ISSN 1870-4069

Extracción semiautomática de metadatos en documentos no estructurados utilizando procesamiento de lenguaje natural y propiedades tipográficas

Alberto Iturbe Herrera[1], Azucena Montes Rendón[2], Juan-Manuel Torres-Moreno[3,4], Gerardo Sierra Martínez[5], Noé Alejandro Castro Sánchez[1], Juan Gabriel González Serna[1]

[1] Tecnológico Nacional de México / Cenidet, Cuernavaca, Morelos, México
[2] Tecnológico Nacional de México / Tecnológico de Tlalpan, San Miguel Topilejo, Ciudad de México
[3] LIA / Université d'Avignon, Avignon, France
[4] Polytechnique de Montréal, Montréal (Québec), Canada
[5] Universidad Nacional Autónoma de México, Instituto de Ingeniería, Ciudad de México, México
{iturbe, ncastro, gabriel, amr}@cenidet.edu.mx, juan-manuel.torres@univ-avignon.fr, gsierram@iingen.unam.mx

Resumen. Hoy en día existen distintos sistemas capaces de extraer metadatos de documento digitales. Sin embargo, la ausencia de una estructura definida en la distribución de los metadatos en documentos de una biblioteca digital de arte presenta un gran problema, esto se debe generalmente al estilo que cada autor o editorial decide utilizar tanto en la portada como en la portadilla del documento. A pesar de que existen herramientas de software que realizan la tarea de extracción de metadatos, éstas se enfocan únicamente en documentos estructurados comopublicaciones de revistas, artículos científicos, etc. Los metadatos no son más que datos estructurados de información, es decir, información de información o datos de datos. Este trabajo introduce el uso de técnicas de lenguaje natural y la información tipográfica del texto en el documento para la extracción de metadatos, tales como: título, autores, editorial y fecha de publicación. Los resultados obtenidos en la evaluación con documentos digitales no estructurados indican el potencial del enfoque propuesto, que es capaz de producir buenos resultados en la extracción de metadatos.

Palabras clave: procesamiento de lenguaje natural, extracción de metadatos, información tipográfrica del texto, extracción de información, documentos no estructurados.

Semiautomatic Metadata Extraction from Unstructured Documents Using Natural Language Processing and Typographic Text Properties

Abstract. Today there are different systems capable of extracting digital document metadata. However, the absence of a structure de fi ned in the distribution of metadata in documents from a digital art library presents a major problem, this is generally due to the style that each author or publisher decides to use both on the cover and in the cover of the document. Although there are software tools that perform the task of extracting metadata, they focus only on structured documents such as journals, scientific articles, etc. Metadata is not more than structured information data, that is, information information or data data. This paper introduces the use of natural language techniques and typographic information of the text in the document for the extraction of metadata, such as: title, authors, publisher and date of publication. The results obtained in the evaluation with unstructured digital documents indicate the potential of the proposed approach, which is capable of producing good results in the extraction of metadata.

Keywords: natural language processing, metadata extraction, typographic information, information extraction, unstructured documents.

1. Introducción

Los metadatos no son más que un dato estructurado sobre información, es decir, información sobre información o datos sobre datos. Los metadatos en la web se consideran datos que pueden ser guardados, intercambiados y procesados por un sistema informático, generando estructuras que proporcionarán una mejor identificación, clasificación, ubicación y descripción del contenido en la web desde documentos o sitios de Internet [8].

Hoy en día, los medios electrónicos y digitales son indispensables en la vida cotidiana, esto se debe a la gran facilidad de manipulación y movilidad que ofrecen estos medios, a diferencia de los objetos físicos como libros o revistas. Actualmente, es posible obtener una vasta cantidad de información digital de libros que se encuentran disponibles físicamente, es aquí donde intervienen las diferentes formas de obtención de información como repositorios institucionales o bibliotecas digitales.

A su vez, estos eliminan los diferentes estragos que pueden ocurrir en los libros con el tiempo. Es por eso que son considerados un medio indispensable para la preservación de la memoria histórica. Sin embargo, la gran variedad de fuentes de información en los medios electrónicos no se encuentra clasificada, motivo por el cual las búsquedas suelen complicarse, por tal razón es necesario establecer mecanismos capaces de ordenar y extraer información apoyándose en el concepto metadato.

Este trabajo se enfoca en realizar la extracción de metadatos en documentos digitales de arte en español, con el objetivo de semiautomatizar la tarea de extracción de información en documentos de esta naturaleza.

