El Test de Turing para la evaluación de resumen automático de texto

The Turing Test for Automatic Text Summarization Evaluation

Alejandro Molina[1], Juan-Manuel Torres-Moreno[1,2]

[1]Laboratoire Informatique d'Avignon - UAPV, France
[2]Ecole Polytechnique de Montréal, Québec

Resumen

Actualmente existen varios métodos para producir resúmenes de texto de manera automática, pero la evaluación de los mismos continua siendo un tema desafiante. En este artículo estudiamos la evaluación de la calidad de resúmenes producidos de manera automática mediante un método de compresión de frases. Abordamos la problemática que supone el uso de métricas automáticas, las cuales no toman en cuenta ni la gramática ni la validez de las oraciones. Nuestra propuesta de evaluación está basada en el test de Turing, en el cual varios jueces humanos deben identificar el origen, humano o automático, de una serie de resúmenes. También explicamos como validar las respuestas de los jueces por medio del test estadístico de Fisher.

Palabras clave

Evaluación de resumen automático, Compresión de frases, Test de Turing, Crowdsourcing

Abstract

Currently there are several methods to produce summaries of text automatically, but the evaluation of these remains a challenging issue. In this paper, we study the quality assessment of automatically generated abstracts. We deal with one of the major drawbacks of automatic metrics, which do not take into account either the grammar or the validity of sentences. Our proposal is based on the Turing test, in which a human judges must identify the source of a series of summaries. We propose how statistically validate the judgements using the Fisher's exact test.

Keywords

Text Summarization Evaluation, Sentence Compression, Turing Test, Crowdsourcing

1 Introducción

La compresión de frases consiste en eliminar las partes menos importantes de una oración o de una frase de manera automática. Aunque es un tema relativamente reciente, ya se han propuesto diversas aplicaciones para su uso, por ejemplo en los dispositivos móviles que cuentan con pantallas reducidas en tamaño y donde el número de caracteres mostrados es limitado. La compresión de frases permitiría reducir la extensión del texto mostrado y, de esta manera, incluir más información en un espacio reducido.

Otra aplicación es la de traducción automática de subtítulos. Un modulo de traducción automática puede estar acoplado con módulo de compresión de frases de manera que se garantice una longitud específica del texto traducido. La compresión de frases también puede servir para ayudar a las personas con problemas visuales. Grefenstette (1998) presenta un método de reducción de textos que tiene por objetivo disminuir el tiempo de lectura de un sintetizador para ciegos.

Si bien es cierto que el tema de resumen automático continúa siendo de mucho interés, la evaluación presenta aún muchos retos que necesitan ser considerados y estudiados. Por ejemplo, se sabe poco acerca de la subjetividad en los criterios para calificar un resumen. En este artículo abordaremos metódicamente el tema de la calidad mínima esperada para un resumen. Esto es, dejando de lado que el resumen sea bueno o malo, verificar que éste cumple con las expectativas mínimas esperadas. Dicho de otra manera, que no se pueda distinguir si el resumen ha sido producido por una máquina o por una persona.

En la sección 2 pondremos en evidencia la necesidad de considerar cierta calidad mínima en la evaluación. Después, en la sección 3 discutiremos algunos métodos de evaluación de resúmenes y puntualizaremos por qué algunos de ellos no resultan adecuados para evaluar resúmenes por

compresión. En las secciones siguientes propondremos un método de evaluación basado en el test de Turing cuyos resultados pueden ser validados estadísticamente con una prueba de hipótesis.

2 Resumen por compresión de frases

La compresión de frases fue definida por Knight & Marcu (2000) como un método de reducción de oraciones. Los autores proponen algoritmos para eliminar palabras de una frase, sin cambiar el orden, de manera que la secuencia resultante, considerada como una compresión de la original, puede o no ser una oración válida en Inglés.

En (Molina, 2013) se plantea usar la compresión de frases como un método para generar resúmenes de manera automática. La idea es eliminar ciertos elementos de las frases de un texto pero considerando su contexto original en lugar de comprimir las frases aisladas. Para esto, se propone dividir la oración en segmentos discursivos y luego, mediante un algoritmo basado en aprendizaje de máquina, se decide cuáles de los segmentos se pueden eliminar. Los criterios para generar el resumen son que éste sea más corto, informativo y gramaticalmente correcto.

