed.). Madrid, Spain (2001)
10. Steinwart I., Christmann A.: Support Vector Machines (1st ed.). Springer Publishing Company, Incorporated (2208)
11. Sutton C., McCallum A.: An introduction to conditional random fields for relational learning, in: L. Getoor, B. Taskar (Eds.), Introduction to Statistical Relational Learning, Ch.1, MIT Press (2007)
12. TES (Terminology Extraction Suite): Distibución para Windows—Traducció, Traduccio.blogs.uoc.edu, urlhttp://traduccio.blogs.uoc.edu/2012/04/13/52/
13. TreeTagger, urlhttp://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
14. Wonsever D., Malcuori M., Rosá-Furman A.: SIBILA: Esquema de anotación de eventos". Reportes Técnicos 08-11. UR. FI INCO (2008)