Sin embargo, los documentos no estructurados, denominados así por las siguientes características: la ausencia de una estructura en la distribución de

los metadatos, los errores generados al utilizar técnicas de OCR o la omisión de las mismas, el tipo y calidad de impresión, y el deterioro de estos al haber sido digitalizados dificulta la extracción de información, motivo por el cual, los sistemas actuales de uso comercial no lograrán extraer la información de forma apropiada.

Este documento está estructurado de la siguiente manera: la Sección 2 presenta información de antecedentes sobre la extracción de información y el procesamiento del lenguaje natural. La sección 3 presenta trabajos relacionados con la extracción de metadatos. En la sección 4 se describen los detalles sobre el sistema de extracción de metadatos propuesto. Los experimentos y resultados se muestran en la sección 5. Finalmente, la sección 6 describe las conclusiones y el trabajo futuro.

2. Background

2.1. Metadato

Por definición, un metadato es un dato estructurado con el objetivo de proporcionar información, describir el contenido, la condición, la calidad y una variedad de características de los datos según sea necesario. El propósito de los metadatos es describir varios atributos de los objetos de información, es decir, otorga significado, contexto y organización, además de mejorar la navegación, búsqueda y administración de archivos en la web.

Etimológicamente los metadatos surgen del griego, $\mu\epsilon\tau\alpha$, meta, que significa "más allá de, después de", en conjunto con el latín *datum*, que significa "lo que se da". Conjunta y literalmente se traduce como ïnformación sobre información.º "datos sobre datos"[2].

2.2. Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural (PLN) se considera un campo interdisciplinario cuyas tareas principales son: la traducción automática, el resumen automático, las búsquedas semánticas, la extracción y recuperación de información, entre otras. El procesamiento del lenguaje natural tiene como objetivo crear sistemas y mecanismos capaces de emular actividades del lenguaje natural, principalmente considerando aspectos lingüísticos.

2.3. Extracción de Información

Esta tarea consiste en extraer las entidades, los eventos y la relación que tienen con respecto a los elementos de un texto o un conjunto de textos. La extracción de información consiste en localizar las partes del texto que contienen información relevante, de acuerdo con las necesidades de los usuarios. Además, esta información se obtiene para ser procesada de forma manual o automática [1].

2.4. Entidades Nombradas

Una entidad nombrada (EN) se define como una unidad léxica que consiste en una secuencia de palabras contiguas que se refieren a una entidad específica, es decir, este elemento referenciado puede ser una persona, un lugar, una organización, cantidades, porcentajes o una fecha [4].

2.5. OCR

OCR, por sus siglas en inglés *Optical Character Recognition*, es un proceso que permite al usuario convertir documentos escaneados, imágenes capturadas previamente con una cámara digital o documentos en formato PDF en documentos con información editable y con opción de búsqueda en ellos.

3. Trabajos relacionados

Una de las tareas de Procesamiento del lenguaje natural es la extracción de información (EI), que implica la extracción automática o semiautomática de información de documentos digitales estructurados, semiestructurados o no estructurados.

Por ejemplo, en [11], Rendón propone una metodología para extraer metadatos en objetos de conocimiento con contenido no estructurado. Este trabajo realiza la extracción de metainformación utilizando las librerías LA-PDFText y GROBID, obteniendo como salida un documento XML con la información identificada a través de etiquetas.

Tkaczyk et. al [12], presentan un sistema integral para la extracción de metadatos en artículos escolares basados en el análisis de la estructura del documento, de encabezado a pie de página. Este trabajo implementó las siguientes bibliotecas: biblioteca iText y LibSVM; y los algoritmos: Docstrum, algoritmos basados en heurísticas de abajo a arriba, agrupación KMeans y campos aleatorios condicionales.

Por otra parte, Guo et. al [6] presentan el sistema SemreX para la extracción de metadatos en artículos científicos. Esta metodología se centra en los documentos IEEE, ACM y LNCS, implementa una serie de heurísticas definidas por los autores, así como el uso de información del texto en el documento como: posición, contenido (texto extraído), tipo de fuente, tamaño de fuente y número de página en el que se encuentra el texto.

Del mismo modo, Huynh et. al [7] presenta la herramienta GATE utilizada para realizar la extracción de metadatos con el propósito de enriquecer una ontología. Este trabajo se centra solo en artículos científicos y utiliza un enfoque basado en las propiedades del texto que implementa Apache PDFBox, JAPE Grammar y el complemento ANNIE.

