UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

INGENIERÍA EN SOFTWARE

“Evaluación de resúmenes de actualización guiados”

TESIS

PARA OBTENER EL TÍTULO DE
INGENIERO EN SOFWARE

PRESENTA:
Brisa Faridi Hernández Castañeda

DIRECTORES DE TESIS
DRA. YULIA NIKOLAEVNA LEDENEVA
M. EN C.C. JONATHAN ROJAS SIMÓN

TIANGUISTENCO, ESTADO DE MÉXICO NOVIEMBRE 2021

Resumen

Actualmente, el manejo de la información digital es indispensable para la realización de tareas que implican el análisis de grandes cantidades de textos. Una de las tareas que se encuentra dentro del área procesamiento automático de lenguaje natural es la generación automática de resúmenes. En esta tesis se va a trabajar la parte de la evaluación de resúmenes automáticos.

La evaluación automática de resúmenes mide automáticamente la calidad de los resúmenes. En esta tarea, la evaluación automática de resúmenes se realiza para la tarea de múltiples documentos, específicamente hacia los resúmenes de actualización guiados. La evaluación automática de resúmenes se realiza sin referencias humanas. La tesis abarca el problema de cómo evaluar los resúmenes de actualización guiados utilizando los métodos sin referencias humanas y los índices de correlación Pearson, Spearman y Kendall.

En esta tesis, los métodos de evaluación permitirán asignar puntajes de calidad a cada resumen. Básicamente, los métodos de evaluación que se utilizaron son parte de los sistemas ROUGE-C (ROUGE-C-1, 2, 3, L y SU4) y SIMetrix (Divergencia Jensen-Shannon suavizada y no suavizada). Los resultados reportados son prometedores dentro de los métodos del estado del arte, ya que nos da un enfoque hacia las evaluaciones realizadas y así verificar que métodos fueron los mejores.

Contenido

Ilustraciones

Tablas

CAPÍTULO 1

Introducción

En este capítulo se introduce al área de evaluación automática de resúmenes de actualización guiados, se presenta el planteamiento del problema, objetivo general, objetivos específicos, hipótesis y finalmente la estructura de la tesis.

1.1 Relevancia de la información

Actualmente, existe una gran cantidad de información digital. Debido al crecimiento constante, puede mostrarse de manera repetitiva o redundante, ocasionando que no se pueda tener acceso con facilidad. Este tipo de problemas suelen ocasionar que las personas tarden demasiado tiempo en leer toda la información, que sea de su interés contenida en el documento. Para ello, es importante contar con métodos y herramientas que permitan la evaluación de resúmenes de manera automática.

Un resumen puede ser definido de distintas maneras. Sin embargo (Ledeneva & García-Hernández, 2017) lo definen como:

> *El resumen es un documento escrito, el cual presenta las ideas más importantes de uno o más documentos, sin omitir información relevante.*

Uno de los objetivos principales del resumen de texto es dar a conocer la información más importante del documento, de manera más compacta del texto original, pero manteniendo su contenido principal, evitando que el usuario tarde mucho tiempo realizando la búsqueda de la información requerida (Suanmali et al., 2011).

Según (Ruvalcaba, C. y Alfredo Cerda Muñoz, 2004) un resumen representa el texto de manera breve, en el cual se extraen las ideas principales, dejando de lado contenido que no es útil.

Existen diferentes maneras de clasificar la generación de resúmenes, basados de acuerdo a la cantidad de documentos de entrada, según el propósito o el tipo de resumen a generar y se clasifican de la siguiente manera:

- **Resumen extractivo**

  Los resúmenes extractivos se encargan de reducir el contenido textual de un conjunto de documentos mediante la extracción de frases u oraciones (Vilchis Sepúlveda & Ledeneva, 2019). Según (Ledeneva, 2008), los resúmenes extractivos se basan en la selección de oraciones más relevantes de los textos originales.

- **Resúmenes abstractivos**

  Los resúmenes abstractivos se centran en la descripción de contenido y enfoque de un documento original, integrando conceptos con una cantidad menor de palabras (Vilchis Sepúlveda & Ledeneva, 2019). Además, son los encargados de entender el texto original, de tal manera que, al reescribir su contenido con menor número de palabras, no se pierde el enfoque del mismo (Ledeneva, 2008).

  Los resúmenes abstractivos tienen bastante similitud con los que son creados por los humanos, ya que son elaborados mediante conceptos específicos y combinación de palabras, los cuales requieren de algún recurso lingüístico.

Según (Matias Mendoza, 2013), para generar un resumen mediante una herramienta automática se necesita saber cuáles serán las características que se deben tomar en cuenta para que el resumen sea de calidad. Por ejemplo, debe ser legible, el contenido no debe ser redundante, que no exista complejidad en cuanto a los términos, para que sea parte del resumen.

Una de las tareas más importantes para este trabajo, es la de resúmenes guiados que han sido generados en los talleres de TAC, especialmente en TAC 2010, considerando que un resumen guiado es un documento breve que expresa las ideas más relevantes del documento fuente, basado en generar un resumen escrito de 100 palabras de

artículos de distintos temas, tales como, Desastres Naturales, Ataques, Seguridad, Salud, Recursos Naturales, entre otros.

De acuerdo con estudios realizados en las últimas décadas, se dice que la tarea de evaluación de resúmenes automáticos ha sido de suma importancia, ya que actualmente la comunidad científica ha estado utilizando ROUGE como un paquete estándar (Rojas-Simón, 2019), basado en la comparación automática entre un resumen y un conjunto de documentos de referencia generados por expertos humanos. Para este proceso, es necesario contar con resúmenes elaborados por humanos, es decir con referencias humanas, para que la evaluación pueda ser realizada.

Al tener una cantidad extensa de posibles resúmenes, se complica la evaluación de los mismos de manera independiente, ya que se tiene que realizar la evaluación de todos para encontrar el mejor. Es decir, el más parecido al de un humano.

Por lo tanto, se debe realizar una comparación entre los juicios de un evaluador automático respecto a una serie de juicios dados por humanos. Es decir, juicios humanos contra juicios de la máquina. Para que la medición sea realizada, normalmente se ocupan tres índices de correlación Pearson, Spearman y Kendall.

1.2 Planteamiento del problema

Hoy en día la evaluación de resúmenes sin referencias humanas ha sido un tema de alto impacto debido a que la generación de referencias humanas suele ser laboriosa y costosa, de aquí surge el problema de esta tesis.

¿Cómo evaluar los resúmenes de actualización guiados utilizando los métodos sin referencias humanas y los índices de correlación (Pearson, Spearman y Kendall)?

1.3 Objetivos

1.3.1 Objetivo general

Determinar la calidad de los resúmenes de actualización guiados, utilizando los métodos de evaluación que no utilizan referencias humanas y los índices de correlación de Pearson, Spearman y Kendall.

1.3.2 Objetivos específicos

- Utilizar resúmenes de la colección TAC 2010.
- Utilizar diferentes métodos de evaluación de resúmenes automáticos.
- Evaluar los métodos de evaluación, mediante los índices de correlación de Pearson, Spearman y Kendall.
- Comparar el desempeño de los evaluadores utilizados mediante los índices de correlación de Pearson, Spearman y Kendall.

1.4 Hipótesis

Si se realiza la evaluación de los resúmenes mediante los métodos que no utilizan referencias humanas, se podría conocer la eficacia de dichos métodos de evaluación con los índices de correlación (Pearson, Spearman y Kendall).

1.5 Estructura de la tesis

La estructura de la tesis está descrita de la siguiente manera:

Capítulo I

Introducción, planteamiento al problema, los antecedentes y los objetivos, la hipótesis.

Capítulo II

Descripción de marco teórico.

Capítulo III

Descripción de estado de arte.

Capítulo IV

Descripción de la metodología propuesta.

Capítulo V

Descripción del corpus TAC 2010

- Evaluación de los métodos
- Comparación con los métodos del estado de arte
- Descripción de los resultados y experimentos
- Presentación de resultados, análisis y otros.

Capítulo VI

conclusiones, trabajo futuro y comentarios finales.

CAPÍTULO 2

Marco teórico

En este capítulo se presentan los conceptos más importantes que permitirán un mejor entendimiento, en cuanto al contenido de cada apartado que se trabaja en la tesis.

2.1 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) es una rama de la inteligencia artificial y la ingeniería lingüística computacional, que entiende el lenguaje de la máquina para procesar la información comunicada. Es decir, trata de procesar el texto por su sentido y no como un archivo binario (Gelbukh, 2010).

El objetivo principal del PLN es construir sistemas y mecanismos que permitan la comunicación entre usuario y máquina por medio de un lenguaje natural (Ledeneva & García-Hernández, 2017).

Según (Alexander Gelbukh & Sidorov, 2006), el esquema que generalmente utilizan los sistemas y métodos que involucran el PLN es el siguiente:

- El texto no es procesado directamente. Primero debe ser transformado en una representación formal que conserva sus características relevantes para realizar la tarea o el método especificado.
- El programa principal es el encargado de manipular dicha representación, para que posteriormente sea transformada según la tarea.
- En caso de ser necesario, los cambios hechos en la representación de la respuesta generada se transforman en lenguaje natural.

De acuerdo con (Amigó et al., 2005), dentro de las principales tareas del PLN, se encuentran:

- Recuperación de información
- Extracción de información
- Traducción automática
- Sistemas de búsqueda de respuestas

- Generación automática de resúmenes de texto
- Minería de datos
- Análisis de sentimientos

En el PLN, se ha implementado la generación automática de resúmenes, como una tarea que busca la extracción de ideas principales del contenido de un texto. Los resúmenes generados a partir de esta tarea generalmente se clasifican de acuerdo a la estrategia de condensación. Por un lado, se encuentran los extractivos, los cuales se centran en la descripción de contenido y enfoque de un documento original, empleando una selección de oraciones sin alterar el contenido de ellas. Por otro lado, se encuentran los abstractivos, los cuales son generados a partir de una reducción del contenido textual de un conjunto de documentos mediante la fusión de frases y oraciones (Ledeneva & García-Hernández, 2017).

2.2 Resumen y tipo de resúmenes

Según (Rojas-Simón et al., 2021), un resumen automático de texto es una de las tareas de amplio estudios en el PLN, debido a que día a día se presenta un gran crecimiento de información digital. Por ello, se dice que la evaluación de resúmenes de texto es de suma importancia para medir el desempeño de distintos métodos, tomando en cuenta que este proceso presenta una alta complejidad.

Un resumen de texto automático es un proceso que selecciona un documento fuente y extrae el contenido más importante para que sea entendible para el usuario.

Es sumamente importante tener un sistema de resumen de texto ya que existe una gran expansión de información digital (Steinberger & Ježek, 2009).

2.2.1 Resumen extractivo

Los resúmenes extractivos se encargan de la selección y reducción del contenido textual de un conjunto de documentos mediante la extracción de frases u oraciones más relevantes de los textos originales (Vilchis Sepúlveda & Ledeneva, 2019) (Ledeneva, 2008). Por ejemplo, se tiene que los resúmenes extractivos son generados por una selección de oraciones como se observa en la ilustración 1.

Columbine Yearbook Honors Victims LITTLETON, Colo.
(AP) This year's Columbine High School yearbook features a special insert honoring the victims of the massacre at the school.
Columbine students began picking up their yearbooks Monday.
The yearbook shows the senior pictures of Eric Harris and Dylan Klebold, who killed 12 classmates and a teacher inside the school April 20 before taking their own lives.
It also has a special insert remembering the slain students and teacher.
Senior Megan Fasano was among the students who picked up a yearbook Monday night at a church near the still-closed school.
She and her friends gave the yearbook called Rebelations for the school's team name, the Rebels a quick glance to see the photos of Harris and Klebold.
It's hard to see Dylan's and Eric's face, Ms. Fasano said.
Another senior, Lesley Pech, objected to the photos.
I don't think they should have their photos in the yearbook, she said.
They look happy and innocent, but it's them who ruined everything.
But Terra Oglesbee, a senior who had classes with Harris, said the two deserve photos because they were students.
They were people, too.
While their senior pictures are included, the two are not pictured in the special four-page memorial supplement.
It has pictures of their 13 victims with their names and birthdates.
The front has the words We are still Columbine and a picture of the flower that gives the school its name.
On the back, the lyrics of Columbine, Friend of Mine, the song written in memory of the tragedy, are superimposed on a picture of the school.
The yearbook club had the option of stopping publication of the book after the rampage, but opted to go ahead.
The main book, featuring a solid red cover, was essentially completed in February.
The club added one insert with pictures of the prom and other activities that took place since February and another for those killed in the rampage.
The staff didn't want to start all over, said Eric Friesen, an English teacher who is yearbook director at Columbine.
Friesen said he expects about 90 percent of Columbine's 1,900 students to buy this year's book, about the same percentage as in previous years.
For some of those students, seeing pictures of their slain friends will be bittersweet.
It feels different and weird to look at a yearbook now, Ms. Fasano said.

Ilustración 1. Ejemplo de resúmenes extractivos generados por selección de oraciones.

2.2.2 Resumen abstractivo

Los resúmenes abstractivos se centran en entender y describir el contenido y enfoque de un documento original, una vez entendido el contexto del documento y al ser reescrito, no se pierda el enfoque del texto, aun con la integración de nuevos conceptos con una cantidad menor de palabras (Vilchis Sepúlveda & Ledeneva, 2019) (Ledeneva, 2008).

Los resúmenes abstractivos tienen bastante similitud con los que son creados por los humanos, ya que son elaborados mediante conceptos específicos y combinaciones de palabras, los cuales requieren de algún recurso lingüístico.

Por ejemplo, para el caso de resúmenes abstractivos se muestra que las palabras fueron cambiadas en comparación con el resumen extractivo sin perder el enfoque, tal como se muestra en la ilustración 2.

LITTLETON, Colo.
(AP) -- This year's Columbine High School yearbook features a special insert honoring the victims of the massacre at the school.
Columbine students began picking up their yearbooks Monday.
The yearbook shows the senior pictures of Eric Harris and Dylan Klebold, who killed 12 classmates and a teacher inside the school April 20 before taking their own lives.
It also has a special insert remembering the slain students and teacher.
Senior Megan Fasano was among the students who picked up a yearbook Monday night at a church near the still-closed school.

Ilustración 2. Ejemplo de resúmenes abstractivos con palabras actualizadas.

2.2.3 Resumen guiado

El resumen guiado es un documento breve que expresa las ideas más importantes de un documento fuente, que ha sido actualizado a través del tiempo.

La tarea de resúmenes guiados se basa en generar un resumen de 100 palabras, a partir de un conjunto de artículos de noticias que hablen de un tema en particular, tomando en cuenta que cada tema pertenece a una categoría predefinida, el apartado que le da sentido a esta tarea, es que se le agrega un componente de actualización.

Tomando en cuenta que, para la evaluación de resúmenes de actualización guiados, el taller Text Analysis Conference 2010 (TAC 2010) cuenta con una plantilla que responde una serie de preguntas de acuerdo a la categoría del tema, la cual nos ayudará a verificar si realmente el resumen actualizado del documento fuente mantiene el mismo enfoque. A continuación, se presentan los apartados de dicha plantilla:

Accidentes y desastres naturales:

1. ¿Qué?: qué pasó.
2. ¿Cuándo?: fecha, hora, otros marcadores de ubicación temporal.
3. ¿Dónde?: ubicación física.
4. ¿Por qué?: motivos del accidente / desastre.
5. ¿Quiénes son los afectados?: víctimas (muerte, lesiones) o personas afectadas negativamente por el accidente / desastre.
6. Daños: daños causados por el accidente / desastre.
7. Contramedidas: contramedidas, esfuerzos de rescate, esfuerzos de prevención, otras reacciones al accidente / desastre.

