Generación automática de resúmenes

Alejandra Carolina Cardoso y M. Alicia Pérez Abelleira
Universidad Católica de Salta, Facultad de Ingeniería e IESIING

Abstract
La generación automática de resúmenes de texto suele formar parte de las aplicaciones de minería de textos, como una manera de presentar el resultado de ésta al usuario. El problema principal consiste en identificar la información más importante en el documento de origen. En este trabajo comenzamos analizando aspectos generales del problema, para centrarnos en la generación de resúmenes por extracción utilizando técnicas de aprendizaje automático supervisado. Se describen las características o atributos del texto más útiles para construir el modelo que clasifica fragmentos del texto como relevantes o no para el resumen y se evalúan técnicas de aprendizaje de dichos modelos. Sobre un corpus de documentos de texto correspondientes a resoluciones rectorales, la construcción de árboles de decisión obtiene resúmenes de calidad adecuada, que sirven como resúmenes indicativos para el usuario de un buscador semántico en dicho corpus.

Palabras Clave
Generación automática de resúmenes, minería de texto, UIMA

Introducción

El interés en la minería de textos ha crecido enormemente en los últimos años, debido a la creciente cantidad de documentos disponibles en forma digital y la también creciente necesidad de organizarlos y aprovechar el conocimiento contenido en ellos. La minería de textos es el proceso de extraer información y conocimiento interesante y no trivial de texto no estructurado. Es un campo relativamente reciente con un gran valor comercial que se nutre de las áreas de recuperación de la información (IR), minería de datos, aprendizaje automático, estadística y procesamiento del lenguaje natural. La minería de textos incluye una serie de tecnologías, entre otras: extracción de la información, seguimiento de temas (*topic tracking*), generación automática de resúmenes de textos, categorización, agrupamiento, vinculación entre conceptos, visualización de la información, y respuesta automática de preguntas. El presente trabajo se centra en la generación automática de resúmenes de textos usando técnicas de aprendizaje automático y complementa trabajos anteriores en la categorización de documentos y en la búsqueda semántica en el contenido de los mismos [1] y en la extracción de entidades con nombre (NER) [2]. Los resúmenes obtenidos se presentan al usuario junto al resultado de sus consultas en la colección de documentos mediante el buscador semántico. El dominio de aplicación es un corpus de más de 8000 documentos que contienen 9 años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos (Word, texto plano, PDF).

Este trabajo comienza describiendo el problema de la generación automática de resúmenes y algunos enfoques utilizados.

Además se describe la arquitectura de gestión de información no estructurada en que se ha implementado el buscador semántico y en particular la tarea de extracción de resúmenes. A continuación se describe el enfoque utilizado para resolver el problema y los experimentos realizados junto a algunos ejemplos, para terminar con algunas conclusiones.

La generación automática de resúmenes

Un resumen es una transformación reductiva de un texto fuente a un texto resumen por reducción de su contenido mediante selección y/o generalización de lo que es importante en el texto fuente [3]. La generación automática de resúmenes se define como el proceso de destilar la información más importante de una fuente (o de varias fuentes) para producir una versión abreviada destinada a un usuario (o

conjunto de usuarios) determinado y para una tarea (o conjunto de tareas) determinada [4].

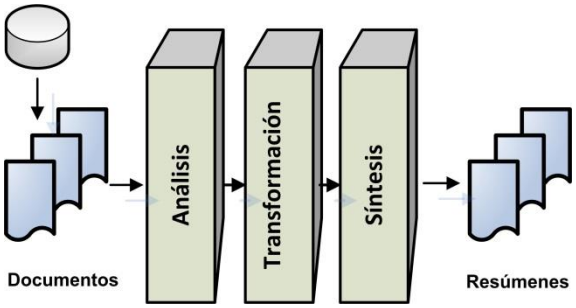

Figura 1 Arquitectura general de un sistema de generación de resúmenes

La Figura 1 muestra a muy alto nivel la arquitectura tradicional de un sistema de generación automática de resúmenes [3] [4]. La entrada al proceso puede ser un solo documento, o una colección de ellos. Las tres fases del proceso son: analizar o interpretar el texto de entrada, posiblemente convirtiéndolo a una representación interna, transformarlo en una representación que lo resuma, y generar una forma apropiada de salida. Estas tres fases pueden emplear una o más de tres operaciones básicas de condensación del texto: la selección de la información más destacada o que no sea redundante, la agregación de información de diferentes partes de la fuente, y la generalización de información específica usando información más general.