Sin embargo, la evaluación de un resumen por compresión de frases es un tema que merece ser tratado cuidadosamente. A diferencia del método de resumen automático por extracción, el resumen por compresión de frases puede modificar la estructura gramatical de las oraciones.

Por ejemplo, considere el Cuadro 2 que presenta las 2^3 compresiones posibles de una frase con 3 segmentos. Sea φ = [En casa es útil tener un termómetro]$\_s_1$ [para saber con precisión]$\_s_2$ [si alguien de la familia tiene fiebre.]$\_s_3$. Note que las compresiones $\widetilde{\varphi}_3$, $\widetilde{\varphi}_4$, $\widetilde{\varphi}_6$ y $\widetilde{\varphi}_7$ no son gramaticalmente correctas o cambian el sentido original de la frase. Esto nos lleva a concluir que la evaluación de un resumen por compresión de frases debe considerar la validez de las frases resultantes. En la sección 3 discutiremos más en detalle algunos métodos de evaluación de resúmenes.

3 La evaluación de resúmenes

La evaluación del resumen automático ha sido una cuestión compleja, que ha propiciado el surgimiento de varios enfoques. En (Amigó et al., 2005) se discuten ampliamente varios métodos de evaluación pero nuestro interés principal es con respecto a la fuente de validación de los resúmenes que puede ser: manual o automática.

3.1 Evaluación manual

La evaluación manual consiste básicamente en la lectura y comparación de los resúmenes automáticos con respecto a los resúmenes producidos por humanos (Edmundson, 1969). Su principal ventaja es que el criterio humano es garantía de validez y pertinencia. Su principal desventaja es que a partir de un mismo texto se puede producir una infinidad de resúmenes válidos y esto puede provocar que los evaluadores no muestren acuerdo.

Como parte de los trabajos representativos de la evaluación manual esta el de Mani et al. (1999) en el que los autores proponen dar a los anotadores tanto los resúmenes producidos con métodos automáticos como los documentos originales. La hipótesis es que los originales, contienen frases-clave que determinan la temática del texto y por lo tanto éstas deben estar incluidas en un resumen. Los anotadores deben comparar los documentos con los resúmenes y verificar que estos últimos en efecto contengan dichas frases-clave.

Saggion & Lapalme (2000) propone otro método en la misma línea pero aplicado a resúmenes mono-documento. La variante es que en lugar de frases, se entrega a los anotadores una lista de conceptos-clave que deben ser mencionados en los resúmenes automáticos.

Orasan & Hasler (2007) propone evaluar la calidad de un resumen con un test comparativo. Los anotadores deben elegir el mejor resumen de entre un par tal que uno de ellos fue elaborado con una herramienta de resumen asistido por computadora (*Computer-Aided Summarization*) y el otro sin esta herramienta. Su hipótesis es que no existe diferencia, estadísticamente significativa, entre ambos tipos y por lo tanto, los anotadores son incapaces de distinguirlos.

3.2 Evaluación automática

La evaluación automática consiste en que un programa evalúe los resúmenes. Su principal ventaja es que permite tratar cantidades masivas de documentos. Sin embargo, muchos métodos automáticos, no consideran ni la coherencia ni la validez gramatical ni la sucesión retórica de las ideas. Es decir que para muchos de estos métodos no importa el orden lógico de las palabras sino simplemente si aparecen o no. En las subsecciones siguientes, mencionaremos las características de algunos métodos automáticos.

$\widetilde{\varphi}_1$	(s_1, s_2, s_3)	En casa es útil tener un termómetro para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_2$	(s_1, s_3)	En casa es útil tener un termómetro si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_3$	(s_1, s_2)	En casa es útil tener un termómetro para saber con precisión.
$\widetilde{\varphi}_4$	(s_2, s_3)	Para saber con precisión si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_5$	(s_1)	En casa es útil tener un termómetro.
$\widetilde{\varphi}_6$	(s_2)	Para saber con precisión.
$\widetilde{\varphi}_7$	(s_3)	Si alguien de la familia tiene fiebre.
$\widetilde{\varphi}_8$	$()$	

Cuadro 1: Ejemplo de las compresiones posibles de una frase.