Ataques (criminales / terroristas):

1. ¿Qué?: qué pasó.
2. ¿Cuándo?: fecha, hora, otros marcadores de ubicación temporal
3. ¿Dónde?: ubicación física.

4. Perpetradores: individuos o grupos responsables del ataque
5. ¿Por qué?: motivos del accidente / desastre.
6. ¿Quiénes son los afectados?: víctimas (muerte, lesiones) o personas afectadas negativamente por el accidente / desastre.
7. Contramedidas: contramedidas, esfuerzos de rescate, esfuerzos de prevención, otras reacciones al accidente / desastre.

Salud y seguridad:

1. ¿Qué?: qué pasó.
2. ¿Cuándo?: fecha, hora, otros marcadores de ubicación temporal.
3. ¿Dónde?: ubicación física.
4. Perpetradores: individuos o grupos responsables del ataque.
5. ¿Por qué?: motivos del ataque.
6. ¿Quiénes son los afectados?: bajas (muerte, heridas) o personas afectadas negativamente por el ataque.
7. Daños: daños causados por el ataque.
8. Contramedidas: contramedidas, esfuerzos de rescate, esfuerzos de prevención, otras reacciones al ataque (por ejemplo, investigaciones policiales).

Recursos en peligro:

1. ¿Qué?: descripción del recurso.
2. Importancia: importancia del recurso.
3. Amenazas: amenazas al recurso.
4. Contramedidas: contramedidas, esfuerzos de prevención.

Investigaciones y juicios (penales / legales / otros):

1. ¿Quién?: quién es el acusado o está bajo investigación.

2. ¿Quién investiga?: quién está investigando, procesando o juzgando.
3. ¿Por qué?: razones generales de la investigación / juicio.
4. Cargos: cargos específicos al acusado.
5. Declaración: reacción del acusado a los cargos, incluida la admisión de culpabilidad, la negación de cargos o explicaciones.
6. Sentencia: sentencia u otras consecuencias para el imputado.

Para generar un resumen de actualización guiado, normalmente se siguen los siguientes pasos.

Paso 1: Tomando en cuenta la plantilla antes mencionada, primero se debe identificar el ID del documento fuente, posteriormente se identifica el título y finalmente tenemos el contenido del mismo, el cual presenta toda la información de la noticia. Como se puede observar de la ilustración 3, el documento fuente contiene una extensa cantidad de información obtenida del tema de la noticia.

Paso 2. De acuerdo con el documento fuente, como se puede observar de la ilustración 3, contiene una cantidad extensa de información. Por ello es importante la tarea de resúmenes de actualización guiados, ya que se genera un resumen únicamente de 100 palabras, el cual abarca todo el contenido de la plantilla antes mencionada, mostrando como resultado un resumen actualizado con menor cantidad de información del documento fuente sin perder el enfoque del mismo. La Ilustración 4 muestra un ejemplo de un resumen guiado, el cual fue obtenido a partir de la plantilla mostrada anteriormente.

<DOC>
<DOCNO> XIE19980718.0080 </DOCNO> ID
<DATE_TIME> 1998-07-18 </DATE_TIME>
<BODY>
<HEADLINE> PNG Defense Force and Police Standby to Help Tsunami </HEADLINE> Título
<TEXT>
<P>
CANBERRA, July 18 (Xinhua) -- The Papua New Guinea (PNG) Defense Force, the police and health services are on standby to help the victims of a tsunami that wiped out several villages, killing scores of people, on PNG's remote north-west coast Friday
night.
</P> <P>
"Prime Minister Bill Skate has directed that immediate action be taken to arrange urgent supplies of food, water, shelter and cooking utensils while detailed assessments and reports are being prepared," Robert Igara, chief secretary of the PNG government, said Saturday in a statement, Australian Associated Press reported.
</P> <P>
Igara said the PNG Red Cross had confirmed arrangements to provide food supplies and authorities had asked the Australian High Commission in Port Moresby for immediate air transport support.
</P> <P>
The death toll so far from the tsunami that struck the coastline near the town of Aitape in the West Sepik (Sandaun) province Friday night, was officially announced 64 and many more people were missing.
</P>
<P>
Igara said reports so far indicated that a community school, government station, Catholic mission station and the Nimas village in the Sissano area west of Aitape had been completely destroyed, where 30 people were dead.
</P>
<P>
And Warapu village had also been completely destroyed, with 11 confirmed deaths and many missing.
</P>
<P>
Sixteen people in the devastated Arop village were confirmed dead, and two children had died in the Malalo village, which was wiped out.
</P>
<P>
Another five people died in the Teles-Lambu villages.
</P>
<P>
Igara said the population in the area affected by the tsunami was 8,000 to 10,000 people.
</P>
<P>
"Officers are now visiting the affected villages and areas to establish the whereabouts, safety and health of the population," he said.
</P>
<P>
Meanwhile, Prime Minister Bill Skate, National Disaster Services Chairman Colin Travertz and National Disaster Services Director-General Ludwig Kembu were reported to inspect the affected areas Sunday.
</P>
</TEXT>
</BODY>
</DOC>

Contenido

Ilustración 3. Ejemplo documento fuente.

A tsunami spawned by a 7.0 magnitude earthquake crashed into Papua New Guinea's north coast, crushing villages and leaving hundreds missing, officials said Sunday.
The death toll in Papua New Guinea's (PNG) tsunami disaster has climbed to 599 and is expected to rise, a PNG disaster control officer said Sunday.
Authorities at Aitape in the West Sepik province, on Papua New Guinea's northwest coast, said the tsunami that hit the coast west of Aitape on Friday night had wiped out three villages and had almost completely destroyed another.
Dalle said that at Warapu 500 people had been confirmed dead.

Ilustración 4. Ejemplo de resumen de actualización guiado.

2.3 Evaluación de resúmenes

La evaluación de la calidad de un resumen es una tarea muy ambiciosa. Ya que se encarga de responder preguntas serias que permanecen en relación con los métodos y tipos de evaluación adecuados. El resumen humano puede ser proporcionado por el autor del artículo, por un juez solicitado para construir un resumen, o por un juez al que se le pide que extraiga sentencias (Steinberger & Ježek, 2009).

De acuerdo con (Amigó et al., 2005), la calidad de un resumen automático puede utilizar dos enfoques, ya sea por evaluaciones humanas que se basa en el resultado de una serie de sistemas comparados por el hombre utilizando un conjunto de evaluación por pautas, o mediante la similitud que existe en el contenido del documento fuente.

Con las evaluaciones humanas, es posible identificar cuando un sistema está trabajando de manera adecuada y cuando no lo está haciendo. No obstante, un sistema de evaluación no suele ser viable cuando sus criterios no se asemejan a los juicios humanos.
Un resumen que tiene mayor similitud con el contenido de texto fuente es considerado el más factible que uno con menor similitud. Entre más similar sea un resumen a su texto

fuente, mejor será su contenido, ya que el resumen presenta menor pérdida de información durante el proceso (Louis & Nenkova, 2013).

2.4 Índices de correlación

Una forma común de conocer el desempeño de un evaluador es mediante una comparación entre los puntajes del evaluador automático, respecto a una serie de juicios humanos. Dicha comparación puede ser calculada mediante los índices de correlación de Pearson, Spearman y Kendall.

2.4.1 Correlación de Pearson

El coeficiente de correlación de Pearson es el encargado de medir el grado de relación entre dos variables relacionadas de manera lineal. Este coeficiente de correlación utiliza la Ecuación 1 donde r representa el coeficiente de correlación de Pearson; N representa el tamaño de los vectores X y Y; $\sum x_i y_i$ es la suma de los productos de cada par de valores $x_i\ y_i$; Σxi es la suma de los valores calculados a partir del vector X; $\sum y_i$ es la suma de los valores calculados a partir del vector Y; $\sum {x_i}^2$ es la suma de los valores de x_i al cuadrado y $\sum y_i$ es la suma de los valores de yi al cuadrado (Rojas, 2019).

$$r = \frac{\mathrm{N} \sum x_i y_i - (\sum x_i \sum y_i)}{\sqrt{[N \sum x_i^2 - (\sum x_i)^2][N \sum y_i^2 - (\sum y_i)^2]}} \qquad (1)$$

2.4.2 Correlación de Spearman

El coeficiente de correlación de rango de Spearman es el encargado de medir un grado de asociación no paramétrico entre dos variables independientes. Este coeficiente de

correlación se mide en una escala ordinal para ser clasificadas en dos series ordenadas. En la Ecuación 2, se define el coeficiente de correlación de rango de Spearman, donde r_s representa el coeficiente de correlación de Spearman; di representa la diferencia entre cada valor de x_i y y_i; N representa el tamaño de las variables X y Y (Rojas, 2019).

$$r_s = 1 - \frac{6 \text{ X } \sum d_i^2}{n(n^2 - 1)} \qquad (2)$$

2.4.3 Correlación de Kendall

El coeficiente de correlación de rango de Kendall (también conocido como Kendall tau, τ) es una medida que determina el grado de asociación entre dos variables utilizando un vector ordenado. Las observaciones de una variable se ordenan de mayor a menor para medir el número de acuerdos y desacuerdos entre las clasificaciones de ambas variables. El coeficiente de correlación de rango de Kendall se calcula a partir de la Ecuación 3, donde S representa el acuerdo o desacuerdo total de las observaciones no ordenadas. Si la relación de cada valor x_i y y_i es directamente proporcional, entonces cada contribución es +1. De lo contrario, cada contribución disminuye en una tasa de -1; n representa el tamaño de las variables X y Y (Rojas, 2019).

$$\tau = \frac{\text{S}}{\frac{1}{2}n(n-1)} \qquad (3)$$

2.5 Niveles de evaluación

Los niveles de evaluación son formas de comparación entre la evaluación automática y la manual que ayudan a medir la similitud entre ambos. En específico, la micro y macro

evaluación, las cuales son formas conocidas para realizar la comparación y el desempeño de métodos automáticos con referencia a juicios humanos. A continuación, se explican los niveles de evaluación antes mencionados.

2.5.1 Micro evaluación

La micro evaluación se centra principalmente en medir la similitud de la evaluación, en comparación con la evaluación manual, utilizando los puntajes de evaluación de cada resumen candidato.

Según (Louis & Nenkova, 2013) y (Rojas-Simón et al., 2021), la micro evaluación se caracteriza por considerar los siguientes aspectos.

- Todos los resúmenes candidatos son evaluados de manera automática y manual.
- Posteriormente, se compara la puntuación obtenida de ambas evaluaciones, mediante los coeficientes de correlación de Pearson, Spearman y Kendall.

2.5.2 Macro evaluación

A diferencia de la micro evaluación, la macro-evaluación utiliza el promedio de la puntuación de los métodos de resumen de texto automático, evaluados de manera automática y manual. Tomando en cuenta, que el resultado de ambas evaluaciones se compara a través de los índices de correlación de Pearson, Spearman y Kendall. Generalmente, la macro evaluación se encarga de medir el grado de previsibilidad de la evaluación automática contra la evaluación manual, utilizando las puntuaciones de los métodos de texto automático (Rojas-Simón et al., 2021), (Louis & Nenkova, 2013).

CAPÍTULO 3

Estado del arte

En este capítulo se presentan los trabajos relacionados del estado del arte sobre la evaluación automática de resúmenes guiados.

3.1 ROUGE: A Package for Automatic Evaluation of Summaries (Lin, 2004)

ROUGE son las siglas de Recall-Oriented Understudy for Gisting Evaluation, el cual ha sido un paquete de evaluación de resúmenes automáticos que puede realizar una evaluación de varios resúmenes, comparando su contenido con respecto a un conjunto de referencias humanas. Actualmente, se encuentra disponible para la investigación, a través de otros métodos de evaluación.

3.1.1 Descripción

En este trabajo, se utilizaron diferentes métodos para determinar automáticamente la calidad de un resumen en comparación con resúmenes creados por humanos. Tales métodos emplean diferentes modelos de representación de texto que ayudan a calcular la similitud entre el resumen generado por la computadora y los resúmenes creados por humanos. En general, ROUGE emplea los métodos ROUGE-N, ROUGE-L, ROUGE-W y ROUGE-S se han utilizado en los talleres Document Understanding Conferences (DUC). Además, se presenta una compilación extensa de los corpus y conjuntos de datos que están disponibles para la investigación y trabajar dentro del área de evaluación de resúmenes.

3.1.2 Evaluación

Generalmente, todos los métodos de ROUGE miden la cantidad de información relevante a partir de un conjunto de referencias humanas. De esta manera, el resumen deberá contener toda la información relevante cuando es lo suficientemente largo y semejante a las referencias humanas.

Como se ha mencionado previamente, ROUGE emplea diferentes métodos de evaluación, los cuales se describen a continuación:

3.1.3 ROUGE-N: Estadísticas de ocurrencia de N-gramas

ROUGE-N mide la co-ocurrencia estadística n-gramas entre un resumen candidato y un conjunto de resúmenes de referencia, como se muestra en la ecuación 4.

$$ROUGE - N = \frac{\sum_{S\epsilon\ (Referencias\ Humanas)} \sum_{grama_n\ \epsilon S} \sum Conteo_{concordantes}(grama_n)}{\sum_{S\epsilon\ (Referencias\ Humanas)} \sum_{grama_n\ \epsilon S} \sum Conteo\ (grama_n)} \quad (4)$$

donde n es la longitud del n-grama y $Conteo_{concordantes}(grama_n)$ el número máximo de n-gramas que co-ocurren en el resumen candidato y el conjunto de resúmenes de referencia.

Además, ROUGE-N se considera una medida basada en la especificidad (recall en inglés) (Sidorov, 2013), porque el denominador de la ecuación es el total de la suma del número de n-gramas ocurrentes en los resúmenes de referencia.

3.1.4 ROUGE-L

ROUGE-L considera que una secuencia, es una serie de elementos que suceden unos a otros y guardan relación entre sí, la subsecuencia común más larga entre dos resúmenes es la continuación de una secuencia, $Z = [z_1, z_2, z_n]$ es una subsecuencia de otra secuencia $X = [x_1, x_2, x_m]$, dado que si existe una secuencia creciente estricta $[i_1, i_2, i_k]$ de índices de X, tal que para $j = 1, 2, k$, se tiene $x_{ij} = z_j$. Dadas dos secuencias X y Y, se considera que la subsecuencia común más larga (LCS) de X y Y es una subsecuencia común de longitud máxima.

3.1.5 ROUGE-W

La subsecuencia común más larga ponderada de ROUGE (ROUGE W) tiene muchas propiedades. Desafortunadamente, la LCS también tiene el problema de no diferenciar las relaciones espaciales dentro sus secuencias de incrustación. Por ejemplo, dada una secuencia de referencia X y dos secuencias candidatas Y_1 y Y_2 de la siguiente manera:

X: [A B C D E F G]
Y1: [A B C D H I K]
Y2: [A H B K C I D]

En este caso, Y_1 obtiene un puntaje bajo al ser evaluada, y debería ser mejor opción que Y_2 porque Y_1 tiene coincidencias consecutivas.