Morales-Solares et. al presentan en [9] una metodología para extraer los metadatos de las cubiertas de los documentos. Para esto se utilizaron dos métodos: regiones extremas máximas estables (MSER) para texto e imágenes con

fondos complejos, y campos aleatorios condicionales (CRF) para elementos de etiquetado lógico en el documento.

Finalmente, y el enfoque más cercano a este trabajo es presentado por Gao et. al [5] quienes desarrollaron una metodología que utiliza propiedades espaciales del texto para determinar cada metadata. Sin embargo, los documentos que procesaron estaban en idioma chino estructurado.

Analizando los trabajos descritos anteriormente, la mayoría de estos realizar la extracción de metadatos en documentos estructurados, tales como: artículos científicos o de investigación. Basado en esta observación, el presente trabajo introduce el uso de técnicas de procesamiento del lenguaje natural y las propiedades tipográficas del texto en documentos no estructurados en español. La metodología propuesta se evaluó utilizando un corpus de 300 documentos en español en formato PDF, de los cuales se obtuvieron los siguientes metadatos: título, autores, editorial y fecha de publicación.

4. Sistema propuesto

Esta sección presenta las herramientas utilizadas en este trabajo y describelos métodos desarrollados en esta metodología y los corpus utilizados en los experimentos.

4.1. Implementación de la herramienta Freeling

Freeling es una biblioteca desarrollada en el lenguaje de programación C++, esta herramienta nos permite realizar tareas de procesamiento del lenguaje natural, tales como: etiquetado gramatical (PoS, *Part of Speech*), reconocimiento y clasificación de entidades nombradas, desambiguación semántica, análisis morfológico, Detección automática de lenguaje, entre otras.

Además, Freeling puede realizar todas las tareas ya mencionadas en los idiomas: español, inglés, portugués, italiano, francés, alemán, ruso, catalán, galés, croata, esloveno, entre otros. Más información y la documentación de esta herramienta se describen en [10].

4.2. Implementación de la herramienta PDFMiner

PDFMiner es una herramienta para extraer información de documentos en formato PDF, proporcionando una serie de propiedades como: obtener la ubicación exacta del texto en una página y el tipo de fuente correspondiente. Además, esta herramienta permite la conversión de archivos a formato HTML y XML.

Esta herramienta utiliza un enfoque basado en el diseño en el que se utilizan los siguientes parámetros: ***M****: char_margin*, esta variable corresponde a la separación entre caracteres (letras, números o signos), ***L***: representa la variable de separación por líneas (*line_margin*), por lo tanto, si dos fragmentos de texto específicos están en líneas diferentes, pero la distancia entre ellos se aproxima al valor de L, los elementos se agruparán como un cuadro de texto. Finalmente, ***W***:

es la variable responsable de determinar la distancia mínima entre dos palabras (word_margin) para asignar posteriormente los respectivos espacios en blanco.

Las variables mencionadas anteriormente tienen los valores predeterminados de $\boldsymbol{M = 1.0}$, $\boldsymbol{L = 0.3}$ y $\boldsymbol{W = 0.2}$. De acuerdo con las necesidades del usuario, estos valores pueden modificarse [3]. La figura 1 ejemplifica lo mencionado anteriormente.

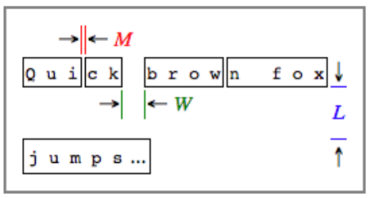

Fig. 1. Parámetros basados en el diseño.

Se requirió una serie de pruebas para determinar que la información extraída de los documentos PDF retenía la mayor cantidad de información posible. Para esto, se utilizó la herramienta en cuestión para extraer el texto usando una línea de comandos. El comando para realizar la extracción es el siguiente:

pdf2txt.py -o destination_directory/myfile.html -t html -p 1,3 myfile.pdf

En el que podemos utilizar diferentes opciones como:

- ***o*** esta variable indica el nombre del archivo de salida.
- ***p*** representa las páginas separadas por comas a partir de las cuales se realiza la extracción.
- ***t*** corresponde a la variable de formato de salida de la información. Este puede ser texto plano, html o xml.
- ***O*** indica el directorio que almacena las imágenes extraídas.
- ***P*** Esta variable se usa para especificar la contraseña del documento si es necesario.