Radev et al [5] hablan de cuatro procesos: extracción es el proceso de identificar el material importante en el texto, abstracción es el proceso de reformularlo de manera nueva, fusión es el proceso de combinar fragmentos extraídos, y compresión es el proceso de eliminar material que no es importante. En todos estos procesos aparece la necesidad de mantener cierto grado de coherencia y gramaticalidad.

Se pueden considerar dos grandes grupos de sistemas que generan resúmenes automáticamente [6]: los que realizan resúmenes por extracción y los que realizan resúmenes por abstracción. Un resumen por abstracción se obtiene "comprendiendo" el documento original y reescribiéndolo con menos palabras. Esto implica una nueva redacción que puede contener términos o frases que no necesariamente estaban en el documento original. En la generación automática de resúmenes abstractivos, se utilizan métodos lingüísticos, que permiten describir mejor al documento. Si bien esta forma parece ser la mejor manera de obtener un resumen, supone un análisis en profundidad del texto que identifique fragmentos claves y genere un ensamblado en un texto coherente. Por esta razón este tipo de resumen solo se ha aplicado en ámbitos muy concretos.

Un resumen por extracción divide el texto en fragmentos (oraciones, párrafos, etc.) y selecciona los más importantes; los fragmentos elegidos no sufren modificación respecto del texto original. Para identificar los fragmentos clave puede considerarse la estructura del texto; por ejemplo, si se compone de capítulos o secciones podría inferirse que en la sección "Conclusiones" se encuentra información importante. Por otro lado la longitud del texto nos aporta la forma en que se puede dividir; en los textos largos como libros, los fragmentos a considerar podrían ser los párrafos, en cambio para textos más cortos sería suficiente considerar las oraciones. La mayor ventaja de este enfoque es que resulta muy robusto y fácilmente aplicable a contextos de propósito general por su independencia del dominio y género y es el que aplicamos en este trabajo.

Existen en la literatura [5] [7] [8] diversas clasificaciones de los resúmenes y de estrategias para obtenerlos. A continuación se mencionan simplemente las que tienen relevancia para nuestro problema. Atendiendo al alcance, un resumen puede obtenerse de un único documento, o de un conjunto de documentos relacionados. En cuanto al propósito al que el resumen está destinado, algunos resúmenes son indicativos, cuando dan una idea de qué se trata el texto pero sin incluir contenido específico, ya que su objetivo es ayudar a un lector a decidir sobre la relevancia del documento original. Por otro lado los

resúmenes informativos proveen una versión abreviada del contenido y como tales podrían reemplazar al documento original. Atendiendo al destinatario, podemos distinguir entre resúmenes genéricos, que recogen los temas principales del documento y están destinados a un conjunto amplio de lectores, y resúmenes adaptados al usuario, cuando son elaborados de acuerdo a los intereses o perfil del lector (por ejemplo sus conocimientos previos, sus temas de interés o necesidades de información).

Considerando ahora las estrategias utilizadas para extraer resúmenes, éstas pueden clasificarse según la profundidad del procesamiento. Las estrategias poco profundas en general no analizan el texto más allá del nivel sintáctico y utilizan características superficiales del mismo tales como [8]:

- frecuencia de los términos: tales medidas estadísticas pueden capturar el tema del texto, asumiendo que las frases importantes son las que contienen palabras que ocurren frecuentemente en el documento
- ubicación: la intuición aquí es que las frases importantes están situadas en ubicaciones particulares, que dependen del género del texto, aunque hay algunas reglas que podrían ser generales como tomar las primeras frases del texto o los encabezados. La ubicación relevante según el tipo de texto puede ser identificada con técnicas de aprendizaje automático [9]. Muchos de los sistemas que utilizan aprendizaje automático para extraer resúmenes utilizan la ubicación de las unidades de texto para estimar su importancia hacia su inclusión en el resumen.
- sesgo: la relevancia de ciertas frases puede depender de que incluyan términos que aparecen en el título o en encabezados del documento, o hasta en la consulta del usuario que requiere el resumen.
- palabras clave como “en resumen”, “en conclusión”, o este trabajo describe” u otras dependientes del dominio pueden señalar la relevancia (o irrelevancia) de una cierta frase en el texto. Algunos sistemas usan listas de tales palabras generadas manualmente; otros las han generado automáticamente.