3.1.6 ROUGE-S

ROUGE-S mide la co-ocurrencia estadística de ocurrencias de bigramas, es decir, utiliza bigramas de saltos que permiten espacios arbitrarios. Los bigramas son estadísticas de co-ocurrencia que miden la superposición de los bigramas entre un resumen candidato y un conjunto de resúmenes de referencias humanas.

3.1.7 Resultados

Para este apartado se utilizaron 17 medidas de ROUGE para cada una de las ejecuciones, utilizando los paquetes de evaluación ROUGE antes mencionados, tomando en cuenta los índices de correlación de Pearson, Spearman y Kendall.

En la Tabla 1 se muestran los coeficientes de correlación de Pearson de 17 medidas ROUGE en comparación a los juicios humanos en las colecciones DUC 2001 y 2002, a partir de resúmenes de un solo documento de 100 palabras. Como se puede observar, el conjunto de datos ROUGE-2 obtuvo un mejor desempeño en comparación con los

paquetes de ROUGE-N, ROUGE-L, ROUGEW y ROUGE-S. En general, todas las medidas de ROUGE lograron una buena correlación con los juicios humanos en datos de DUC 2002.

Tabla 1. Correlaciones de Pearson de 17 medidas de ROUGE.

	Duc 2001 Documento único de 100 palabras						Duc 2002 Documento único de 100 palabras					
	1 REFERENCIA			3 REFERENCIAS			1 REFERENCIA			2 REFERENCIAS		
ROUGE-1	0.76	0.76	0.84	0.8	0.78	0.84	0.98	0.98	0.99	0.98	0.98	0.99
ROUGE -2	0.84	0.84	0.83	0.87	0.87	0.86	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -3	0.82	0.83	0.8	0.86	0.86	0.85	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -4	0.81	0.81	0.77	0.84	0.84	0.83	0.99	0.99	0.98	0.99	0.99	0.99
ROUGE -5	0.79	0.79	0.75	0.83	0.83	0.81	0.99	0.99	0.98	0.99	0.99	0.98
ROUGE -6	0.76	0.77	0.71	0.81	0.81	0.79	0.98	0.99	0.97	0.99	0.99	0.98
ROUGE -7	0.73	0.74	0.65	0.79	0.8	0.76	0.98	0.98	0.97	0.99	0.99	0.97
ROUGE -8	0.69	0.71	0.61	0.78	0.78	0.72	0.98	0.98	0.96	0.99	0.99	0.97
ROUGE -9	0.65	0.67	0.59	0.76	0.76	0.69	0.97	0.97	0.95	0.98	0.98	0.96
ROUGE -L	0.83	0.83	0.83	0.86	0.86	0.86	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -S	0.74	0.74	0.8	0.78	0.77	0.82	0.98	0.98	0.98	0.98	0.97	0.98
ROUGE -S4	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -S9	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -SU	0.74	0.74	0.81	0.78	0.77	0.83	0.98	0.98	0.98	0.98	0.98	0.98
ROUGE -SU4	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE -SU9	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
ROUGE-W-1.2	0.85	0.85	0.85	0.87	0.87	0.87	0.99	0.85	0.85	0.85	0.87	0.87

3.2 Evaluation of text summaries without human references based on the linear optimization of content metrics using a genetic algorithm (Rojas-Simón et al., 2021)

3.2.1. Descripción

En este trabajo, se describe un conjunto de métodos de evaluación para el análisis del contenido de un resumen, utilizando sus documentos fuente como referencia. Además,

se propone una optimización lineal de 31 métricas de contenido utilizando un Algoritmo Genético (AG).

3.2.2. Evaluación

Para este trabajo, se propusieron métricas de evaluación utilizando un algoritmo genético para la mejora entre la correlación de una evaluación de resumen automático y la evaluación manual.

3.2.3. Métricas de ROUGE-C-N

Se utilizaron cinco métricas de evaluación derivado del método ROUGE-C-N (ROUGE-C-1, 2, 3, 4 y 5). Además de las métricas ROUGE-C-N, también se utilizaron otras métricas basadas en la extracción y evaluación de LCS y bigramas con saltos de 4 palabras (SU4) para obtener ROUGE-C-L y ROUGE-C-SU4, respectivamente (Rojas, 2021).

3.2.4. Métricas de LSA

El Análisis Semántico Latente (LSA, por sus iniciales en inglés) es un método algebraico utilizado para evaluar resúmenes con o sin referencias humanas. Específicamente, el LSA se ha utilizado para la comparación de temas principales del resumen candidato con su documento fuente. En dicha comparación, ambos documentos se representan en dos matrices A de m x n dimensiones, donde n representa el número de oraciones y m representa el número de términos. El valor interno de cada matriz (a_{i_j}) se calcula a partir de la multiplicación de la ponderación de términos local y global, como se observa en la ecuación 5, donde L_{i_j} denota el peso local para el término j de la oración i, y G_{i_j} representa el peso global del término j en todo el documento.

$$a_{i_j} = L_{i_j} \times G_{i_j} \qquad (5)$$

Para la generación de métricas del método LSA es indispensable la combinación de esquemas globales de ponderación. Por ello, se utilizó el nombre LSA (LW, GW) para diferenciar cada métrica, donde LW representa el peso local y GW representa el peso global.

A continuación, se muestran los esquemas de ponderación de términos que se utilizan para asignar el peso a cada término (L_{i_j}):

- Peso de frecuencia (FW): $L_{i_j} = tf_{i_j}$, donde tf_{i_j} representa el número de veces que el término j se incluye en la oración i.
- Peso binario (BW): L_{i_j} = 1, si el término j se incluye al menos una vez en la oración i. De lo contrario, L_{i_j} = 0.
- Peso aumentado (AW): L_{i_j} = 0.5 + 0.5 × (tf_{i_j} / tf_{max_i}), tf_{max_i} representa el valor de frecuencia del término más frecuente en la oración i.
- Peso logarítmico (LW): $L_{i_j} = log\ (1 + tf_{i_j})$

Los siguientes esquemas de ponderación de términos se utilizan para asignar la ponderación global para cada j^{th} término de cada i^{th} oración (G_{ij}):

- Sin peso (NW): Este esquema no establece cambios para ningún término j ($G_{ij} = 1$).
- Frecuencia de oración inversa (ISF): $G_{ij} = log\ (N\ /\ n_j) + 1$, donde N representa el número de oraciones en el documento y n_j es el número de oraciones que contienen el término j.
- GFidf (GF): $G_{ij} = gf_j\ /\ sf_j$, donde sf_j (frecuencia de la oración) representa el número total de oraciones en las que aparece el término j, gf_j representa el número de veces que se incluye ese término j en todo el documento.

- Frecuencia de entropía (EF): $G_{ij} = 1 - \sum_{i=1}^{n}\left(\frac{p_{ij\log(P_{ij})}}{\log(n_{sent})}\right)$, donde $P_{ij} = \frac{tf_{ij}}{gf_j}$ y n_{sent} es el número de oraciones en el documento.

3.2.5. Métricas de SIMetrix

Es una evaluación basada en 10 métricas de similitud / distancia, para la evaluación de resúmenes sin referencias humanas. Sin embargo, se tiene que las métricas con los mejores resultados de correlación se derivan de Kullback-Leibler (D_{KL}) y Jensen Shannon (D_{JS}) como medidas de divergencia. Tomando en cuenta que D_{KL} y D_{JS} fueron propuestas inicialmente para medir la magnitud de información entre dos señales de comunicación, ambas medidas han sido utilizado en el ETS para medir la pérdida de información del candidato resumen.

Las mejores métricas de evaluación de SIMetrix se derivan de las divergencias D_{KL} y D_{JS}. Sin embargo, debido a los problemas de D_{KL} solo utilizó la métrica D_{JS} en la experimentación. De D_{JS}, se han extraído n-gramas con longitudes de 1 a 4 para generar diferentes métricas. Además, los autores utilizaron funciones de probabilidad suavizadas y no suavizadas. En total, ocho métricas derivadas de la divergencia D_{JS} se utilizaron en la optimización lineal.

3.2.6. Resultados

Los experimentos se ejecutaron a nivel de micro-evaluación para probar el desempeño del AG, observando que el rendimiento del AG en el conjunto de prueba (DUC01) aumenta cuando el AG ha obtenido altos valores de aptitud en el entrenamiento del conjunto (DUC02).

Tabla 2. Correlación de métricas de evaluación a nivel de micro evaluación utilizando el conjunto de datos DUC 2001.

Métricas	Pearson	Spearman	Kendall
Optimizada	0.46735	0.42968	0.30012
ROUGE-C-2	0.44499	0.40609	0.28256
ROUGE-C-3	0.44388	0.40571	0.28229
ROUGE-C-SU4	0.44678	0.41110	0.28618
LSA(*AW,NW*)	0.45454	0.41748	0.29042
LSA(*AW,GF*)	0.45195	0.41299	0.28722
LSA(*AW,EF*)	0.45946	0.42337	0.29473
D_{JS1}	0.47534	0.44154	0.30827
D_{JS2}	0.45127	0.40524	0.28227
D_{JS3}	0.43763	0.39191	0.27260
$D_{JS\ Smth1}$	0.47393	0.44136	0.30790

Como se puede observar en la Tabla 2, la métrica optimizada, D_{JS1} y $D_{JS\ Smth1}$ tuvieron mejor correlación que las otras métricas evaluadas.

3.3 Evaluación de resúmenes automáticos con y sin resúmenes de referencia para el idioma inglés (Vilchis Sepúlveda & Ledeneva, 2019)

3.3.1. Descripción

Se realiza una evaluación de los resúmenes automáticos generados por las herramientas comerciales, con los evaluadores que utilizan un resumen modelo humano y los que utilizan el texto original, donde se observó que cada evaluador toma una parcialidad diferente.

3.3.2. Evaluación

En este apartado, se realizó la evaluación de resúmenes automáticos, resúmenes de referencia humana, medidas, Jensen Shannon, ROUGE-C, ROUGE-1, ROUGE-2, con herramientas comerciales.

3.3.3. ROUGE-N

Este sistema trabaja mediante la medición de especificidad (conocida como Recall en inglés) de n-gramas entre un resumen candidato y un resumen ideal creado por un humano y se calcula mediante la ecuación 4,

Por otro lado, la evaluación de resúmenes sin referencias humanas con ROUGE-C-N está basado bajo la siguiente ecuación 6:

$$ROUGE - C - N = \frac{\sum s \in \{Resumen\ candidato\} \sum grama_{n \in s}\ Conteo_{concordantes}(grama_n)}{\sum s \in \{Documentos\ fuente\} \sum grama_{n \in s}\ Conteo(grama_n)} \quad (6)$$

donde n determina el uso del n-grama a evaluar a través de su longitud, $Conteo_{concordantes}(grama_n)$ es el máximo número de co-ocurrencia de n-gramas entre el resumen generado por el sistema y el documento fuente, tal como se observa en la ecuación 6.

En el método ROUGE se plantea el intercambio entre documentos, es decir, el documento fuente se coloca como resumen de evaluación y el resumen generado se coloca como un resumen de evaluación, mientras que el resumen arrojado por el sistema es colocado como un resumen de referencia como se muestra en la ilustración 5.

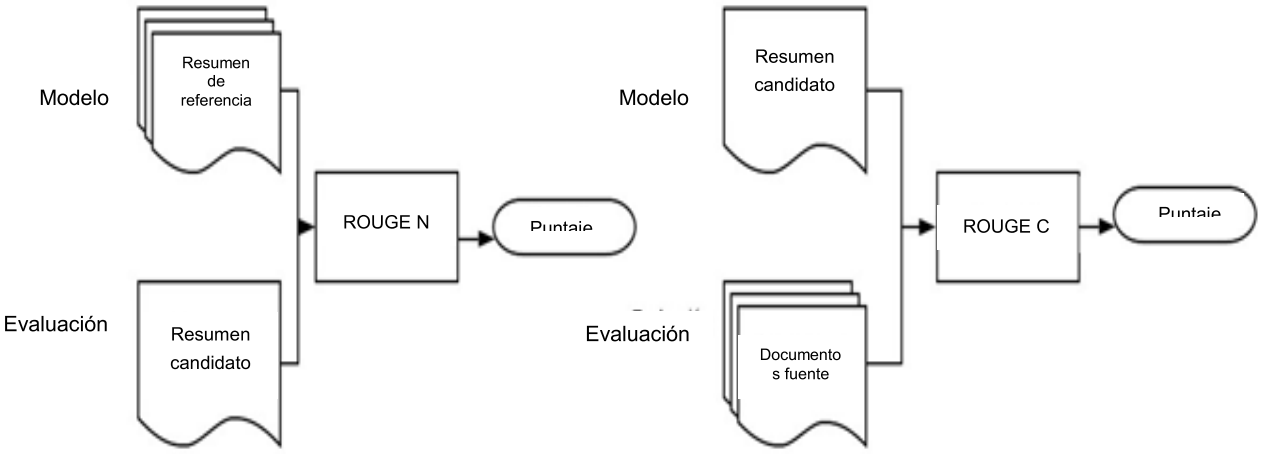


Ilustración 5. Evaluación de ROUGE y ROUGE-C.

3.3.4. Resultados

Para el primer paso, se realizó la generación de los resúmenes automáticos por las herramientas comerciales. Posteriormente, una vez obtenidos los resúmenes automáticos, se realizó la primera evaluación con el evaluador ROUGE-1. Los resultados obtenidos de la Tabla 3 e ilustración 6, muestran que la herramienta que obtuvo los resultados con mayor similitud contra los generados por un humano fue SweSum.

La segunda evaluación de resúmenes automáticos fue realizada con el evaluador ROUGE-2, los resultados que se obtuvieron de la Tabla 4 demostraron que SweSum sigue siendo la herramienta en línea encargada de generar resúmenes automáticos con mayor similitud a los generados por un humano, sin embargo, el puntaje obtenido con este evaluador disminuyó.

Tabla 3. Evaluación de diferentes herramientas comerciales usando ROUGE-1.