La Figura 2 muestra un ejemplo de la información extraída en formato HTML de un documento PDF, en la que se puede apreciar la sintaxis tradicional de las páginas escritas en este formato. Al mismo tiempo, este documento contiene la información correspondiente a la página y la ubicación del texto en ella. Dicha información se encuentra en la línea 11 para este caso; asimismo, las líneas 12 y 15 contienen la información correspondiente al tamaño y el tipo de fuente para los fragmentos de texto indicados. Más información y documentación de esta herramienta se describe en [3].

4.3. Módulo de pre-procesamiento de documentos PDF

Usando la herramienta mencionada anteriormente, se procesaron los archivos y se almacenaron en las carpetas correspondientes, y seguido de esto, la informa-

```
<html>
<head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
</head>
<body>
    <span style="position:absolute; border: gray 1px solid; left:0px; top:50px; width:595px; height:842px;">
    </span>
    <div style="position:absolute; top:50px;">
        <a name="1">Page 1</a>
    </div>
    <div style="position:absolute; border: textbox 1px solid; writing-mode:lr-tb; left:286px; top:356px; width:158px;
    height:52px;">
        <span style="font-family: KEAHPJ+BodoniBT-Roman; font-size:26px">Conceptos claves
            <br>
        </span>
        <span style="font-family: KEAHPJ+BodoniBT-Roman; font-size:26px">de museología
            <br>
        </span>
    </div>
```

Fig. 2. Información tipográfica extraída en formato HTML.

ción obtenida se almacenó de acuerdo con el formato del archivo. Es necesario mencionar que se desarrollaron algoritmos para el procesamiento de documentos como: la eliminación de saltos de línea en los títulos de los documentos, la eliminación de páginas con información irrelevante y la detección y eliminación de páginas sin contenido.

4.4. Módulo de etiquetado XML

En este módulo, los archivos XML se generaron con el objetivo principal de identificar entidades nombradas tales como: personas, organizaciones, ubicaciones, fechas, entre otras. Sin embargo, al someter las diferentes copias al proceso de etiquetado de PoS utilizando Freeling, fue posible ver que, la mayoría de las palabras que comienzan con letras mayúsculas se identificaron erróneamente como nombres propios, ya sea como persona u organización.

La Figura 3 (a) muestra los ejemplos relacionados con el problema mencionado anteriormente, donde el token con id: **t1.1** contiene la palabra *conceptos* con una clasificación errónea, indicando esto como una persona. De manera similar, el token con id: **t1.5**, que contiene la misma palabra que el token **t1.1**, recibió una clasificación diferente. Otro ejemplo es el token **t1.9** con la palabra *bajo* y la asignación a la clase de organizaciones de manera incorrecta.

Por lo tanto, se desarrolló un conjunto de métodos de mejora de etiquetado de PoS utilizando repositorios construidos. Estos métodos son responsables de modificar los errores mencionados anteriormente para las clases de persona y fecha.

En la gran mayoría de los libros, la fecha de publicación contiene solo el año sin especificar el día y el mes, debido a esto, la herramienta asignó la etiqueta numérica a todos los números encontrados, generando errores con respecto a los números de página y las posibles fechas de publicación. Por esta razón, se desarrolló un método para identificar números no menores de 1800 y no mayores que 2017. Cabe mencionar que estos valores se pueden adaptar a las necesidades.

Una vez que cada método de corrección identifica errores en los textos, extraen información indispensable del fragmento de texto como: id, form y lemma. Posteriormente, si el atributo de form, que representa la palabra en su forma original no presenta incidentes en el repositorio de nombres de personas,

Fig. 3. Errores iniciales y correcciones con el algoritmo de optimización NER / NEC de las herramientas de Freeling.

palabras clave de organización, organización, número de cuatro dígitos entre los rangos ya mencionados o no es ciudad o país, se modifica a letras minúsculas y, posteriormente, se vuelve a enviar al proceso de etiquetado de PoS y la información anterior se reemplaza con los datos obtenidos, al tiempo que se retiene solo la identificación del token para respetar la posición original. Toda esta información se almacena en archivos temporales separados para cada archivo original respectivamente. La Figura 3 (b) muestra la corrección de los errores en los tokens: **t1.1**, **t1.5**, **t1.9**.