Otras estrategias exploran el texto a mayor profundidad, modelando las entidades que aparecen en el texto y sus relaciones, o hasta al nivel de discurso, modelando la estructura global del texto y su relación con metas de comunicación. En general, las estrategias poco profundas generan resúmenes que son extractos, mientras que las más profundas son necesarias para generar abstractos [10].

De la literatura (ver [8] para un panorama) solo mencionamos algunos trabajos representativos y relevantes para nuestro problema. En [7] mediante una serie de heurísticas que se aplican a fragmentos de texto y se combinan en una función, se determina la inclusión o no del fragmento en el resumen, que por tanto se construye por extracción. Dos de las heurísticas son generales (dar mayor puntuación a las cinco primeras frases, determinar las palabras más significativas del texto y asignar mayor valor a las frases con mayor cantidad de esas palabras) y la tercera depende del usuario, potenciando los fragmentos con mayor similitud a las preferencias del usuario. Tanto las heurísticas como la función que las combina son preestablecidas. Diversos trabajos, incluyendo el descrito en este artículo, proponen en cambio aprender automáticamente cuáles son las características relevantes. Por otro lado [6] plantea el problema como uno de aprendizaje no supervisado y usa el algoritmo k-medias de agrupamiento para extraer frases claves. La hipótesis es que al formar grupos de oraciones similares; puede formarse un resumen tomando de cada grupo la oración más representativa

En base a lo descrito, caracterizamos nuestro problema de generación automática de resúmenes para el buscador semántico, como uno en que es suficiente obtener un resumen de cada documento por extracción. Este resumen puede ser meramente indicativo ya que su objeto es presentar al

usuario los resultados de la búsqueda, aunque el resumen contendrá todos los elementos significativos del texto. Para ello es suficiente adoptar una estrategia poco profunda. Así plantearemos la obtención de un resumen como un problema de clasificación para el cual se construirán modelos mediante técnicas de aprendizaje automático. Se tomarán frases o fragmentos de textos como unidades de análisis. El objetivo a aprender es si incluir o no un fragmento en el resumen. Las características o atributos de dichos fragmentos tendrán que ver con la presencia o ausencia de anotaciones correspondientes a entidades con nombre (que en el corpus de resoluciones rectorales, descrito en la sección siguiente, corresponden a personas, instituciones, carreras, unidades académicas, fechas, identificador de la resolución), la ubicación dentro del texto (sección de la resolución rectoral en que aparece) y la clase (categoría) asignada a la resolución [1]. La Figura 2 muestra este proceso.

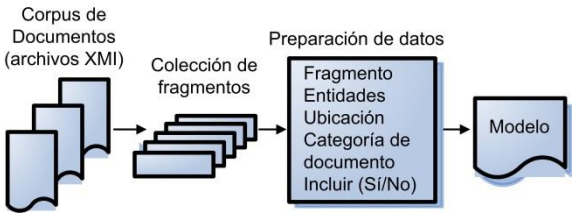

Figura 2 Proceso para obtener el modelo para generar resúmenes

Arquitectura

Conceptualmente las aplicaciones de gestión de información no estructurada, como es el caso de los documentos de texto, suelen organizarse en dos fases. En la fase de análisis se recogen y analizan colecciones de documentos y los resultados se almacenan en algún lenguaje o depósito intermedio. La fase de entrega hace accesible al usuario el resultado del análisis, y posiblemente el documento original completo mediante una interfaz apropiada. La Figura 3 muestra la aplicación de este esquema a nuestro dominio [1], en el que partimos de más de 8000 resoluciones rectorales en archivos de texto de distinto tipo: Word, PDF, texto plano. Previo al análisis, se procede a la extracción del texto de cada archivo utilizando herramientas de software libre (*poi.apache.org* y *tm-extractors*). El texto se normaliza eliminando acentos para facilitar los procesos de búsqueda y equiparación de cadenas. También se divide en partes la resolución extrayendo el encabezado (texto que contiene el número y la fecha de la resolución) y el cuerpo con la mayor parte de la información, y descartando en lo posible el texto "de forma".