Herramienta	Recuerdo	Precisión	F-measure
SweSum	**0.4363**	**0.4331**	**0.4346**
OTS	0.4258	0.4239	0.4246
Text-Compactor	0.4224	0.4195	0.4208
T-Conspectus	0.4160	0.4107	0.4132
Summarizing	0.4105	0.4073	0.4087
Summarizer	0.4095	0.4065	0.4079
Word Office 2007	0.3960	0.3973	0.3965
Word Office 2003	0.3947	0.3962	0.3953

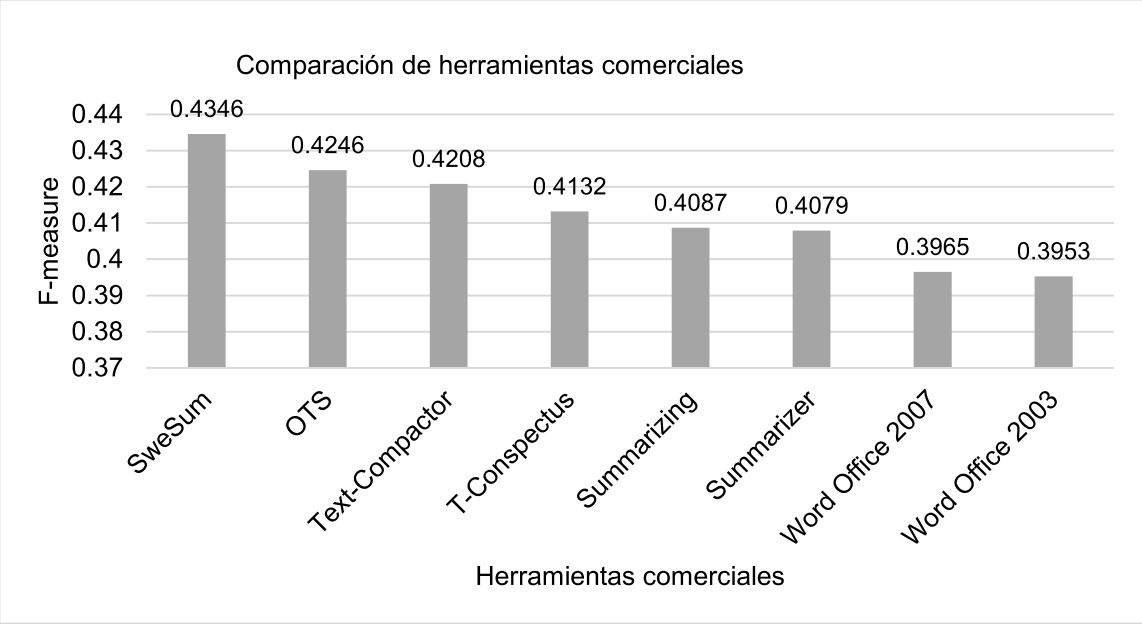


Ilustración 6. La herramienta con mayor puntaje fue SweSum.

Tabla 4. Evaluación de diferentes herramientas comerciales usando ROUGE-2.

Herramienta	Recuerdo	Precisión	F-measure
SweSum	**0.43626**	**0.43313**	**0.43459**
OTS	0.42577	0.42386	0.42458
Text-Compactor	0.42243	0.41954	0.42084
T-Conspectus	0.41595	0.410709	0.41321
Summarizing	0.410508	0.40726	0.40874
Summarizer	0.40947	0.40646	0.40786
Word Office 2007	0.39597	0.397309	0.39651
Word Office 2003	0.39469	0.39618	0.39533

Por otro lado, las evaluaciones realizadas con el evaluador ROUGE-C demuestran que la herramienta compactadora de texto tiene mayor similitud al texto original mientras que SweSum, por lo que resulta ser una herramienta de poca competencia para esta evaluación.

En la Tabla 5 e ilustración 7 se observa que la herramienta SweSum se posicionó en el quinto lugar, mientras que en las evaluaciones pasadas tomando resúmenes de referencias humanas se posicionaba en primer lugar.

Tabla 5. ROUGE-C-1, ROUGE-C-2, ROUGE-C-3, ROUGE-C-L, ROUGE-C-SU4

Herramienta	C-1	C-2	C-3	C-L	C-SU4
Text-Compactor	**0.2565**	**0.2503**	**0.2458**	**0.2563**	**0.2453**
Summarizer	0.2295	0.2039	0.1785	0.2294	0.1754
Word 2007	**0.2257**	**0.2124**	**0.1997**	**0.2255**	**0.1971**
OTS	0.2257	0.2136	0.2034	0.2249	0.2018
SweSum	0.2108	0.1999	0.1892	0.2106	0.1873
Conspectus	0.2091	0.1844	0.1595	0.2076	0.1555
Word 2003	0.2035	0.1915	0.1797	0.2034	0.1769
Summarizing	0.1914	0.1780	0.1663	0.1871	0.1643

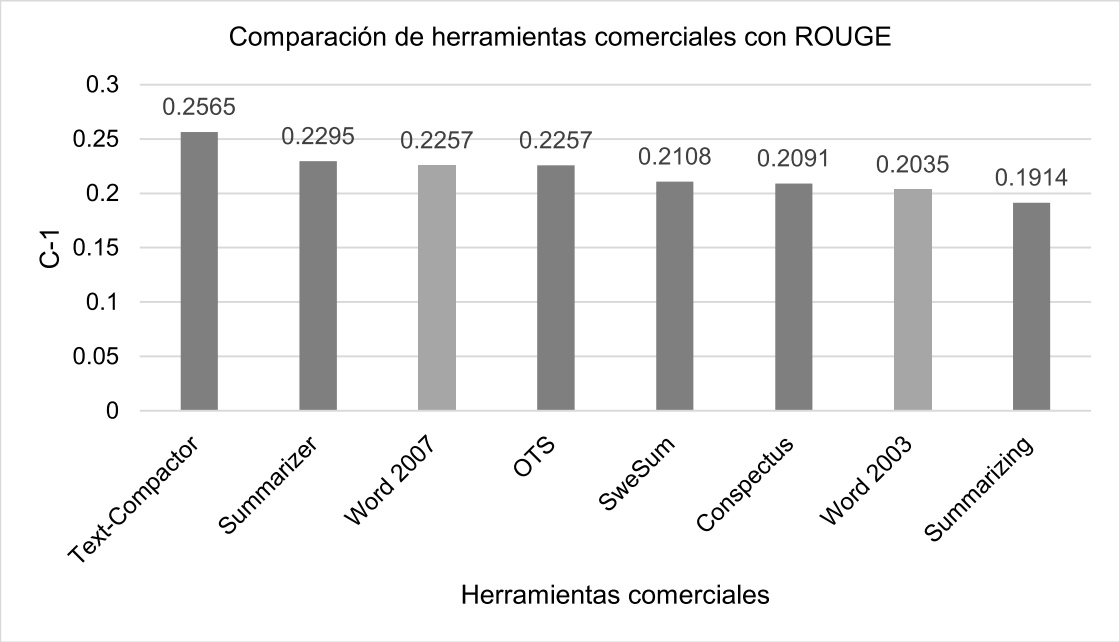


Ilustración 7. Evaluación de diferentes herramientas comerciales usando ROUGE-C sin referencia.

3.3.5. Jensen-Shannon

Este método evalúa un resumen generado de manera automática con el documento original. Incorpora la idea de la distancia entre dos distribuciones no puede ser muy diferente del promedio de las distancias de su distribución media. La evaluación está basada bajo la siguiente expresión.

$$D_{JS}(\mathrm{P}||\mathrm{Q}) = \frac{1}{2}\left[D(P||A) + D(Q||A)\right] \qquad (7)$$

Donde $\mathrm{A} = \frac{P+Q}{2}$ es la distribución media de P y Q. La divergencia D_{JS} es simétrica y siempre está definida.

En la evaluación realizada con D_{JS} se compara el resumen automático generado contra el texto original y de acuerdo con los datos obtenidos de la Tabla 6 Text Compactor fue

la mejor herramienta comercial para la GART, con dicho evaluador se tiene que la herramienta Word 2007 toma la tercera posición.

Tabla 6. Evaluación de diferentes herramientas comerciales usando Jensen Shannon.

Herramienta	Suavizada	No suavizada
Text-Compactor	**0.7363**	**0.71277**
Summarizer	0.72611	0.69972
Word Office 2007	**0.72378**	**0.69679**
OTS	0.72226	0.69674
SweSum	0.71815	0.69063
Word Office 2003	0.71165	0.6818
T-Conspectus	0.70925	0.68131
Summarizing	0.70185	0.67105

3.4 Automatically Assessing Machine Summary Content Without a Gold Standard (Louis & Nenkova, 2013)

3.4.1. Descripción

De acuerdo con este trabajo, los métodos actuales de evaluación de calidad del contenido de un resumen, se define por dos aspectos clave tales como, el contenido y la calidad lingüística. Un bien resumen debe tener la información más importante tanto en la entrada como en la estructura del contenido, presentándolo como un resumen bien escrito.

En este trabajo, se presentan métricas de evaluación para el contenido resumido que pueden o no tener participación humana, así como también se proponen métodos totalmente automáticos para la evaluación de contenido que se puede utilizar en

ausencia de resúmenes humanos. También exploramos métodos para promover la mejora y el desempeño de la evaluación cuando solo se dispone de un resumen modelo.

3.4.2. Evaluación

3.4.2.1. Evaluación de pirámide

La evaluación del método de pirámide, ha sido desarrollado para una evaluación confiable y de diagnóstico de la calidad de selección de contenido del resumen y se ha utilizado en varias evaluaciones a gran escala. Utilizando múltiples modelos humanos, a partir de los cuales los anotadores se identifican semánticamente como Unidades de contenido resumido definidas (SCU). A cada SCU se le asigna un peso igual al número de resúmenes de modelos humanos que se expresan en SCU, como se muestra en la ecuación 8.

$$py\,(S) = \frac{suma\ de\ pesos\ de\ SCU\ expresada\ en\ S}{suma\ de\ pesos\ de\ un\ resumen\ ideal\ con\ el\ mismo\ número\ de\ SCU\ que\ S} \qquad (8)$$

3.4.2.2. Capacidad de respuesta

Esta evaluación, se centra en la capacidad de respuesta de un resumen siendo una medida de calidad que combina contenidos selección y calidad lingüística. Mide hasta qué punto los resúmenes transmiten información apropiada contenido de forma estructurada. La capacidad de respuesta se evalúa mediante calificaciones directas otorgadas por los jueces. Por ejemplo, una escala del 1 (resumen pobre) al 5 (resumen muy bueno) es utilizado y estas evaluaciones se realizan sin referencias a ningún modelo de resumen.

3.4.2.3. Jensen-Shannon

La divergencia de D_{JS} incorpora que la idea acerca de la diferencia entre dos distribuciones no puede ser muy diferente del promedio de distancias de su distribución media. Se define formalmente como

$$D_{JS}\,(P\,||\,Q\,) = \frac{1}{2}\,[\,D\,(P||A) + (Q||A)] \qquad (9)$$

donde $\mathrm{A} = \mathrm{P} + \frac{Q}{2}$ es la distribución media de P y Q. En contraste con la divergencia D_{KL}, la distancia D_{JS} es simétrica y siempre definida. Se calcularon tanto versiones suavizadas como versiones sin suavizar de la divergencia como puntajes resumidos.

El enfoque de este análisis es comprender si la divergencia D_{JS} y ROUGE-2 están cometiendo errores al ordenar los sistemas o si sus errores son diferentes.

Este resultado también ayudó a comprender si la divergencia D_{JS} y ROUGE tienen fortalezas que se pueden combinar. Para esto, consideramos pares de sistemas y se calculó el mejor sistema en cada par de acuerdo con las puntuaciones de la pirámide. Posteriormente, para la divergencia D_{JS} y ROUGE, se registró la frecuencia con la que proporcionaron los juicios para los pares según lo indicado por la evaluación piramidal. Habiendo 1.653 pares de sistemas a nivel macro y los resultados se encuentran en la Tabla 7. Esta tabla muestra que una gran mayoría (80%) de la evaluación se predice correctamente por divergencia D_{JS} y ROUGE. Otro 6% de los pares son métricas que no proporcionan el juicio correcto. Por lo tanto, la divergencia de D_{JS} y ROUGE coinciden en una gran mayoría de los sistemas de pares.

Tabla 7. Divergencia correcta e incorrecta.

	D_{JS} correcto	D_{JS} incorrecto
ROUGE-2 correcto	1.319 (79,8%)	133 (8,1%)
ROUGE-2 incorrecto	96 (5,8%)	105 (6,3%)

Para este enfoque, se consideró que los resúmenes se generan según la distribución de palabras en la entrada. Entonces la probabilidad de una palabra en la entrada sería indicativa de la probabilidad de que se emitiera en un resumen. Además, los autores calcularon tanto la probabilidad unigrama de un resumen como su probabilidad bajo un modelo multinomial como se muestra en la Tabla 8.

Tabla 8. Resultado de correlación de las métricas de Simetrix.

Características	Pirámide	Sensibilidad
D_{JS} div	-0.880	-0.736
D_{JS} div suavizado	-0.884	-0.737
Porcentaje de palabras del tema de entrada	0.795	0.627
Resumen de div de D_{KL} entrada	-0.763	-0.694
Similitud de coseno con todas las palabras	0.712	0.647
Porcentaje de resumen = palabras del tema	0.699	0.602
Similitud de coseno con palabras temáticas	0.688	0.629
D_{KL} div entrada de resumen	0.222	-0.585
Probabilidad multinomial del resumen	-0.188	0.235
Probabilidad de unigramas del resumen	0.867	-0.101
Regresión	0.879	0.705
ROUGE-1	0.859	0.806
ROUGE-2	0.905	0.873

3.4.3. Resultados

El rendimiento de la evaluación completamente automática sigue siendo alto para su uso durante el desarrollo del sistema. Las mejores métricas, tanto la divergencia y regresión de D_{JS} resultan ser útiles con poca diferencia en el rendimiento entre ellos.

La selección local proporciona el mismo rendimiento que la selección global para las puntuaciones piramidales, aunque disminuye la calidad de la evaluación a nivel micro para la capacidad de respuesta.

3.5 Automatically Evaluating Content Selection in Summarization without Human Models (Louis & Nenkova, 2009)

3.5.1 Descripción

En este trabajo, se presentan métricas de evaluación para el contenido resumido que hacen uso de poca o ninguna participación humana y métodos de evaluación como pirámides y ROUGE, las cuales están basadas en comparar resúmenes con sus respectivas referencias humanas.

3.5.2 Evaluación

3.5.2.1 Divergencia Kullback-Leibler

La divergencia de D_{KL} entre dos distribuciones de probabilidad de distribuciones P, Q y w que hace referencia a cada palabra del documento y viene dado por

$$D_{KL}\,(P||Q) = \frac{1}{2}\sum_{w} P_w\, log_2 \frac{P_w}{Q_w} \qquad (10)$$

Se define como el número medio de bits desperdiciados al codificar muestras pertenecientes a P usando otra distribución Q, un aproximado de P.

La divergencia D_{KL} no es simétrica, ya que las divergencias de entrada de resumen se introducen como métricas. Además, la divergencia no está definida cuando $p^{P(W)} > 0$ pero $p^{Q(W)} = 0$. Para este problema, los autores realizaron un suavizado a la divergencia D_{KL}, como se muestra en la ecuación 11.

$$p\,(\,w\,) = \frac{C + \delta}{N + \delta * B} \qquad (11)$$

La variable C es el recuento de la palabra w y N es el número de tokens; $B = 1.5|V|$, donde V es el vocabulario de entrada y δ se estableció en un valor pequeño de 0,0005 para evitar cambios en la probabilidad de masa a eventos invisibles.

3.5.3 Probabilidad de resumen

Para este proceso, se consideró que los resúmenes se generan según la distribución de palabras en la entrada. Por ello, la probabilidad de que una palabra se encuentre en la entrada, será indicativo de que sea transmitido en un resumen y así mismo, pueda ser calculado utilizando diferentes métodos esperando que la probabilidad sea mayor para obtener resúmenes de mejor calidad.

De acuerdo con lo anterior se calculó tanto la probabilidad de resumen de unigrama como su probabilidad bajo un modelo multinomial.