3.2.1 Evaluación usando referencias: ROUGE

Uno de los métodos de evaluación automática más utilizados es ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Éste se utiliza incluso durante las campañas internacionales de *Document Understanding Conferences - Text Analysis Conference* desde el año 2008. La idea general es comparar un resumen candidato (automático) con varios resúmenes elaborados por expertos, llamados resúmenes modelo o referencias. Así, la métrica de evaluación se basa en la coocurrencia de n-gramas entre el resumen candidato y los de referencia. La ecuación (1) corresponde a la fórmula para calcular la métrica ROUGE-N (R_n), donde n es el tamaño del n-grama y $Cnt_m(gram_n)$ es el número máximo de n-gramas que aparecen tanto en el resumen candidato como en las referencias (refs).

$$R_n = \frac{\sum_{S\epsilon\{\text{refs}\}} \sum_{gram_n \epsilon S} Cnt_m(gram_n)}{\sum_{S\epsilon\{\text{refs}\}} \sum_{gram_n \epsilon S} Cnt(gram_n)} \quad (1)$$

Si bien ROUGE ha tenido gran aceptación entre la comunidad, no resulta adecuado para evaluar resúmenes con frases comprimidas. Note que la métrica se basa en la cobertura entre el resumen candidato y el conjunto de los resúmenes de referencia. Dado que los resúmenes con frases comprimidas contienen menos palabras, estos son penalizados aunque su contenido sea pertinente. Por esta razón, ROUGE no es adecuado para evaluar la calidad de un resumen producido mediante la compresión de frases.

3.2.2 Evaluación inspirada en la traducción automática: BLEU

En (Molina et al., 2010), se estudia la evaluación de frases comprimidas mediante una métrica semiautomática, desarollada originalmente por IBM para la tarea de la traducción automática, llamada BLEU (Papineni et al., 2002). Inspirada por ROUGE, esta métrica está basada en la precisión entre los n-gramas de una frase candidata y un conjunto de frases de referencia. La idea general es calcular la proporción de n-gramas de la frase candidata presentes en las referencias y el número total de n-gramas de la frase candidata. En la ecuación (2), C corresponde a la frase candidata y Cnt_{clip}(n) es el número máximo de veces que un n-grama de la frase candidata fue encontrado en alguna de las referencias. Una penalización con respecto a la longitud (*Brevity Penalty*, BP) se impone a las frases demasiado largas o demasiado cortas en la ecuación (3). Cuanto más corta sea la frase mayor será su penalización. En consecuencia, la frases comprimidas obtienen un score bajo. Por lo tanto, BLEU no es adecuado para evaluar resúmenes producidos por compresión de frases.

$$P_n = \frac{\sum_{C\epsilon\{\text{Cands}\}} \sum_{\text{n-gram}\epsilon C} Cnt_{clip}(\text{n})}{\sum_{C\epsilon\{\text{Cands}\}} \sum_{\text{n-gram}\epsilon C} Cnt(\text{n})} \quad (2)$$

$$\text{BLEU} = BP \times e^{\left(\sum_{n=1}^{N} \frac{1}{N} \mathbf{log}(P_n)\right)} \quad (3)$$

3.2.3 Evaluación sin referencias: FRESA

En (Molina et al., 2012), se intenta evaluar resúmenes con frases comprimidas utilzando la métrica FRESA (Torres-Moreno et al., 2010; Saggion et al., 2010; Torres-Moreno, 2014), la cual no requiere resúmenes de referencia dado que solamente utiliza el documento de origen. La idea es calcular las divergencias entre las distribuciones de frecuencias de términos entre el resumen que se quiere evaluar y el texto de origen. Estas divergencias corresponden a las de Kullback-Leibler (KL) y Jensen-Shannon (JS) como se describe en (Louis & Nenkova, 2008).