La fase de análisis incluye tokenización y detección de entidades en documentos individuales tales como personas, fechas, organizaciones, unidades académicas y datos sobre la resolución (fecha y número). Además con la ayuda de un clasificador

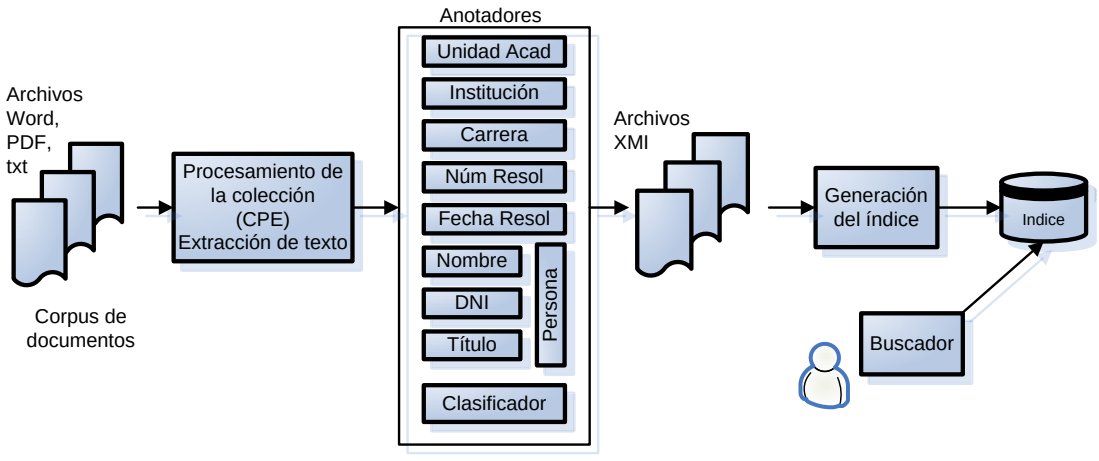

Figura 3 Arquitectura del sistema

aprendido automáticamente del corpus de resoluciones se anota cada documento con una categoría. Existen 21 categorías que fueron obtenidas del personal especializado en la elaboración de resoluciones. Algunos ejemplos son: designación de planta docente, convenio de pasantías, convenio de colaboración, o llamado a concurso docente.

El resultado de la fase de análisis es un conjunto de archivos en formato XMI (*XML Metadata Interchange*). Estos archivos contienen, además de las partes relevantes del texto original, metadatos en forma de anotaciones correspondientes a las entidades existentes y a la categoría de documentos. Los archivos serán procesados para construir el índice de un motor de búsqueda que contiene los tokens (en nuestro caso, las palabras que aparecen en el texto) y las entidades y categorías extraídas automáticamente.

En la fase de entrega existe una interfaz para hacer búsquedas en el índice. El usuario puede buscar documentos que contengan combinaciones booleanas de entidades, categorías y tokens mediante un motor de búsqueda semántica.

Las dos fases están desarrolladas sobre UIMA (*Unstructured Information Management Architecture*), una arquitectura basada en componentes para construir sistemas de procesamiento de información no estructurada [11]. En UIMA, el componente que contiene la lógica del análisis se llama anotador. Cada anotador realiza una tarea específica de extracción de información de un documento y genera como resultado anotaciones, que son añadidas a una estructura de datos denominada CAS (*common analysis structure*). A su vez, esas anotaciones pueden ser utilizadas por otros anotadores. Los anotadores pueden ser agrupados en anotadores agregados. La mayoría de nuestros anotadores realizan reconocimiento de entidades con nombre (NER), a saber: personas, unidades académicas, carreras, instituciones; además hay otros que extraen fechas, número y año de las resoluciones. Para detectar entidades correspondientes a personas se agregan otras (nombres propios, DNIs y títulos) obtenidas por los anotadores correspondientes. Un último anotador asigna la categoría de documento en base al modelo aprendido automáticamente [1]. Todos estos anotadores son modelos aprendidos automáticamente [2].

Extracción de resúmenes de resoluciones rectorales

Para plantear el problema de generación de resúmenes como uno de clasificación (Figura 2), es preciso convertir los documentos de texto, en nuestro caso las resoluciones, en conjuntos de instancias o ejemplos de entrenamiento. Cada ejemplo va a corresponder a un fragmento de texto y para cada ejemplo se determinará el valor de una serie de características o atributos.

El primer paso de procesamiento fue extraer los fragmentos de cada resolución. Inicialmente se utilizaron oraciones como fragmentos, extraídas con la herramienta *OpenNLP Sentence Detector* que detecta oraciones o fragmentos de textos en inglés. Considera una oración como la secuencia de caracteres más larga entre dos puntos. Aunque la herramienta distingue cuando un punto es parte de un número, abreviatura de nombre o siglas (ej D.N.I.), pero no cuando forma parte de un título. Por ej. el texto “El Dr. Gonzalez visitó la universidad.” es separado en “El Dr” y “Gonzalez visitó la universidad”. Para evitar este problema se preprocesó el texto eliminando los puntos de las abreviaturas de título (Lic., Ing., etc.). El texto disponible no es perfecto y en muchos casos omite el punto al final de párrafos o artículos, produciendo fragmentos muy extensos. Por ello se decidió usar también el punto y coma como separador de textos.