3.5.3.1 Probabilidad de resumen de unigrama

La probabilidad de que una palabra se encuentre en la entrada, será indicativo de que la probabilidad se transmita en un resumen.

$$(Pinp^{w_1})^{n_1}\ (Pinp^{w_2})^{n_2}\ (Pinp^{w_r})^{n_r} \qquad (12)$$

Donde $N = n_1 + n_2 + n_r$ es el número total de palabras en el resumen.

3.5.3.2 Probabilidad de resumen multinomial

La probabilidad del contenido de un resumen se puede calcular utilizando diferentes métodos esperando que la probabilidad sea mayor para resúmenes de mejor calidad y por ello se calcula la probabilidad multinomial, como se muestra en la ecuación 13.

$$\frac{N}{n\,1\,!\,n2\,!\,nr\,!}\ (Pinp^{w_1})^{n_1}\ (Pinp^{w_2})^{n_2}\ (Pinp^{w_r})^{n_r} \qquad (13)$$

donde $N = n_1 + n_2 + n_r$ es el número total de palabras en el resumen.

3.5.4 Similitud coseno

La tercera métrica es la similitud del coseno entre $tf.idf$ representaciones vectoriales (con normalización max-tf) de los contenidos de entrada y resumen.

$$COS\ \theta = \frac{V\ inp\ .Vsumm}{||\ V\ inp\ ||||\ V\ summ\ ||} \qquad (14)$$

3.5.5 Regresión lineal

Además, se evaluó el desempeño de una métrica de regresión lineal que combina todas las métricas previamente mencionadas. Durante el desarrollo, el valor de la puntuación basada en la regresión para cada resumen se obtuvo utilizando un nuestro enfoque.

Para una combinación particular de entrada y resumen del sistema, el conjunto de entrenamiento constaba únicamente de ejemplos que no incluían la misma entrada ni el mismo sistema.

Por lo tanto, durante el entrenamiento, no se vieron ejemplos de la entrada de prueba ni del sistema.

3.5.6 Resultados

Se obtuvieron buenos resultados con el análisis de correlación mediante juicios humanos, mostrando que la información puede sustituir a los resúmenes modelo y los esfuerzos manuales en la evaluación resumida.
Las mejores correlaciones se obtuvieron mediante la divergencia D_{JS} con puntajes piramidales de -0.88 y -0,73 con capacidad de respuesta a nivel de sistema, así como también se muestra la correlación de Spearman entre puntajes manuales y automáticos. Por lo tanto, las mejores medidas se pueden utilizar para evaluar el rendimiento de la selección de contenido de los sistemas, en un nuevo dominio donde los resúmenes de referencias humanas no están disponibles.

Además, se tiene que la combinación de regresión lineal obtiene altas correlaciones con puntajes manuales, pero no conduce a mejores resultados que la divergencia D_{JS}.

A nivel micro evaluación, la combinación de características con regresión lineal obtiene mejores resultados en contraste con los resultados a nivel macro. Dicho resultado tiene implicaciones ya que no se puede predecir de forma fiable un buen contenido para una entrada en particular.

Sin embargo, debemos tener en cuenta que las características se basan únicamente en la distribución de términos de entrada, por lo tanto, es menos probable que se obtenga un buen contenido para todos los tipos de entrada. Por ejemplo, dado un conjunto de

documentos cada uno describiendo diferentes opiniones sobre un problema dado, probablemente tendrá menos repetición a nivel léxico y de unidad de contenido.

En general, los resultados a nivel micro evaluación sugieren que las medidas automáticas que examinaron no serán útiles para proporcionar información sobre la calidad del resumen para una entrada individual. Para el caso de promedios superiores de varios conjuntos de prueba, se considera que las evaluaciones automáticas dan resultados más fiables y útiles, correlacionados con características producidas por evaluaciones con referencias humanas.

De la Tabla 9 se obervan las correlaciones de Spearman a nivel micro evaluación con el mínimo y máximo de consultas, así como también, los valores de las correlaciones significativas junto con el número y porcentaje de correlaciones significativas.

Tabla 9. Correlaciones de Spearman a nivel micro (tarea centrada en consultas). Solo el mínimo y el máximo.

Características	maximo	minimo	número significativo (%)	maximo	minimo	número significativo (%)
Divergencia D_{JS}	-0.714	-0.271	35 (72.9)	-0.654	-0.262	35 (72.9)
Divergencia D_{JS} suavizada	-0.712	-0.269	35 (72.9)	-0.649	-0.279	33 (68.8)
Porcentaje de palabras de entrada	-0.736	-0.276	35 (72.9)	-0.628	-0.261	35 (72.9)
Resumen de divergencia D_{KL} de entrada	0.701	0.286	31 (64.6)	0.693	0.279	29 (60.4)
Superposición de coseno, todas las palabras	0.622	0.276	31 (64.6)	0.618	0.265	28 (58.3)
Porcentaje de resumen = palabras del tema	-0.628	-0.262	28 (58.3)	-0.577	-0.267	22 (45.8)
Superposición de coseno, palabras temáticas	0.597	0.265	30 (62.5)	0.689	0.277	26 (54.2)
Divergencia D_{KL} entrada-resumen	0.607	0.269	23 (47.9)	0.534	0.272	23 (47.9)
Probabilidad de resumen multinomial	0.434	0.268	8 (16.7)	0.459	0.272	10 (20.8)
Probabilidad de resumen de unigrama	0.292	0.261	2 (4.2)	0.466	0.287	2 (4.2)
Regresión	0.736	0.281	37 (77.1)	0.642	0.262	32 (66.7)
ROUGE-1	0.833	0.264	47 (97.9)	0.754	0.266	46 (95.8)
ROUGE-2	0.875	0.316	48 (100)	0.742	0.299	44 (91.7)

CAPÍTULO 4

Metodología propuesta

En este capítulo, se describe la metodología propuesta que se llevó a cabo para realizar la evaluación de resúmenes guiados, así como la comparación entre diferentes métodos de evaluación. En general, la metodología está conformada por las siguientes estapas mostradas en la ilustración 8:

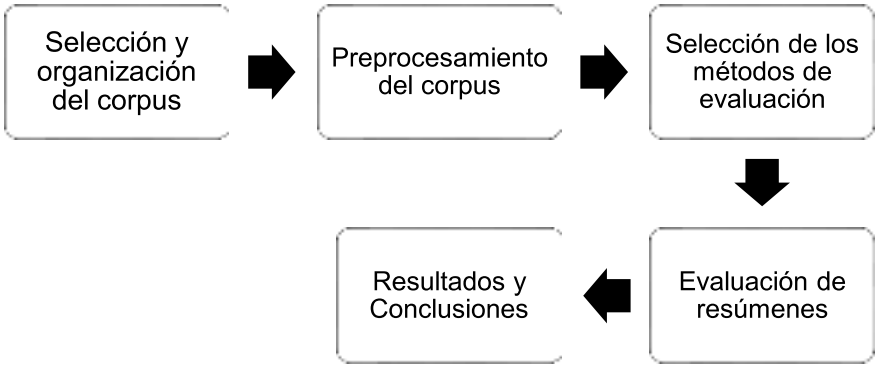


Ilustración 8. Metodología propuesta.

A continuación, se describe detalladamente cada una de las etapas previamente mostradas de la metodología propuesta.

4.1 Selección y organización del corpus

Antes de comenzar con la evaluación de resúmenes, es importante tener en cuenta, conocer y describir a detalle el corpus de documentos que se utilizó.

4.2 Preprocesamiento del corpus

El preprocesamiento del corpus involucra una serie de tareas que ayudan a proporcionar información relevante en relación con el contenido de los documentos del corpus. Las tareas que se manejaron en esta etapa fueron las siguientes:

- Limpieza de documentos: El contenido los documentos fuente y los resúmenes fue depurado mediante una eliminación de etiquetas HTML. A continuación, se muestra un ejemplo de eliminación de tales etiquetas a partir de un documento fuente.

```
<DOC>
<DOCNO> APW19990503.0128 </DOCNO>
<DATE_TIME> 1999-05-03 15:55:11 </DATE_TIME>
<BODY>
<CATEGORY> washington </CATEGORY>
<HEADLINE> Congress Looking at Youth Violence </HEADLINE>
<TEXT>
<P>
WASHINGTON (AP) -- Pressured to help stop kids from killing, Congress is opening hearings on the causes of a ``crisis among our young'' amid a thorny political question of what government should
do to prevent massacres like the one in Littleton, Colo.
</P>
<P>
``The tragedy at Columbine High and the ongoing carnage on our inner city streets presents us with a complicated cultural moment and an important opportunity to thoroughly examine the root causes
of a crisis among our young,'' House Judiciary Committee Chairman Henry Hyde told reporters on Monday.
</P>
<P>
Hyde's committee will hold a wide-ranging hearing beginning May 13 focusing on urban and suburban violence by kids.
</P>
<P>
The witness list was still in its infancy late Monday, but several people close to the committee said it could include individuals involved in the Columbine tragedy, survivors of other school shootings, entertainment industry executives and interest groups on both sides of the gun control debate.
</P>
<P>
Hyde, who has supported gun control legislation, is committed to keeping the hearing from spiraling into a partisan shoot-out over gun policy, according to two officials close to the panel who asked not to be named. They acknowledged, however, that the discussion would have to touch on such issues as new firearms technology and whether access to guns should be further restricted.
</P>
<P>
Though Hyde's support of gun control legislation makes him no friend of the National Rifle Association, these officials would not rule out testimony by the group's president, actor Charlton Heston, or any other of the group's officials.
</P> <P>
In the Senate, the Commerce Committee on Tuesday opens a hearing on the marketing of violent material to children.
</P> <P>
The Senate hearing was scheduled before the April 20 Columbine massacre.
</P>
</TEXT>
</BODY>
</DOC>
```

Ilustración 9. Ejemplo de documento con etiquetas HTML.

Congress Looking at Youth Violence

WASHINGTON (AP) Pressured to help stop kids from killing,Congress is opening hearings on the causes of a crisis among our young amid a thorny political question of what government should do to prevent massacres like the one in Littleton, Colo.
The tragedy at Columbine High and the ongoing carnage on our inner city streets presents us with a complicated cultural moment and an important opportunity to thoroughly examine the root causes of a crisis among our young, House Judiciary Committee Chairman Henry Hyde told reporters on Monday.
Hyde's committee will hold a wide-ranging hearing beginning May 13 focusing on urban and suburban violence by kids.
The witness list was still in its infancy late Monday, but several people close to the committee said it could include individuals involved in the Columbine tragedy, survivors of other school shootings, entertainment industry executives and interest
groups on both sides of the gun control debate.
Hyde, who has supported gun control legislation, is committed to keeping the hearing from spiraling into a partisan shoot-out over gun policy, according to two officials close to the panel who asked not to be named. They acknowledged, however, that the discussion would have to touch on such issues as new firearms technology and
whether access to guns should be further restricted.
Though Hyde's support of gun control legislation makes him no friend of the National Rifle Association, these officials would not rule out testimony by the group's president, actor Charlton Heston, or any other of the group's officials.
In the Senate, the Commerce Committee on Tuesday opens a hearing on the marketing of violent material to children.
The Senate hearing was scheduled before the April 20 Columbine massacre.

Ilustración 10. Ejemplo de eliminación de etiquetas HTML.

- Clasificación de documentos: Primeramente, los documentos fuente son agrupados por tipo (A o B). Posteriormente, los resúmenes a evaluar son clasificados de acuerdo con el sistema generador de resúmenes por el que fue generado, y además por tipo de resumen (A representa el resumen inicial y B representa el resumen actualizado).

- Segmentación de textos en oraciones: Se aplica la segmentación textos en oraciones para cada resumen y documento fuente. En el siguiente ejemplo, se

- muestra un ejemplo de segmentación de texto en oraciones en un documento fuente.

Congress Looking at Youth Violence WASHINGTON (AP) Pressured to help stop kids from killing,Congress is opening hearings on the causes of a crisis among our young amid a thorny political question of what government should do to prevent massacres like the one in Littleton, Colo.
The tragedy at Columbine High and the ongoing carnage on our inner city streets presents us with a complicated cultural moment and an important opportunity to thoroughly examine the root causes of a crisis among our young, House Judiciary Committee Chairman Henry Hyde told reporters on Monday. Hyde's committee will hold a wide-ranging hearing beginning May 13 focusing on urban and suburban violence by kids. The witness list was still in its infancy late Monday, but several people close to the committee said it could include individuals involved in the Columbine tragedy, survivors of other school shootings, entertainment industry executives and interest groups on both sides of the gun control debate. Hyde, who has supported gun control legislation, is committed to keeping the hearing from spiraling into a partisan shoot-out over gun policy, according to two officials close to the panel who asked not to be named. They acknowledged, however, that the discussion would have to touch on such issues as new firearms technology and whether access to guns should be further restricted. Though Hyde's support of gun control legislation makes him no friend of the National Rifle Association, these officials would not rule out testimony by the group's president, actor Charlton Heston, or any other of the group's officials. In the Senate, the Commerce Committee on Tuesday opens a hearing on the marketing of violent material to children. The Senate hearing was scheduled before the April 20 Columbine massacre.

Ilustración 11. Ejemplo de documento sin segmentación de oraciones.

- Normalización: Cada documento fuente y resumen es normalizado, eliminando signos de puntuación, y cada palabra es convertida en minúsculas o mayúsculas.

- Stemming: Se realiza la reducción de diferentes variantes de una palabra a su forma raíz mediante el *stemming* de Porter. Por ejemplo, si un documento contiene los términos organizes, organized, organization y organizing, el stemming de Porter los reduce por su término raíz en común: organize.

```
Congress Looking at Youth Violence WASHINGTON (AP) Pressured to help
stop kids from killing, Congress is opening hearings on the causes of
a crisis among our young amid a thorny political question of what
government should do to prevent massacres like the one in Littleton,
Colo.
The tragedy at Columbine High and the ongoing carnage on our inner
city streets presents us with a complicated cultural moment and an
important opportunity to thoroughly examine the root causes of a
crisis among our young, House Judiciary Committee Chairman Henry Hyde
told reporters on Monday.
Hyde's committee will hold a wide-ranging hearing beginning May 13
focusing on urban and suburban violence by kids.
The witness list was still in its infancy late Monday, but several
people close to the committee said it could include individuals
involved in the Columbine tragedy, survivors of other school
shootings, entertainment industry executives and interest groups on
both sides of the gun control debate.
Hyde, who has supported gun control legislation, is committed to
keeping the hearing from spiraling into a partisan shoot-out over gun
policy, according to two officials close to the panel who asked not
to be named.
They acknowledged, however, that the discussion would have to touch
on such issues as new firearms technology and whether access to guns
should be further restricted.
Though Hyde's support of gun control legislation makes him no friend
of the National Rifle Association, these officials would not rule out
testimony by the group's president, actor Charlton Heston, or any
other of the group's officials.
In the Senate, the Commerce Committee on Tuesday opens a hearing on
the marketing of violent material to children.
The Senate hearing was scheduled before the April 20 Columbine
massacre.
```

Ilustración 12. Ejemplo de segmentación de oraciones.