Sea T el conjunto de términos contenidos en el documento de origen. Para cada $t \in T$, C_t^T es el número de apariciones de t en el documento de origen y C_t^S es el número de apariciones de t

en el resumen que se quiere evaluar. En la ecuación (4), se calcula la diferencia absoluta entre las divergencias de dichas distribuciones (en el espacio *log*). Los valores altos (poca divergencia) están asociados a la similitud entre el resumen y el texto de origen mientras que los balores bajos (alta divergencia) implican disimilitud entre ellos.

$$\mathcal{D} = \sum_{t \in T} \left| \log\left(\frac{C_t^T}{|T|} + 1\right) - \log\left(\frac{C_t^S}{|S|} + 1\right) \right| \quad (4)$$

La interpretación de esta métrica no es trivial ni intuitiva. La única conclusión que se puede sacar a partir de los valores de la métrica es que el valor de divergencia entre un texto y su resumen es elevado, pero esto siempre es aplicable a las frases comprimidas como se muestra en los experimentos de (Molina et al., 2012). Así, FRESA asocia valores de alta divergencia independientemente de la estrategia utilizada, incluyendo la compresión aleatoria. Por lo tanto, tampoco resulta una manera adeacuada de evaluación de resúmenes por compresión de frases.

Después de presentar tres diferentes medidas automáticas en esta sección, queda claro que ninguna de ellas toma en cuenta la estructura gramatical de las frases comprimidas. En efecto, una de las principales desventajas de las evaluaciones automáticas es que no consideran ni la gramática ni la retórica ya que se basan solamente en las apariciones de elementos léxicos como los n-gramas.

En la siguiente sección, proponemos afrontar la problemática de la evaluación de otra manera, usando el test de Turing.

4 El juego de la imitación

Las ideas que tuvo Alan Turing, acerca de las máquinas y el pensamiento, siguen generando polémica. Sin embargo, vamos a explorar cómo pueden resultar ventajosas para la evaluación de algunas tareas del Procesamiento de Lenguaje Natural (PLN) y en concreto del resumen automático.

Nos referimos al famoso test de Turing descrito en el artículo (Turing, 1950) en el cual se discute la cuestión: "¿Y si las máquinas pudieran pensar?"

Para evitar la complicación de tener que definir qué significa pensar, Turing estableció el juego de la imitación, hoy conocido como el test de Turing. En el juego hay dos jugadores y un juez. El primer jugador es un ser humano (A) y el segundo es una máquina (B). Otra persona que funge como el juez (C) debe adivinar la identidad de cada uno de los jugadores sin verlos. Unicamente se permite que los jugadores interactúen con el juez mediante una terminal. Por ejemplo, el juez escribe preguntas con la ayuda de un teclado y lee las respuestas de los jugadores en una pantalla. Al final del juego, el juez debe indicar quién es la máquina y quién es el humano a partir de las respuestas obtenidas durante el intercambio.

Por supuesto, el objetivo de este experimento hipotético propuesto por Turing no era el de engañar a alguien en particular, sino el de plantear cuestiones filosóficas en torno al pensamiento. Concretamente, sobre la posibilidad de recrear artificialmente las funciones cognitivas del cerebro humano y sobre la posibilidad de evaluar si dichas funciones corresponden a lo que podríamos esperar de "algo"que piensa.

Para nuestro estudio, hemos rescatado algunos aspectos del protocolo del test que nos parecen aplicables a la evaluación de una tarea compleja de procesamiento del lenguaje y para la cual no se ha propuesto ningún método eficaz. Concordamos con Harnad (2000) sobre el hecho de que Turing privilegió, en el test, la comunicación por medio del lenguaje natural. ¿No es acaso la lengua uno de los principales medios para vehicular el pensamiento? No obstante, las cuestiones filosóficas del test de Turing no conciernen el presente estudio. Nuestro objetivo es simplemente validar, mediante una adaptación del test de Turing, un tipo específico de función lingüística: la generación de resúmenes.