Para obtener el conjunto de entrenamiento se han utilizado las resoluciones en formato XMI (Figura 3). Estos archivos contienen, además del texto completo de la resolución, las anotaciones. Usando éstas podemos determinar para cada fragmento qué

anotaciones, en particular entidades con nombre, están presentes. Esta información será una de las características o atributos (Tabla 1).

Tabla 1. Ejemplos de fragmentos y anotaciones

Fragmento	**Atributos**
Art 1.- Disponer que con todas las formalidades del caso, se tome juramento de rigor a la egresada, ANA MARIA TORANZO, DNI: 21.182.376	tieneUA : NO tienePersona: SI tieneDNI: SI tiene Titulo: NO tieneCarrera: NO
la solicitud elevada por las autoridades de la FACULTAD DE CIENCIAS JURIDICAS, para la cobertura del cargo de Profesor Auxiliar Docente;	tieneUA : SI tienePersona: NO tieneDNI: NO tieneTitulo: SI tieneCarrera: NO

Además se han utilizado características asociadas con la ubicación del fragmento. En una resolución se identifican cuatro partes: (a) el encabezamiento, que incluye la fecha y número de la resolución; (b) el texto donde se detalla las causas de la resolución y que comienza con la palabra *VISTO*; (c) el texto donde se fundamenta la razón de la resolución y que comienza con la palabra *CONSIDERANDO*; (d) la última parte donde se enumeran los artículos. En el corpus se observa que en muchas resoluciones el texto de la sección (a) es casi idéntico al de la sección (d). Teniendo esto en cuenta se evita que en el resumen obtenido aparezcan frases relevantes duplicadas. Esta información se ha incluido en la característica *parteDeResol* con cuatro valores posibles.

El atributo *Clase* corresponde al tema de la resolución (existen 21 categorías posibles) asignada por un modelo aprendido automáticamente [1], como se indicó anteriormente y está también disponible en el archivo XMI.

Finalmente, se incluye el atributo que funcionará como atributo de clase a la hora de construir los modelos de clasificadores, en este caso *Incluir*, que indica si el fragmento debe o no estar en el resumen, con valores SÍ/NO. El etiquetado de los ejemplos del conjunto de entrenamiento con este atributo se realizó en forma manual.

La colección de fragmentos así obtenida y sirve para entrenar modelos usando diferentes técnicas y evaluarlos, como se describe en la sección siguiente.

Experimentos

Para realizar los experimentos se determinó preparar cuatro conjuntos de entrenamiento: tres conjuntos con resoluciones de las categorías Juramento, Designación de Planta Docente y Aprobación de Curso (las categorías más numerosas en el conjunto de entrenamiento, de entre las 21 existentes [1]). Un cuarto conjunto incluye resoluciones de todas las categorías. En total se trata de 500 resoluciones del año 2007 (Tabla 2), disponibles como 500 archivos en formato XMI. Estos archivos fueron preprocesados obteniéndose 3822 fragmentos, correspondientes a ejemplos del conjunto de entrenamiento que fueron etiquetadas manualmente con el atributo *Incluir*, de acuerdo a lo que se entendía que debía estar en el resumen.

Tabla 2 Descripción de los conjuntos de entrenamiento.

Conjunto	**Docu-mentos**	**Frag-mentos**	**Incluir =Sí**	**Incluir =No**
Juramento	112	978	320	658
Desig Planta Doc	135	825	228	597
Curso	34	359	86	273
Todas	500	3822	738	3084

Para los experimentos se utilizó la herramienta de software libre Weka, que incluye una amplia colección de técnicas de clasificación adecuadas a este problema. Los resultados se muestran en Tabla *3* y 4.