4.3 Selección de los métodos de evaluación

Posterior al preprocesamiento del corpus, es importante seleccionar los métodos de evaluación. Los métodos de evaluación permitirán asignar puntajes de calidad a cada

resumen. Básicamente, los métodos de evaluación que se utilizaron son parte de los sistemas ROUGE-C y SIMetrix y están enlistados a continuación:

- ROUGE-C-1: evaluación a base de palabras (unigramas) encontradas entre el resumen y sus documentos fuente.
- ROUGE-C-2: evaluación a base de bigramas encontrados entre el resumen y sus documentos fuente.
- ROUGE-C-3: evaluación a base de trigramas entre el resumen y sus documentos fuente.
- ROUGE-C-L: considera las subsecuencias comunes más largas (LCSs) de palabras entre el resumen y sus documentos fuente.
- ROUGE-C-SU4: utiliza bigramas de palabras con saltos de cuatro términos para calcular la semejanza entre el resumen a evaluar y sus documentos fuente.
- Divergencia Jensen-Shannon no suavizada de SIMetrix: emplea la divergencia Jensen-Shannon para medir cuán diferente es el resumen con respecto a su documento fuente. En este método de evaluación, no se considera el manejo de las palabras del resumen que no se encuentran en los documentos fuente.
- Divergencia Jensen-Shannon suavizada de SIMetrix: a diferencia del método de evaluación anterior, la divergencia Jensen-Shannon suavizada si considera las palabras del resumen que no están en sus documentos fuente.

4.4 Evaluación de resúmenes

Para esta etapa, se describe cómo se está comparando cada resumen con su respectivo documento fuente. Además, se muestra el número total de resúmenes evaluados.

4.5 Resultados y conclusiones

Se menciona el proceso por el cual se mide la proximidad de cada evaluador utilizado del paso 3 con respecto a una serie de juicios humanos, utilizando los índices de correlación de Pearson, Spearman y Kendall. Posteriormente, se muestran los resultados obtenidos en términos de correlación, así como las conclusiones.

CAPÍTULO 5

Experimentos y resultados

En este capítulo se describe el corpus TAC 2010 y los experimentos realizados durante el proceso de evaluación.

5.1 Descripción del corpus

TAC 2010 es un corpus de documentos destinado al desarrollo e implementación de generadores y evaluadores de resúmenes. Básicamente, está constituido por 920 documentos fuente en inglés divididos en 46 colecciones, donde cada colección cuenta con documentos iniciales (Tipo A) y de actualización (Tipo B). Por lo tanto, todos los documentos de tipo A preceden cronológicamente a los documentos de tipo B. Dichos documentos fueron recolectados de diferentes fuentes periodísticas y además abordan las siguientes cinco categorías:

- Accidentes y desastres naturales
- Ataques
- Salud y seguridad
- Recursos en peligro
- Juicios e investigaciones

Posterior a la definición de documentos fuente, se requirió que diversos generadores de resúmenes produjeran dos resúmenes guiados de 100 palabras para cada colección. Un resumen corresponde a la versión sintetizada del documento fuente inicial (tipo A) y el otro corresponde a la versión sintetizada del documento fuente actualizado (tipo B). Por lo tanto, TAC 2010 cuenta con 3956 resúmenes distribuidos en 43 sistemas generadores de resúmenes. Además, cuenta con 368 resúmenes manuales generados por ocho resumidores humanos, obteniendo un total de 4324 resúmenes. Finalmente, es importante mencionar que se seleccionaron tanto los documentos fuente como sus respectivos resúmenes para el proceso de evaluación.

Dentro de TAC 2010, la generación de resúmenes guiados tiene como objetivo alentar a los sistemas generadores de resúmenes realizar un análisis lingüístico (semántico) de los documentos de origen, en lugar de depender únicamente de las frecuencias de palabras del documento para seleccionar conceptos importantes. Por lo tanto, cada

resumen generado debe seguir una plantilla, la cual responde una serie de preguntas tales como: ¿Qué pasó? ¿Cuándo ocurrió? ¿Dónde ocurrió? ¿Por qué?, ¿Quién o quiénes estuvieron involucrados?, fecha, hora, daños ocurridos y el esfuerzo realizado por cada persona para sobrellevar el desastre o accidente (para los resúmenes de accidentes y desastres naturales). Todo este proceso se realiza para cada colección de documentos a resumir.

Por otro lado, TAC 2010 abordó la evaluación automática de resúmenes por pares (AESOP), que consiste en generar métodos de evaluación capaces de predecir la calidad de un resumen automático o manual. Para poder realizar esta tarea, fue necesario culminar con la generación de resúmenes guiados.

La principal meta de AESOP consiste en producir dos conjuntos de evaluaciones:

1. Evaluación de todos los resúmenes por pares (All peers): Consiste en asignar un puntaje de evaluación para cada resumen automático o manual.

2. Evaluación de resúmenes automáticos (No models): Únicamente, se encarga de evaluar resúmenes automáticos, excluyendo los resúmenes manuales.

5.2 Experimentación de resúmenes de tipo A

Una forma de conocer el alcance de un evaluador es mediante la comparación que emite el evaluador de juicios humanos, es decir, juicios humanos contra juicios de la computadora. Para realizar esta medición, se ocupan los índices de correlación de Pearson, Spearman y Kendall. Cada uno de estos índices cuentan con diferentes criterios para medir la proximidad o nivel de predicción del evaluador automático hacia una evaluación manual. Por ejemplo, la correlación de Pearson estima cuan linealmente se encuentran relacionadas dos variables, entiéndase como variables el evaluador automático y el evaluador manual.

Posteriormente, se tiene el índice de correlación de Spearman. En este caso, la correlación de Spearman mide el grado de igualdad de la evaluación automática respecto a los juicios humanos.

Por último, tenemos el índice de correlación Kendall, el cual establece una clasificación general entre dos variables. A partir de esta clasificación, se mide el número de concordancias y discordancias crecientes/decrecientes entre ambos, generando un valor de correlación.

De cada método de evaluación, se han generado una serie de puntajes que indican la similitud de los resúmenes evaluados con respecto a sus documentos fuente. Posteriormente, los puntajes obtenidos son comparados con puntajes de evaluación manual a través de los índices de correlación previamente mencionados. Los resultados de correlación varían de -1 a 1, donde -1 indica una relación inversa entre la evaluación automática y manual. Es decir, el evaluador automático asigna puntajes altos a malos resúmenes y puntajes bajos a buenos resúmenes.

Por otro lado, si la correlación se aproxima a 1, indica que el evaluador asigna altos puntajes a buenos resúmenes y bajos puntajes a malos resúmenes.

En la Tabla 10 se muestra que en ROUGE-C-1 existe una concordancia positiva bajo los tres índices de correlación. Por lo tanto, ROUGE-C-1 toma un comportamiento de asignar altos puntajes a buenos resúmenes y bajos puntajes a malos resúmenes

Los resúmenes que evalúa ROUGE-C-3 presentaron discordancia, ya que tuvieron mayor puntaje en resúmenes malos, dado que, cuando existe una correlación negativa quiere decir que existe una relación inversa.

Tabla 10. Resultados de correlación de los métodos utilizados en la colección TAC 2010, resúmenes de actualización guiados.

ALLPEERS				
Tipo A	P	S	K	PROMEDIO
ROUGE-C-1	0,2644	0,2225	0,1715	0,2195
ROUGE-C-2	-0,4422	-0,0239	0,0118	-0,1514
ROUGE-C-3	-0,5702	-0,0809	-0,0322	-0,2278
ROUGE-C-L	0,1276	0,1687	0,1322	0,1429
ROUGE-SU4	-0,3593	0,0028	0,0275	-0,1096
D_{JS} suavizada	0,1760	0,3502	0,2975	0,2746
D_{JS} no suavizada	**0,2516**	**0,3584**	**0,3006**	**0,3035**

Por otro lado, el evaluador de divergencia no suavizada es considerado como el mejor para la evaluación de resúmenes de tipo A, ya que obtuvo una correlación mucho mayor que los demás, considerando que en este caso existe una concordancia y puntaje mayor a buenos resúmenes.

En la Tabla 11 se muestran los resultados de correlación de los evaluadores en la colección TAC 2010, usando solamente los resúmenes automáticos. Para el caso de no modelos (NO MODELS), únicamente se toman en cuenta los juicios automáticos. Dado que, en el evaluador ROUGE-C-1, ROUGE-C-2, ROUGE-C-3, ROUGE-C-L y ROUGE-C-SU4 existen correlaciones positivas, obteniendo un mejor resultado mediante la evaluación de resúmenes automáticos.

Cabe resaltar que, en el apartado de no modelos, existe un mejor resultado en cuanto a la evaluación de todos los resúmenes, ya que muestra concordancia y puntajes altos de buenos resúmenes. Esto quiere decir, que la evaluación de resúmenes automáticos sin referencias humanas es la mejor hasta el momento.

Tabla 11. Resultados de correlación de los métodos utilizados en la colección TAC 2010 usando resúmenes automáticos (NO MODELS) de tipo A.

	NO MODELS			
Tipo A	P	S	K	PROMEDIO
ROUGE-C-1	0.7273	0.4755	0.3548	0.5192
ROUGE-C-2	0.7148	0.5331	0.3735	0.5405
ROUGE-C-3	0.6779	0.4950	0.3357	0.5029
ROUGE-C-L	0.5465	0.4938	0.3681	0.4695
ROUGE-SU4	0.7228	0.5574	0.3913	0.5572
D_{JS} suavizada	**0.8987**	**0.8556**	**0.6662**	**0.8068**
D_{JS} no suavizada	0.8676	0.8569	0.6618	0.7954

Tomando en cuenta que el mejor evaluador en los resúmenes automáticos fue D_{JS} suavizada, por esta razón se dice que hubo concordancia entre los evaluadores y los puntajes obtenidos, esto quiere decir que se encontraron más resúmenes buenos que malos.

5.3 Experimentación de resúmenes de tipo B

En esta sección se realizó el mismo procedimiento que el anterior, se evaluaron a los métodos bajo los mismos índices de correlación (Pearson, Spearman y Kendall), pero tomando como referencia el conjunto de datos de tipo B.

Para la Tabla 12, se observa que el evaluador ROUGE-C-2, presentó bajas correlaciones hacia los juicios humanos, es decir, existe una correlación negativa teniendo como resultado una relación inversa.

El evaluador de divergencia suavizada fue el único que presentó concordancia en todos los índices de correlación, puesto que sus puntajes de buenos resúmenes fueron altos, tomando el segundo lugar en la tabla de resultados. Por otro lado, el evaluador de

divergencia no suavizada fue el evaluador con las más altas correlaciones. Es decir, mostró cierta concordancia en sus evaluaciones con respecto a los juicios humanos.

Tabla 12. Resultados de correlación de los métodos utilizados en la colección TAC 2010, resúmenes de actualización guiados tipo B.

ALLPEERS				
Tipo B	P	S	K	PROMEDIO
ROUGE-C-1	0,1735	0,0067	-0,0125	0,0559
ROUGE-C-2	-0,4003	-0,0956	-0,0676	-0,1878
ROUGE-C-3	-0,5477	-0,0393	-0,0157	-0,2009
ROUGE-C-L	0,1218	-0,0450	-0,0408	0,0119
ROUGE-SU4	-0,3272	-0,0355	-0,0220	-0,1282
D_{JS} suavizada	0,0359	0,2578	0,2547	0,1828
D_{JS} no suavizada	**0,1417**	**0,2771**	**0,2641**	**0,2277**

En la Tabla 13 se puede visualizar que hubo correlaciones altas en todos los evaluadores respecto a los juicios humanos, esto quiere decir que los resúmenes automáticos son los más factibles para ser evaluador mediante métodos automáticos.

Tabla 13. Resultados de los índices de correlación Pearson, Spearman y Kendall de los métodos utilizados en la colección TAC 2010, resúmenes automáticos tipo B.

	NO MODELS			
Tipo B	P	S	K	PROMEDIO
ROUGE-C-1	0.6594	0.2833	0.1844	0.3757
ROUGE-C-2	0.6721	0.3926	0.2555	0.4401
ROUGE-C-3	0.6652	0.4877	0.3333	0.4954
ROUGE-C-L	0.6596	0.2901	0.1911	0.3803
ROUGE-SU4	0.6846	0.4940	0.3244	0.5010
D_{JS} suavizada	**0.8643**	**0.8268**	**0.6422**	**0.7778**
D_{JS} no suavizada	0.8041	0.8259	0.6444	0.7582

Cabe mencionar y resaltar que el evaluador de divergencia D_{JS} suavizada es el evaluador con mayor correlación, ya que fue el mejor en la evaluación de resúmenes automáticos. No obstante, se puede visualizar que la mayoría de los evaluadores utilizados manejaron puntajes altos y una mayor concordancia en cuanto a la evaluación de resúmenes automáticos sin referencias humanas. Por lo que, en el mejor de los casos, se sugiere utilizar este tipo de métodos al considerar solamente resúmenes automáticos.

CAPÍTULO 6

Conclusiones

- En este capítulo se describen las consecuencias y aspectos importantes del trabajo realizado.
- Se describe si se lograron cumplir a los objetivos específicos y el objetivo general.
- Se brindan recomendaciones.

6.1 Conclusiones

En esta tesis se realizó la evaluación de los resúmenes de actualización guiados utilizando métodos sin referencias humanas e índices de correlación (Pearson, Spearman y Kendall).

En la tesis se cumplieron los siguientes objetivos:

- Se determinó la calidad de los resúmenes de actualización guiados.
- Se utilizaron diferentes resúmenes de la colección TAC 2010.
- Se realizó una comparación entre métodos de evaluación con los índices de Pearson, Spearman y Kendall.
- Se obtuvo una evaluación de resúmenes mediante herramientas de que no dependen de referencias humanas.
- Se realizó una comparación de resultados de los resúmenes evaluados a partir de métodos de evaluación propuestos en el estado del arte.

6.2 Trabajo futuro

Como trabajo futuro se probarán colecciones en otros idiomas como español, portugués y ruso. La metodología propuesta se aplicará para la evaluación de otras tareas de PLN.

El trabajo futuro con base en esta tesis es implementar un método para evaluación de resúmenes automáticos guiados, utilizando los documentos que están disponibles en el corpus TAC 2010.

Realizar pruebas del corpus TAC 2010 con otros métodos del estado del arte para la evaluación de resúmenes automáticos guiados.

Referencias

Alexander Gelbukh, & Sidorov, G. (2006). Comparación de los coeficientes de las leyes de Zipf y Heaps en diferentes idiomas. In Procesamiento automático del español con enfoque en recursos léxicos grandes.

Amigó, E., Gonzalo, J., Peñas, A., & Verdejo, F. (2005). QARLA. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05, 280–289. https://doi.org/10.3115/1219840.1219875

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus Aplicaciones. Un Cuento de Una M´aquina Parlante, I.

Ledeneva, Y. (2008). Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. In Micron.Gelbukh.Com. http://micron.gelbukh.com/thesis/Yulia Ledeneva - PhD.

Ledeneva, Y., & García-Hernández, R. (2017). Automatic Generation of Text Summaries: Challenges, proposals and experiments.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches out (WAS 2004), 1, 25–26. papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85

Louis, A., & Nenkova, A. (2013). Automatically Assessing Machine Summary Content Without a Gold Standard. Computational Linguistics, 39(2), 267–300. https://doi.org/10.1162/COLI_a_00123

Louis, A., & Nenkova, A. (2009). Automatically evaluating content selection in summarization without human models. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP ’09, 1, 306. https://doi.org/10.3115/1699510.1699550.