4.5. Módulo de extracción y almacenamiento de metadatos

El siguiente módulo describe el enfoque de este trabajo y las heurísticas para determinar cada uno de los cuatro metadatos.

Detección de autores. Combinando el algoritmo de optimización integrado de Freeling y los repositorios de nombres propios, se agregó un algoritmo capaz de generar tres formas diferentes de los registros en el repositorio de nombres propios, es decir, nombre propio capitalizado (Nombre Propio), minúscula (nombre propio) y mayúscula (NOMBRE PROPIO). A su vez, si este contiene algún signo de acentuación como: acento agudo (’), acento circunflejo (ˆ), acento grave (‘), virgulilla de la eñe (˜), virgulilla de la cedilla (Ç / ç), diéresis (¨), entre otros. La palabra se procesa para generar nuevos registros con la ausencia de estos caracteres (Gutiérrez - Gutierrez), en las tres variantes mencionadas anteriormente.

Sin embargo, a pesar de haber reducido en gran medida el rango de error en relación con la identificación de las entidades nombradas, algunos casos continuaron apareciendo en los que no se detectaron correctamente en los casos en que otros fragmentos del texto tenían el tamaño de fuente más grande como se muestra en la Figura 4, por lo que se sugirió optimizar este módulo descartando todos los elementos de una cadena de texto que tuvieran incidentes en el diccionario de la RAE, realizando el proceso de tokenización de las cadenas de texto y luego realizando una búsqueda en el repositorio generado de la RAE. Por otro lado,

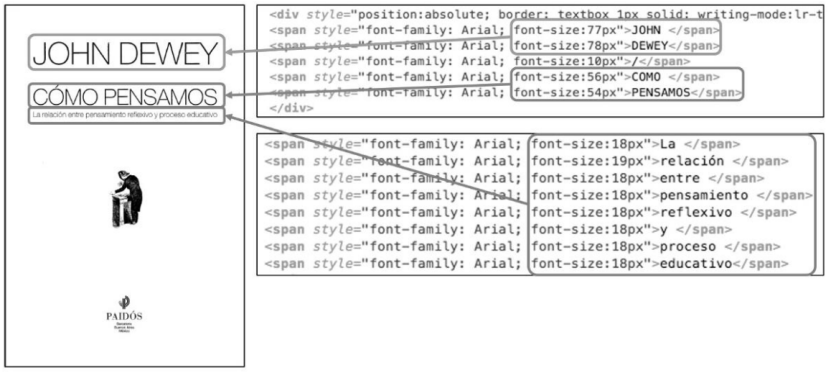

Fig. 4. Ejemplo de distribución de los metadatos.

debido a que existen varios nombres propios con incidencias en el diccionario de la RAE como Abad, Abarca, Bravo, entre otros, se diseñó un algoritmo capaz de evaluar si una palabra determinada como las ya mencionadas, puede o no pertenecer a una entidad nombrada, que recibe como entrada el archivo XML que contiene el etiquetado PoS y la identificación de las entidades nombradas, seguido de un análisis contextual de la palabra, es decir, identifica si el token está precedido o seguido por un nombre propio, ya sea primer nombre o apellido.

Más adelante, se evalúa si la palabra puede o no pertenecer a una entidad nombrada y esta información se actualiza en el archivo XML, agregando: *tag = "NP00SP0çtag = "NP"pos = "noun"type = "proper"neclass = "Person"nec = "PER"* a la etiqueta PoS en caso de que no la contenga y la elimine en caso de que haya una entidad nombrada de tipo de nombre mal identificada.

Una vez que se realiza la actualización del archivo XML, se obtiene la información de todas las etiquetas PoS que contienen solo la etiqueta "NP00SP0", descartando todas aquellas con un tamaño de fuente inferior a 15 píxeles, en caso de que la heurística no pueda extraer la información suficiente, El valor de tamaño de fuente mínimo se modifica automáticamente para continuar la extracción.

La información obtenida se ordena de dos maneras: por tamaño de fuente y por la ubicación espacial del texto en el documento, para identificar a los autores

principales en primer lugar, que deberían tener un tipo de letra más grande que el resto y, posteriormente, a todos los autores. De menor impacto para concluir con la inserción de estos en la base de datos.

Detección de editorial. Al igual que en la detección de autores, se implementó un algoritmo para la detección de editores, sin embargo, a diferencia del anterior, este utiliza una serie de palabras clave como: editorial, impresión, editores, ediciones, S.L., entre otras.