Tabla 3 Resultados experimentales

Conjunto de datos	**ADTree**	**ID3**	**C4.5 con poda**	**C4.5 sin poda**	**Tabla de Decisión**	**Ripper**	**Naïve Bayes**
Curso	89.08	90.00	88.22	89.05	87.57	89.08	89.19
Desig Planta Doc	96.76	96.78	96.82	96.82	96.72	96.79	96.30
Juramento	99.51	99.65	99.49	99.49	99.21	99.49	98.88
Todas	94.06	94.89 +	94.62	95.05 +	93.29	94.45	89.64 -

Para cada conjunto de datos se utilizó validación cruzada con 10 pliegues, que además se repitió 10 veces. Tras nuestra experiencia en otros problemas de aprendizaje para clasificación de textos y algunos experimentos previos, se eligieron para evaluar las siguientes técnicas (ver [12] para más detalles sobre los algoritmos), que se corresponden con las columnas de la Tabla 3:

- aprendizaje de árboles de decisión utilizando el algoritmo C4.5 (implementación J4.8 de Weka), en dos casos: con poda usando un factor de confianza 2.5 para realizar la poda, y sin poda. También se utilizó el algoritmo básico ID3.
- aprendizaje de árboles de decisión alternantes (ADTree), que representan un conjunto de clasificadores obtenidos mediante *boosting*.
- aprendizaje de reglas utilizando poda incrementalmente para reducir el error, con la implementación JRip que hace Weka del algoritmo RIPPER.
- aprendizaje de tablas de decisión mediante un sencillo clasificador por mayoría.
- aprendizaje bayesiano con el algoritmo Naïve Bayes.

La Tabla 3 muestra el porcentaje de instancias correctamente clasificadas (precisión o *accuracy*) utilizando las técnicas más relevantes tras experimentos preliminares. Además se realizó un test de significatividad estadística (*corrected resampled t-test* [12]) de las diferencias entre la primera técnica de la tabla tomada como base frente a las restantes. El símbolo +/- colocado junto a un resultado indica que el esquema correspondiente sobre el conjunto de datos es estadísticamente (nivel de significación 0.05) mejor/peor que el

Tabla 4 Modelos obtenidos para el algoritmo ID3

Aprobación de Curso	**DesigDocPlanta**	**Juramento**
parteResol = p_encab: NO parteResol = p_visto: NO parteResol = p_consid | tieneUA = SI | | tieneTitulo = SI | | | tieneCarrera = SI: NO | | | tieneCarrera = NO | | | | tienePersona = SI | | | | | tieneInstitucion = SI: NO | | | | | tieneInstitucion = NO: SI | | | | tienePersona = NO | | | | | tieneInstitucion = SI: SI | | | | | tieneInstitucion = NO: NO | | tieneTitulo = NO | | | tieneCarrera = SI: SI | | | tieneCarrera = NO | | | | tienePersona = SI: SI | | | | tienePersona = NO: SI | tieneUA = NO | | tieneTitulo = SI: SI | | tieneTitulo = NO | | | tieneCarrera = SI: NO | | | tieneCarrera = NO | | | | tienePersona = SI: NO | | | | tienePersona = NO: NO parteResol = p_artic | tieneUA = SI: SI | tieneUA = NO | | tienePersona = SI: SI | | tienePersona = NO | | | tieneInstitucion = SI: SI | | | tieneInstitucion = NO: SI	parteResol = p_encab: NO parteResol = p_visto: NO parteResol = p_consid: NO parteResol = p_artic | tieneCarrera = SI | | tieneUA = SI | | | tienePersona = SI: SI | | | tienePersona = NO | | | | tieneTitulo = SI: SI | | | | tieneTitulo = NO: SI | | tieneUA = NO: SI | tieneCarrera = NO | | tieneTitulo = SI: SI | | tieneTitulo = NO | | | tieneUA = SI | | | | tienePersona = SI: NO | | | | tienePersona = NO: NO | | | tieneUA = NO | | | | tienePersona = SI: SI | | | | tienePersona = NO: NO	parteResol = p_encab: NO parteResol = p_visto: NO parteResol = p_consid: NO parteResol = p_artic | tienePersona = SI: SI | tienePersona = NO | | tieneCarrera = SI: SI | | tieneCarrera = NO | | | tieneUA = SI: SI | | | tieneUA = NO | | | | tieneDNI = SI: SI | | | | tieneDNI = NO | | | | | tieneInstitucion = SI: SI | | | | | tieneInstitucion = NO: NO

esquema tomado como base (ADTree). Si no aparece un símbolo, no hay diferencia significativa entre la precisión de ambas técnicas. A partir de estos resultados se decidió utilizar el algoritmo ID3.