Matias Mendoza, G. A. (2013). Generación automática de resúmenes usando algoritmos genéticos.pdf. Unidad Académica Profesional Tianguistenco.

Rojas-Simón, J. (2019). Evaluation Of Text Summaries Based On Linear Optimization Of Automatic Metrics. Unidad Académica Profesional Tianguistenco.

Rojas-Simón, J., Ledeneva, Y., & García-Hernández, R. A. (2021). Evaluation of text summaries without human references based on the linear optimization of content metrics using a genetic algorithm. Expert Systems with Applications, 167, 113827. https://doi.org/10.1016/j.eswa.2020.113827.

Ruvalcaba, C. y Alfredo Cerda Muñoz, F. (2004). Taller De Lectura Y Redacción. 44950, Mexico: Ildelisa Arias Arellano.

Sidorov, G. (2013). Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction. International Journal of Computational Linguistics and Applications, 4(2), 169–188. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.644.907&rep=rep1&type=pdf%0Ahttps://www.ijcla.org/2013-2/.

Steinberger, J., & Ježek, K. (2009). Evaluation measures for text summarization. Computing and Informatics, 28(2), 251–275.

Suanmali, L., Salim, N., & Binwahlan, M. S. (2011). Genetic algorithm based sentence extraction for text summarization. International Journal of Innovative Computing, 1(1), 22. http://kp.fsksm.utm.my/ijic/index.php/ijic/article/view/6.

Vilchis Sepúlveda, S. A., & Ledeneva, Y. (2019). Evaluación de resúmenes automáticos con y sin resúmenes de referencia para el idioma inglés. Research in Computing Science, 148(7), 241–252. https://doi.org/10.13053/rcs-148-7-18.

Anexos

1. Lista de stopwords en inglés

A, ABLE, ABOUT, ABOVE, ACCORDING, ACCORDINGLY, ACROSS, ACTUALLY, AFTER, AFTERWARDS, AGAIN, AGAINST, AIN'T, ALL, ALLOW, ALLOWS, ALMOST, ALONE, ALONG, ALREADY, ALSO, ALTHOUGH, ALWAYS, AM, AMONG, AMONGST, AN, AND, ANOTHER, ANY, ANYBODY, ANYHOW, ANYONE, ANYTHING, ANYWAY, ANYWAYS, ANYWHERE, APART, APPEAR, APPRECIATE, APPROPRIATE, ARE, AREN'T, AROUND, AS, ASIDE, ASK, ASKING, ASSOCIATED, AT, AVAILABLE, AWAY, AWFULLY, B, BE, BECAME, BECAUSE, BECOME, BECOMES, BECOMING, BEEN, BEFORE, BEFOREHAND, BEHIND, BEING, BELIEVE, BELOW , BESIDE, BESIDES, BEST, BETTER, BETWEEN, BEYOND, BOTHBRIEF, BUT, BY,C, C'MON, C'S, CAME, CAN, CAN'T, CANNOT, CANT,, CAUSE, CAUSES, CERTAIN, CERTAINLY, CHANGES, CLEARLY, CO, COM, COME, COMES, CONCERNING,CONSEQUENTLY, CONSIDER, CONSIDERING, CONTAIN, CONTAINING, CONTAINS,CORRESPONDING, COULD, COULDN'T, COURSE, CURRENTLY, D, DEFINITELY, DESCRIBED,DESPITE, DID, DIDN'T, DIFFERENT, DO, DOES, DOESN'T, DOING, DON'T, DONE, DOWN,DOWNWARDS, DURING, E, EACH ,EDU ,EG ,EIGHT ,EITHER, ELSE, ELSEWHERE, ENOUGH,ENTIRELY, ESPECIALLY, ET, ETC , EVEN, EVER ,EVERY ,EVERYBODY ,EVERYONE, EVERYTHING,EVERYWHERE, EX, ,EXACTLY, EXAMPLE, EXCEPT, F ,FAR ,FEW ,FIFTH ,FIRST , FIVE ,FOLLOWED,FOLLOWING ,FOLLOWS ,FOR, FORMER, FORMERLY, FORTH, FOUR, FROM, FURTHER,FURTHERMORE ,G, GET ,GETS, GETTING, GIVEN, GIVES, GO, GOES , GOING, GONE, GOT,GOTTEN, GREETINGS ,H, HAD , HADN'T, HAPPENS, HARDLY, HAS, HASN'T, HAVE, HAVEN'T,HAVING ,HE ,HE'S, HELLO, HELP, HENCE ,HER, HERE, HERE'S, HEREAFTER ,HEREBY ,HEREIN,HEREUPON , HERS, HERSELF ,HI ,HIM, HIMSELF, HIS, HITHER, HOPEFULLY, HOW, HOWBEIT,HOWEVER, I, I'D, I'LL, I'M, I'VE, IE ,IF, IGNORED, IMMEDIATE, IN, INASMUCH, INC, INC.,

INDEED,INDICATE, INDICATED, INDICATES, INNER, INSOFAR, INSTEAD, INTO, INWARD, IS, ISN'T, IT, IT'D, IT'LL, IT'S, ITS, ITSELF, J, JUST, K, KEEP, KEEPS, KEPT, KNOW, KNOWS, KNOWN, L, LAST,, LATELY ,LATER,LATTER, LATTERLY, LEAST, LESS, LEST, LET, LET'S, LIKE, LIKED, LIKELY, LITTLE, LOOK, LOOKING, LOOKS,LTD, M, MAINLY, MANY, MAY, MAYBE, ME, MEAN, MEANWHILE, MERELY, MIGHT, MORE,MOREOVER, MOST, MOSTLY, MUCH, MUST, MY, MYSELF, N, NAME, NAMELY, ND, NEAR, NEARLY,NECESSARY, NEED, NEEDS, NEITHER, NEVER, NEVERTHELESS, NEW, NEXT, NINE, NO, NOBODY,NON, NONE, NOONE, NOR, NORMALLY, NOT, NOTHING, NOVEL, NOW, NOWHERE, O,OBVIOUSLY, OF, OFF, OFTEN, OH, OK, OKAY, OLD, ON, ONCE, ONE, ONES, ONLY, ONTO, OR,OTHER, OTHERS, OTHERWISE, OUGHT, OUR, OURS, OURSELVES, OUT, OUTSIDE, OVER, OVERALL,OWN, P, PARTICULAR, PARTICULARLY, PER, PERHAPS, PLACED, PLEASE, PLUS, POSSIBLE,PRESUMABLY, PROBABLY, PROVIDES, Q, QUE, QUITE, QV, R, RATHER, RD, RE, REALLY,REASONABLY, REGARDING, REGARDLESS, REGARDS, RELATIVELY, RESPECTIVELY, RIGHT, S, SAID,SAME, SAW, SAY, SAYING, SAYS, SECOND, SECONDLY, SEE, SEEING, SEEM, SEEMED, SEEMING,SEEMS, SEEN, SELF,SELVES, SENSIBLE, SENT, SERIOUS, SERIOUSLY, SEVEN, SEVERAL, SHALL, SHE,SHOULD, SHOULDN'T, SINCE, SIX, SO, SOME, SOMEBODY,, SOMEHOW, SOMEONE, SOMETHING,SOMETIME, SOMETIMES, SOMEWHAT, SOMEWHERE, SOON, SORRY,, SPECIFIED, SPECIFY,SPECIFYING, STILL, SUB, SUCH, SUP, SURE, T, T'S, TAKE, TAKEN,, TELL, TENDS, TH, THAN, THANK,THANKS, THANX, THAT, THAT'S, THAT’S, THE, THEIR, THEIRS, THEM, THEMSELVES, THEN, THENCE,THERE, THERE'S, THEREAFTER, THEREBY, THEREFORE, THEREIN, THERES, THEREUPON, THESE, THEY,THEY'D, THEY'LL, THEY'RE, THEY'VE, THINK,, THIRD, THIS, THOROUGH, THOROUGHLY, THOSE,THOUGH, THREE, THROUGH, THROUGHOUT, THRU, THUS, TO, TOGETHER, TOO, TOOK, TOWARD,TOWARDS, TRIED, TRIES, TRULY, TRY, TRYING, TWICE, TWO, U, UN, UNDER, UNFORTUNATELY,UNLESS, UNLIKELY, UNTIL, UNTO, UP, UPON, US, USE, USED, USEFUL, USES, USING, USUALLY, UUCP,V, VALUE, VARIOUS, VERY, VIA, VIZ, VS, W, WANT, WANTS, WAS, WASN'T,

WAY, WE, WE'D, WE'LL,WE'RE, WE'VE, WELCOME,, WELL, WENT, WERE, WEREN'T, WHAT, WHAT'S, WHATEVER, WHEN,WHENCE, WHENEVER, WHERE, WHERE'S, WHEREAFTER, WHEREAS, WHEREBY, WHEREIN,WHEREUPON, WHEREVER, WHETHER, WHICH, WHILE, WHITHER, WHO, WHO'S, WHOEVER, WHOLE,WHOM, WHOSE, WHY, WILL, WILLING, WISH, WITH, WITHIN, WITHOUT, WON'T, WONDER, WOULD,WOULDN'T, X, Y, YES, YET, YOU, YOU'D, YOU'LL, YOU'RE, YOU'VE, YOUR, YOURS, YOURSELF,YOURSELVES, Z, ZERO

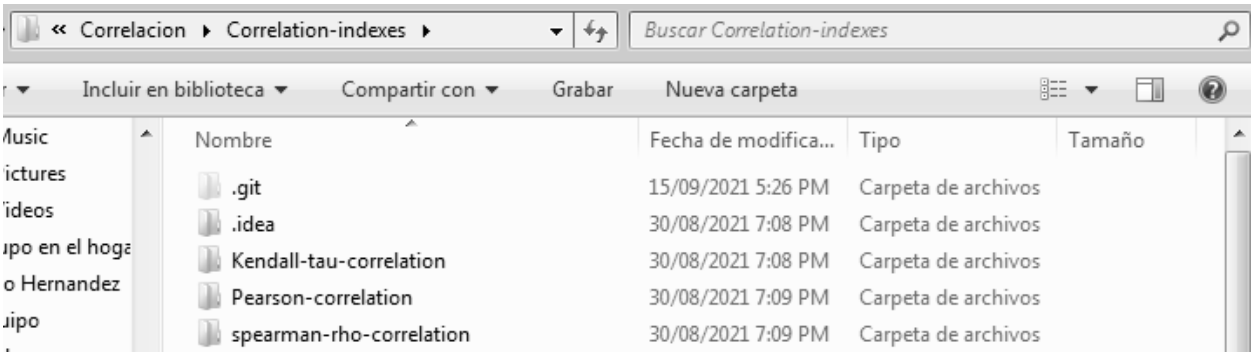
2. Resultados de la evaluación

A continuación se da a conocer como se fue trabajando con los evaluadores utilizados.

Para el caso de evaluador de correlación se tiene:

Carpeta principal:

En los primeros 3 directorios corresponde a los nombres de los índices de correlación (Pearson, Spearman y Kendall), cabe aclarar que en estos primeros 3 directorios no se le mueve nada.



Carpeta principal

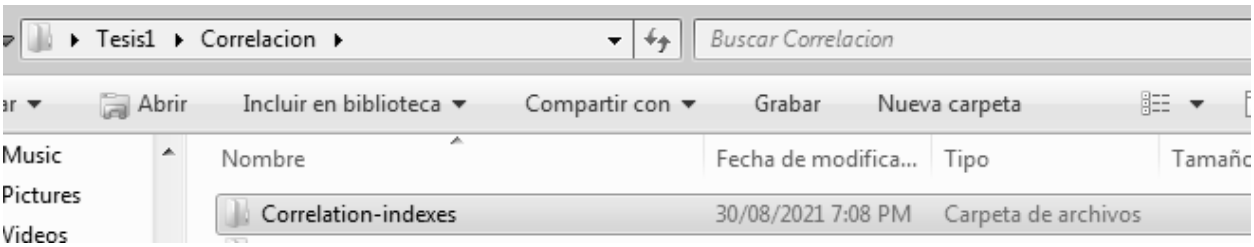


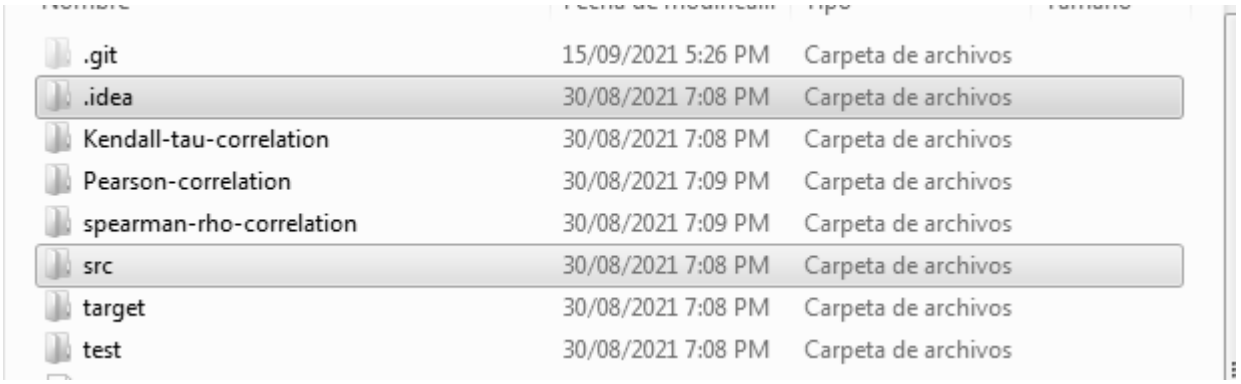
Ilustración 13. Indicadores de correlación

Directorio 1, 2 y 3

Directorio 4. Idea

Directorio 5. Src

Son aquellos que almacenan el código fuente de otro componente encargado de llamar a lost res índices de correlación.

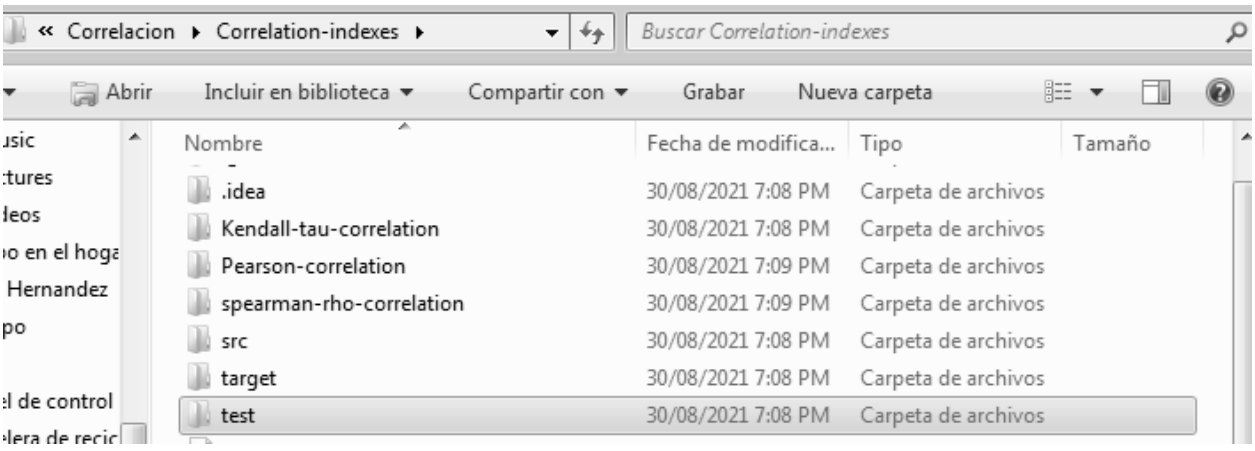


Directorio 4 y 5.