Todas las palabras clave están sujetas a un proceso automático que genera tres formas diferentes de la palabra clave original, es decir, la palabra clave capitalizada (Palabra Clave), minúscula (palabra clave) y finalmente se genera el uso de mayúsculas (PALABRA CLAVE). Además, la palabra clave se procesa en caso de que contenga signos de acentuación como se mencionó anteriormente.

También fue necesario analizar el contexto de la palabra clave que se busca en el documento, es decir, a modo de ejemplo; Si el sistema detecta la palabra editorial y está precedido por signos de puntuación como: ". (Punto) / , (coma)ïmplica en la mayoría de los casos que el nombre del editor está antes del signo de puntuación y antes de la palabra clave identificada, en el caso de identificar ": (dos puntos)ïmplica que el nombre del editor va seguido del signo de puntuación con la opción de continuar en la misma línea o en la siguiente. Finalmente, el último caso ocurre en ausencia de signos de puntuación, lo que implica que en la mayoría de los casos el nombre del editor está en la misma línea.

Posteriormente, en caso de encontrar varias palabras clave, la información obtenida se ordena de dos maneras: por el tamaño de la fuente y por la ubicación espacial del texto en el documento y de esta manera determina qué información corresponde al editor.

Detección de fecha de publicación. Para detectar la fecha de publicación, surge el siguiente problema porque un archivo puede contener una o más fechas a lo largo de las páginas legales del documento y estas pueden ser o no fechas de publicación, es decir, se pueden presentar otras fechas, tales como: Fechas de ediciones anteriores, fechas de traducción, fechas de revisión, entre otras.

Por lo tanto, se aplicó el mismo enfoque para determinar la fecha de publicación, descartando todos los números menores a 1700 y todas las fechas con fuente inferior a 15 píxeles, en caso de que no haya suficiente información, se modifica automáticamente el tamaño mínimo para continuar con la extracción.

En caso de que el sistema identifique más de una fecha, ordenará todos los resultados obtenidos de mayor a menor, priorizando esa fecha que tenga el tamaño de fuente más grande. Si todas las fechas identificadas tienen un tamaño de fuente común, el sistema evaluará el número de veces que se repite una cantidad, eligiendo la que tenga más repeticiones como fecha de publicación.

Finalmente, si no se cumple ninguno de los casos mencionados, el sistema elegirá la fecha más grande identificada y esta se asignará como la fecha de publicación.

Detección de título. Esta tarea es primordial en la extracción y almacenamiento de los metadatos, ya que esta se encargará de realizar las relaciones respectivas entre los autores y los editores con los libros correspondientes.

Al igual que con los metadatos anteriores, se propuso resolver este problema mediante la obtención de las propiedades tipográficas y espaciales del texto en la página, basándose en el análisis de diferentes colecciones digitales, que, en su mayoría, muestran una tipografía diferente del resto, además de poseer la tipografía más grande. Sin embargo, esto no descarta la posibilidad de que algunos autores manifiesten estas características.

Para mitigar estos casos, se realizó un análisis contextual del fragmento de texto, se analizó qué información se encontraba en las líneas anteriores y en las siguientes, es decir, si el nombre propio se puede conectar al resto del texto mediante preposiciones o artículos formando parte del título como: Biografía de Frida Kahlo, Siendo, se es la tesis de Parménides, De Aristóteles a Newton, entre otros. De esta manera, se evita que el sistema tenga en cuenta a esa entidad nombrada como un posible autor.

Posteriormente, la información extraída anteriormente se filtra de tal manera que solo se conservan las líneas del documento con un tamaño de fuente superior a 15 y, en caso de que no haya suficiente información antes de realizar dicha depuración, el sistema corregirá automáticamente el tamaño mínimo. Fuente para continuar la extracción.

Una vez hecho esto, la información se organiza con la tipografía de mayor a menor, y luego se crea la alineación correspondiente a las propiedades espaciales del texto, generando así una cadena de texto legible correspondiente al título del documento.

Finalmente, al igual que con los metadatos anteriores, se realizaron consultas SQL básicas para los motores de base de datos InnoDB y, en el caso de las bases de datos que usan el motor MyISAM, se utiliza la búsqueda FULLTEXT para obtener la puntuación de relevancia y hacer comparaciones basadas en los valores obtenidos.