En la Tabla 4 se muestran los modelos (árboles de decisión) obtenidos por ID3 para las tres categorías específicas de resoluciones. Nótese que para la categoría Aprobación de Curso se seleccionan fragmentos de la sección de considerandos de la resolución y de la sección donde se describen los artículos; esto se debe a que la primera suele contener información interesante para el resumen tal como los objetivos, destinatarios, contenidos y responsables del curso. En el caso de las categorías de Designación de Planta Docente y Juramento los fragmentos elegidos pertenecen a la sección de artículos de cada resolución y en particular son los que contienen entidades de tipo Carrera, UA, Persona o Título.

Generación de resúmenes

El modelo aprendido etiqueta en cada resolución algunos fragmentos con el valor "SI", indicando que son los más relevantes y deben estar en el resumen. En nuestro esquema de generación de resúmenes por extracción, éste se forma por la concatenación de estos fragmentos.

La Tabla 5 muestra una resolución a la que se aplicó el modelo elegido y el resumen

Tabla 5 Ejemplo de resumen de una resolución de la categoría aprobación de curso

Texto de la Resolución Rectoral	Resumen obtenido
RESOLUCIÓN Nº 656/07 En el Campo Castañares, sito en la ciudad de Salta, Capital de la Provincia del mismo nombre, República Argentina, sede de la Universidad Católica de Salta, a los veintiún días del mes de junio de dos mil siete: **VISTO:** la presentación efectuada por las autoridades de la Facultad de Artes y Ciencias y de la Secretaría de Postgrado y Perfeccionamiento Docente, en virtud de la cual solicitan el auspicio de esta Universidad; y **CONSIDERANDO:** que se trata del Curso de Postgrado "IV CONGRESO SOBRE ABUSO SEXUAL INFANTO-JUVENIL", previsto para los días 05, 06 y 07 de Julio de 2007, en la ciudad de Salta; que su objetivo es: Abordar este difícil y complejo tema del abuso sexual, teniendo en cuenta la importancia de articular acciones desde las instituciones públicas y privadas, aunando esfuerzos con un mismo propósito, el de concienciar, desnaturalizar y prevenir situaciones de abuso y violencia sexual; que los ejes temáticos son: Abuso sexual- Jóvenes y adolescentes frente al delito; Aspectos Éticos en el diagnóstico del abuso sexual infanto-juvenil; ¿Buscamos siempre el interés superior del niño?; Psicología del Testimonio; Aspectos relevantes sobre la pericia Psicológica en los delitos sexuales; Aplicación e implementación de la cámara GESSELL; Aspectos Psicológico del abuso; El proceso de la sexuación. Factores filogenéticos, familiares y culturales que influyen en la misma; Delitos Sexuales; que está destinado a: Abogados, Psicólogos, Asistentes Sociales, Comunicadores, Periodistas, Docentes, Actores Sociales y Comunidad en general; que habiéndose dado intervención a la Unidad Académica correspondiente, es necesario dictar el instrumento legal que lo auspicie; que en Reunión de fecha 20.06.07, el Consejo Académico se expidió favorablemente; **EL RECTOR DE LA UNIVERSIDAD CATÓLICA DE SALTA** **RESUELVE** **Art.1°.-** Auspiciar la realización del Curso de Postgrado **"IV CONGRESO SOBRE ABUSO SEXUAL INFANTO-JUVENIL"**, organizado por la Fundación Lapacho, Facultad de Artes y Ciencias, Facultad de Ciencias Jurídicas, Secretaría de Postgrado y Perfeccionamiento Docente de la Universidad Católica de Salta, a llevarse a cabo los días 05, 06 y 07 de Julio de 2007, en la ciudad de Salta. **Art.2°.-** Comunicar a: Vicerrector Académico, Vicerrector Administrativo, Secretaría General, Unidades Académicas y Administrativas correspondientes, a los efectos a que hubiere lugar. **Art.3°.-** Registrar, reservar el original, publicar en el Boletín Oficial de la Universidad Católica de Salta y archivar.	Dos fragmentos: [que está destinado a: Abogados, Psicologos, Asistentes Sociales, Comunicadores, Periodistas, Docentes, Actores Sociales y Comunidad en general;] [Art 1.- Auspiciar la realizacion del Curso de Postgrado IV CONGRESO SOBRE ABUSO SEXUAL INFANTO-JUVENIL, organizado por la Fundacion Lapacho, Facultad de Artes y Ciencias, Facultad de Ciencias Jurídicas, Secretaria de Postgrado y Perfeccionamiento Docente de la Universidad Catolica de Salta, a llevarse a cabo los dias 05, 06 y 07 de Julio de 2007, en la ciudad de Salta.]

obtenido. En general uno o dos fragmentos son suficientes para obtener resúmenes de calidad en nuestro dominio, más aún dado el fin indicativo de los mismos, como ayuda al usuario del buscador semántico.