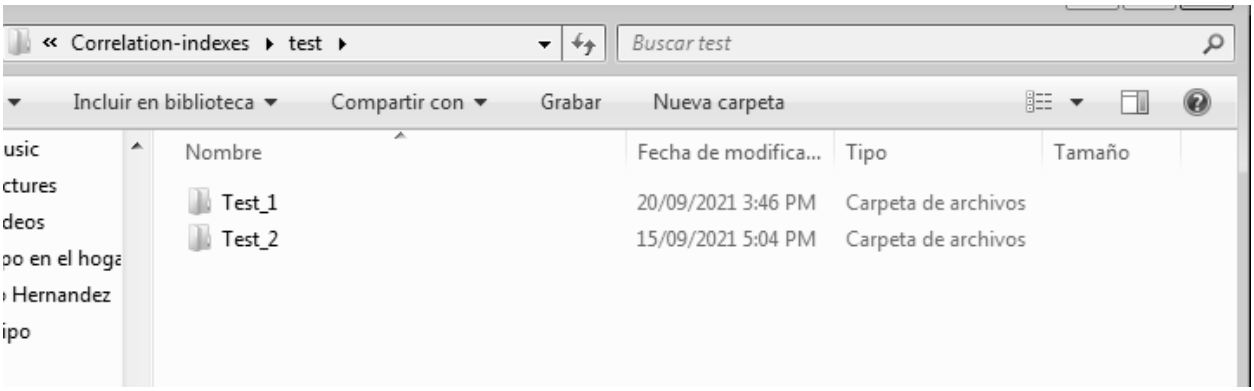
Directorio 6 target: Corresponde al código ejecutable del componente, el cual también hace el llamado a los índices de correlación.

Directorio 6.

Directorio 7 test: Cuenta con dos subcarpetas, en cada una se almacenan los documentos previamente procesados que van a ser evaluados, tomando en cuenta que cada evaluación se realiza por separado.



Directorio 7.



Subcarpetas-Directorio 7.

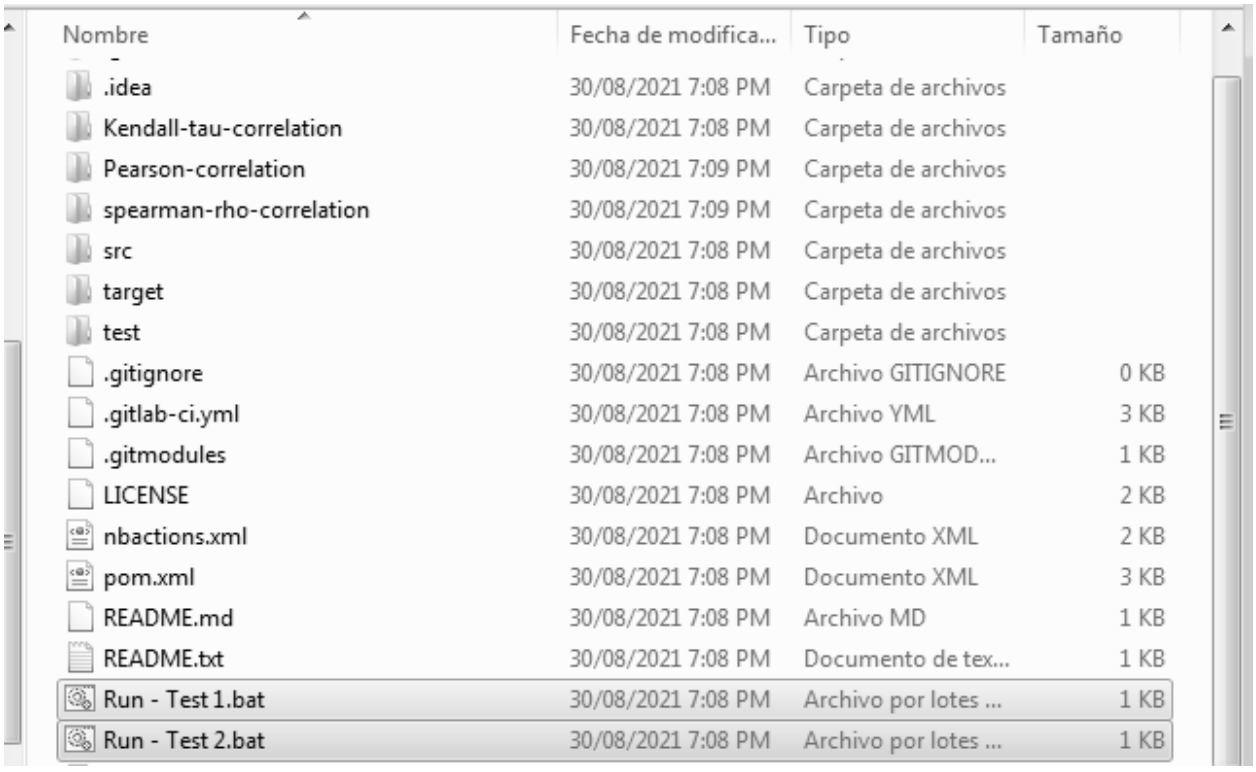
Dentro de la carpeta principal, se tienen dos archivos .bat:

Test 1:

Test 2.

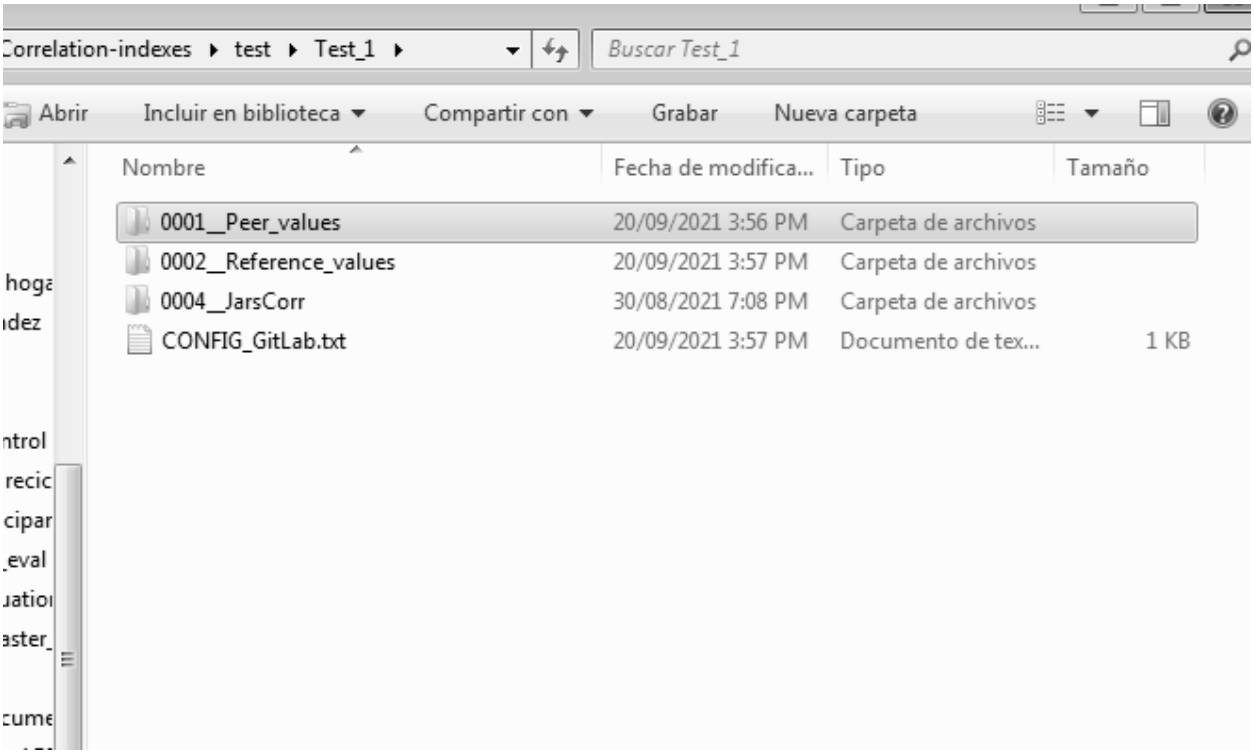
Cada archive .bat se encarga de ejecutar los documentos de manera separada.

Archivos .bat.



Posteriormente dentro del directorio test, se encuentran tres subdirectorios más.

En el directorio 0001__Peer_values: Se encuentran todos los puntajes de la evaluación y dentro del mismo se colocan los 7 archivos de los evaluadores, tanto del conjunto A como del conjunto B.

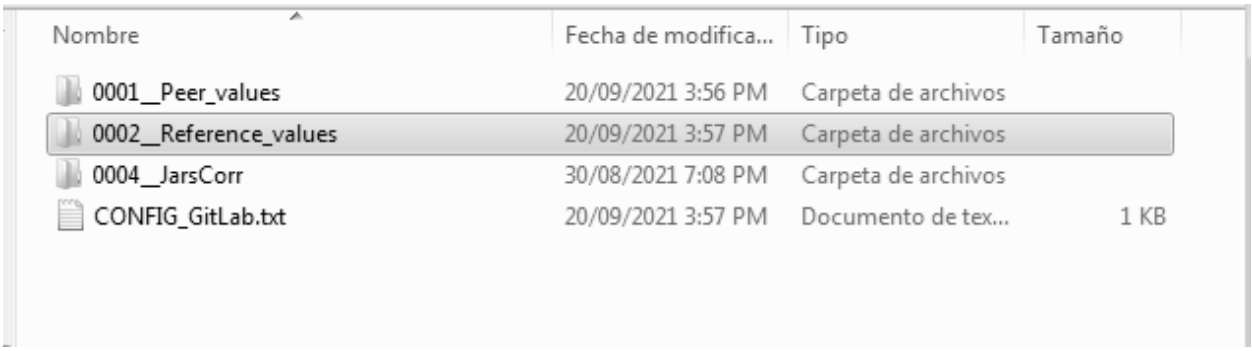


Directorio 0001.

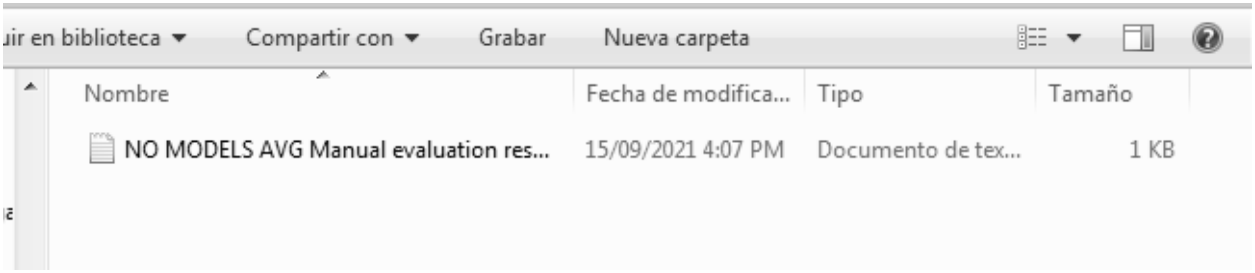
También es importante establecer con quien o con que se van a comparar, para ello tenemos:

Directorio 0002__Reference_values: En este directorio se encuentra un archive delimitado por comas (CSV), en él se almacenan todos los puntajes

de evaluación humana, tanto del conjunto A como del conjunto B, es decir, nos comparamos contra el juicio humano.

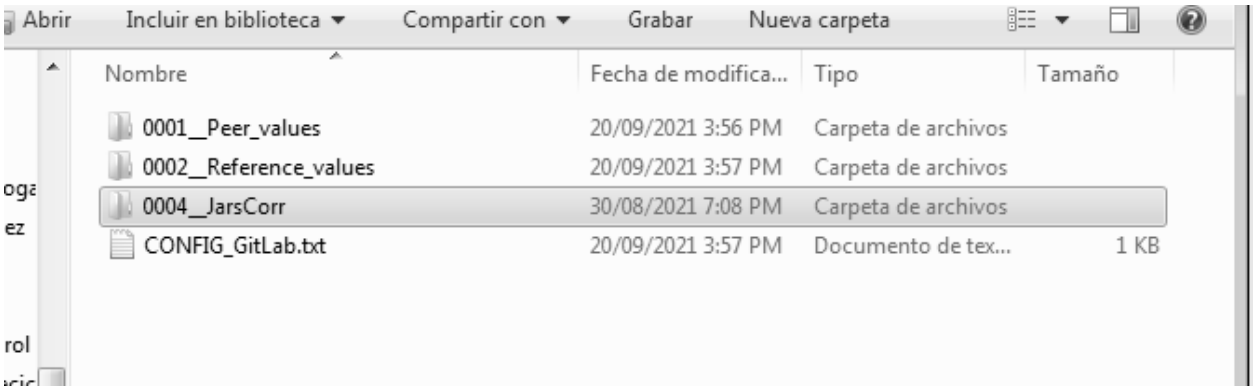


Directorio 0002.



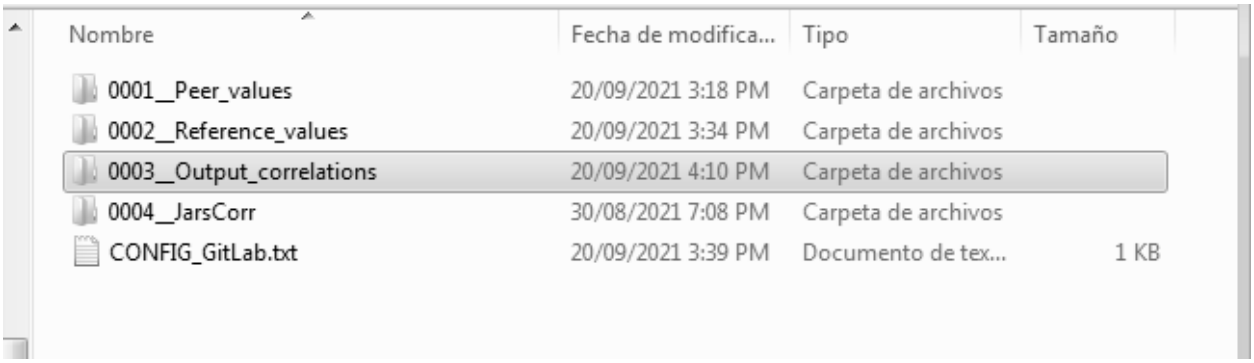
Archivo CSV.

Directorio 0004__JarsCorr: El archive config se encarga de almacenar todas las rutas, teniendo todo lo que se necesita para su funcionamiento.



Directorio 0004.

Directorio 0003: Se crea de manera automática al realizar la ejecución y almacena todos los resultados de correlación.



Directorio 0003.