Este proceso se usa para evitar registros duplicados para cada metadato debido a que algunos editores usan logotipos o acrónimos para referirse a sus nombres, por lo tanto, esto implica que la detección de estos metadatos no se realizará en estos casos debido a la ausencia de palabras clave.

Es necesario mencionar que, si el archivo no tiene permisos de lectura y escritura, no se ha hecho OCR en el documento o está protegido por contraseña, el sistema no podrá generar los archivos necesarios, si es el caso, el sistema realizará la inserción de la ubicación de vista previa del archivo y el tamaño del archivo solamente.

5. Resultados experimentales

Esta sección presenta los experimentos realizados para evaluar la metodología propuesta en este trabajo. Para realizar la evaluación, se seleccionaron 300 documentos no estructurados en español en formato PDF, es necesario mencionar

que este corpus fue construído manualmente obtenidos de internet, estos fueron analizados manualmente para identificar el número total de autores, editores y fechas de publicación de cada libro.

En total, se obtuvieron 300 títulos, 549 autores, 265 editoriales y 282 fechas de publicación. Estos valores se compararon con los datos generados por el sistema para determinar la precisión y cobertura correspondientes utilizando las ecuaciones 1 y 2:

$$Precisión = \frac{tp}{(tp + fp)}, \tag{1}$$

$$Cobertura = \frac{tp}{(tp + fn)}, \tag{2}$$

donde tp: representa los valores llamados verdaderos positivos (*true positive*), es decir, los valores que fueron identificados por el sistema como los metadatos correspondientes y, de hecho, lo son. fp: representa los valores denominados falsos positivos (*false positive*), es decir, los valores identificados por el sistema como metadatos específicos, pero no son los correctos. Finalmente, fn: representa los valores denominados falsos negativos (*false negative*), es decir, aquellos que el sistema **no** identificó como los metadatos correspondientes, pero deben considerarse correctos. La Tabla 1 describe los verdaderos positivos, falsos positivos, falsos negativos, precisión y cobertura para cada metadato.

Tabla 1. Resultados de Precisión y Cobertura para cada metadato.

	Título	**Autor**	**Editorial**	**Fecha de publicación**
True positive	224	406	239	239
False positive	76	81	61	61
False Negative	76	62	61	61
Precisión	74.66 %	83.36 %	79.66 %	82.33 %
Cobertura	74.66 %	86.76 %	79.66 %	82.33 %

A partir de los resultados obtenidos, se determinó que la razón principal de algunos títulos que no se identificaron correctamente se debe a la problemática en la distribución de la información y la variación en los textos, lo que genera textos incompletos. Además, en algunos casos, la información tipográfica de los documentos era la misma, independientemente del nivel visual que fuera diferente, por lo tanto, las heurísticas basadas en estas propiedades no podían realizar la extracción correctamente. Por otro lado, el sistema no pudo identificar a todos los autores debido a los diferentes orígenes de los que pueden provenir los nombres propios. Por lo tanto, ya que estos no se encuentran en los repositorios

y el freeling no ha podido identificarlos, produce la ausencia del etiquetado de la entidad nombrada.

Con respecto a los metadatos editoriales, los errores se produjeron en algunos casos por la ausencia de palabras clave utilizadas para determinar el editor. Además, algunos libros utilizan el logotipo del editor, por lo que están ausentes las palabras clave.

Finalmente, al igual que con los metadatos anteriores, si la información tipográfica en algunos documentos era la misma sin importar si eran visualmente diferentes y había más de una incidencia con las palabras clave, había errores en la extracción de la información.

Es necesario mencionar que los resultados de la extracción de fechas de publicación se vieron afectados principalmente cuando hay dos o más fechas con el mismo número de incidentes entre ellos y cuando el documento tiene demasiadas fechas, que a veces representan en número de reimpresiones o Ediciones de cualquier tema.

Adicionalmente, se realizó la comparación de los valores de precisión y cobertura obtenidos para este trabajo y aquellos mencionados anteriormente, estos datos se muestran en la Tabla 2. Es necesario mencionar que estos trabajos se enfocan en distintos metadatos en distintos idiomas y en distintos tipos de documentos. Si bien los resultados obtenidos en este trabajo no obtuvieron los mejores resultados, estos destacan por el hecho de enfocarse en documentos no estructurados y en idioma español, los cuales no podría ser extraídos con los sistemas enfocados a documentos estructurados.

Tabla 2. Tabla comparativa.