En general en nuestros experimentos la concatenación de estos fragmentos proporciona un texto suficientemente claro y coherente. No obstante es posible, por ejemplo en el caso de la categoría Aprobación de curso, que aparezcan documentos con una estructura o nivel de detalle dispar a los de la mayoría con lo que el resumen obtenido no sea un texto ordenado o suficientemente coherente.

Conclusión

En este trabajo hemos descrito la aplicación de técnicas de aprendizaje automático al problema de extracción de resúmenes. El dominio de aplicación es un corpus de más de 8000 documentos que contienen 9 años de resoluciones rectorales de la Universidad Católica de Salta. Dadas las características de los resúmenes precisados, de tipo indicativo, ya que servirán como ayuda al usuario de un buscador semántico para seleccionar entre los documentos obtenidos como relevantes a su consulta, ha sido suficiente utilizar algoritmos del tipo poco profundo. En particular los modelos de árboles de decisión se han mostrado adecuados. Las características del texto utilizadas para aprender estos modelos han sido la presencia o no de entidades con nombre en un fragmento, la ubicación del mismo en el documento, y el tipo de documento. Los fragmentos seleccionados por el modelo forman el resumen extraído. Los resúmenes obtenidos son de calidad suficiente para el fin pretendido.

La extracción de resúmenes integra un sistema de minería de textos cuyo objetivo es la búsqueda semántica de documentos relevantes en una colección. Complementa así aplicaciones anteriores del aprendizaje automático a la categorización de los documentos y a la extracción de entidades con nombre (NER). El buscador semántico, desarrollado sobre la arquitectura UIMA, sirve de plataforma sobre la que se integran el resto de los sistemas y facilita la experimentación y desarrollo de estas tecnologías. Se está continuando el trabajo en esta línea, explorando otros problemas como la búsqueda de respuestas y la extracción de relaciones entre entidades.

Agradecimientos

Este trabajo ha sido financiado en parte por el Consejo de Investigaciones de la Universidad Católica de Salta (Resol Rect 333/11).

Referencias

1. Perez, A., Cardoso, A. C.: Categorización automática de documentos. In : Simposio Argentino de Inteligencia Artificial, 40 JAIIO, Córdoba (2011)
2. Pérez, A., Cardoso, A.: Extracción de entidades con nombre. In : Simposio Argentino de Inteligencia Artificial, 42 JAIIO, Córdoba (2013)
3. Sparck Jones, K.: Automatic summarizing: factors and directions. In : Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
4. Mani, I., Maybury, M.: Introduction. In : Advances in Automatic Text Summarization. MIT Press (1999)
5. Radev, D., Hovy, E., McKeown, K.: Introduction to the Special Issue on Summarization. Computational Linguistics 28(4), 399-408 (2002)
6. Montiel Soto, R., García-Hernández, R., Ledeneva, Y., Cruz Reyes, R.: Comparación de tres modelos de texto para la generación automática de resúmenes. Procesamiento del Lenguaje Natural(43), 303-311 (2009)
7. Acero, I., Alcojor, M., Díaz, A., Gómez, J.: Generación automática de resúmenes personalizados. Procesamiento del lenguaje natural 27, 281-290 (Sept. 2001)
8. Alonso, L., Castellón, I., Climent, S., Fuentes, M., Padró, L., Rodríguez, H.: Approaches to Text Summarization: Questions and Answers. Revista Iberoamericana de Inteligencia Artificial 20, 34-52 (2003)
9. Lin, C. Y., Hovy, E.: Identifying topics by position. In : Proceedings of the Applied Natural Processing Conference , Washington DC, pp.283-290 (1997)
10. Anaya Sanchez, H., Pons Porrata, A., Berlanga Llavori, R.: Una panorámica de la construcción de extractos de un texto. Revista cubana de Ciencias Informáticas 1(1) (2006)
11. Ferrucci, D., Lally, A.: Building an example

application with the Unstructured Information Management Architecture. IBM Systems Journal 45(3) (2004)

12. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques 2nd edn. Morgan Kaufmann, San Francisco (2005)

Datos de Contacto

Alejandra Carolina Cardoso y M Alicia Pérez Abelleira. Facultad de Ingeniería e IESIING, Universidad Católica de Salta, Castañares. A4400 Salta. {acardoso, aperez}@ucasal.net