	Título		**Autor**		**Editorial**		**Fecha de publicación**	
	PREC	**REC**	**PREC**	**REC**	**PREC**	**REC**	**PREC**	**REC**
Gao et. al [5]	98.1 %	36.4 %	93.9 %	34.2 %	-	-	-	-
Guo et. al [6]	89 %	-	84.5 %	-	-	-	-	-
Huynh et. al [7]	100 %	100 %	92.72 %	89.47 %	95.83 %	92 %	-	-
Tkaczyk et. al [12]	84.07 %	-	87.68 %	-	97.5 %	-	93.67 %	-
Nuestro trabajo	**74.66 %**	**74.66 %**	**83.36 %**	**86.76 %**	**79.66 %**	**79.66 %**	**82.33 %**	**82.33 %**

El objetivo de semiautomatizar el proceso de extracción de metadatos en documentos no estructurados se resolvió utilizando este enfoque. Sin embargo, los errores en la extracción y clasificación de la información continúan presentándose, por lo que, al ser un proceso semiautomático, el usuario final puede verificar la información y corregirla si es necesario mediante una interfaz web.

6. Conclusiones y trabajos futuros

Este trabajo presentó un enfoque para la extracción de metadatos utilizando técnicas de procesamiento de lenguaje natural e información tipográfica del texto en documentos no estructurados en español.

Es necesario mencionar que a lo largo de esta investigación no fue posible determinar un patrón en la organización de los metadatos en este tipo de documentos. Por lo tanto, esta metodología se diseñó bajo el enfoque del análisis de las propiedades tipográficas del texto en el documento para determinar los metadatos que utilizan este enfoque.

Además, se construyeron una serie de repositorios de nombres propios, ciudades, países y palabras clave editoriales, que se pueden usar en otras aplicaciones de procesamiento de lenguaje natural, más específicamente en el Reconocimiento y Clasificación de Entidades Nombradas.

Finalmente, la tarea de extraer metadatos se vuelve bastante compleja porque carece de un orden en la distribución de estos, por lo que esta investigación se centró solo en la extracción de los metadatos: título, autores, editorial y fecha de publicación. Sin embargo, en algunos casos, el OCR aplicado a cada documento generalmente genera errores en la interpretación de los caracteres, por lo tanto, obtiene cadenas de texto incoherentes que causan errores de extracción de información.

Referencias

1. Information Extraction - Universitat Pompeu Fabra. https://www.upf.edu/hipertextnet/numero-5/pln.html, accessed: 2019-03-4
2. Metadata - Universidad Nacional de Colombia Sede Amazonia. http://www.unal.edu.co/siamac/sig/metadatos1.html, accessed: 2019-03-4
3. PDFMiner. https://media.readthedocs.org/pdf/pdfminerdocs/latest/pdfminerdocs.pdf, accessed: 2019-03-4
4. Carreras, X., Màrquez, L., Padró, L.: Named entity recognition for Catalan using Spanish resources. In: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1. pp. 43–50. EACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), https://doi.org/10.3115/1067807.1067815
5. Gao, L., Zhong, Y., Tang, Y., Tang, Z., Lin, X., Hu, X.: Metadata extraction system for Chinese books. In: 2011 International Conference on Document Analysis and Recognition. pp. 749–753 (Sep 2011)
6. Guo, Z., Jin, H.: A rule-based framework of metadata extraction from scientific papers. In: 2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. pp. 400–404 (Oct 2011)
7. Huynh, T., Hoang, K.: Gate framework based metadata extraction from scientific papers. In: 2010 International Conference on Education and Management Technology. pp. 188–191 (Nov 2010)
8. Lapuente, M.J.L.: Hipertextos. http://www.hipertexto.info/documentos/ metadatos.htm, accessed: 2019-03-4
9. Morales-Solares, C., Sierra, G., Escalante-Ramirez, B.: An unsupervised approach for automatic discovery of metadata in document images. In: 2016 Fifteenth Mexican International Conference on Artificial Intelligence (MICAI). pp. 1–7 (Oct 2016)
10. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)

11. Rendón, J.C.: Clasificación automática de objetos de conocimiento con contenido no estructurado para el poblado semiautomático de ontologías multidimensionales. Master's thesis, Centro Nacional de Investigación y Desarrollo Tecnológico - Sin publicar (2014)
12. Tkaczyk, D., Bolikowski, L., Czeczko, A., Rusek, K.: A modular metadata extraction system for born-digital articles. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 11–16 (March 2012)