En el test de Turing el juez no tiene el derecho de ver a los jugadores. Con esta restricción, Turing puso de manifiesto que son los aspectos funcionales y no los aspectos físicos los que deben ser juzgados. Nos parece entonces natural utilizar este test para evaluar tareas del PLN cuyo objetivo es simular la habilidad de los humanos. A este respecto, consideramos una variante del test de Turing destinada a la evaluación de resúmenes automáticos.

5 El test de Turing para evaluar resúmenes automáticos

Supongamos que un ser humano (A) y una máquina (B) producen respectivamente dos resúmenes a partir del mismo documento. (A) y (B) deben respetar las mismas reglas para que las producciones sean homogénas y, en consecuencia, comparables. Un juez humano (C), debe determinar cuál de los resúmenes fue elaborado por (A) y cual fue elaborado por (B). Para esto,

el juez debe revelar la identidad de cada jugador apoyándose unicamente en la lectura de sus resúmenes.

5.1 Protocolo experimental

Para nuestro experimento, hemos convocado 54 jueces quienes leyeron y evaluaron exactamente los mismos resúmenes. Como la evaluación requiere la lectura directa, se eligieron solamente seis resúmenes humanos (A) y seis resúmenes automáticos (B) producidos, en ambos casos, mediante el mismo algoritmo (Algoritmo 1), descrito en (Molina, 2013).

El Algoritmo 1 toma como argumentos un umbral de probabilidad ($\alpha \in [0, 1]$) y el documento a resumir (Doc). Primero, el documento original es segmentado en frases y luego en segmentos discursivos. Posteriormente, se decide para cada segmento si éste debe ser eliminado según un valor de probabilidad. El algoritmo termina cuando se han procesado todas las frases del documento y se ha producido un resumen. La computadora utilizó un modelo de regresión lineal que calcula la probabilidad de eliminar un segmento basándose en el aprendizaje de 60 844 segmentos anotados manualmente. Para el caso de los humanos, la desición está basada simplemente en su criterio para decidir si algún segmento es importante o no lo es.

Algoritmo 1 Generación de resúmenes por eliminación de segmentos.

Argumentos:(α, Doc)
$Segmentar_{\varphi}(Doc)$ //En frases.
$Segmentar_{s}(Doc)$ //En segmentos discursivos.
para todo φ en Doc **hacer**
  para todo s en φ **hacer**
    si ($\mathbf{P_{elim}}(s, \varphi) > \alpha$) **entonces**
      $Eliminar(s)$ de φ
    fin si
  fin para
fin para
  devolver resumen // Doc con φs modificadas.

Las frases fueron segmentadas mediante dos métodos distintos, un segmentador retórico para el español llamado DiSeg (da Cunha et al., 2012) y un segmentador adaptado a la compresión de frases (Molina, 2013). Para cada segmentador seleccionamos tres categorías de resumen de acuerdo con la tasa de compresión τ (Cuadro 2): poca compresión ($\tau < 50\,\%$), compresión media ($\tau \approx 50\,\%$) y mucha compresión ($\tau > 50\,\%$). Para el test, conservamos los que tenían mejores scores en gramática para cada una de las categorías.

Los 54 jueces (C), todos hispanohablantes con nivel de estudios de 4 o más años en la universidad, ignoraban toda la información respecto al juego de la imitación. Unicamente se les otorgaron los doce resúmenes y se les dio una sola instrucción: determinar para cada resumen si éste había sido producido por un humano o por una máquina. El Anexo 1, al final de este artículo, se muestra una copia del documento entregado a los jueces.

	Pals. origen	**Pals. resumen**	τ (%)	
$\tau < 50\,\%$	303	49	16.1	DiSeg
$\tau \approx 50\,\%$	209	104	49.6	DiSeg
$\tau > 50\,\%$	156	119	76.3	DiSeg
$\tau < 50\,\%$	217	57	26.2	CoSeg
$\tau \approx 50\,\%$	165	76	43.4	CoSeg
$\tau > 50\,\%$	234	186	79.4	CoSeg

Cuadro 2: Criterios de selección para la evaluación de resúmenes con un test de Turing.

6 La catadora de té

Para validar estadísticamente nuestros resultados, nos inspiramos en el experimento de la "dama del té", descrito en (Agresti, 2002), por medio del cual Ronald A. Fisher desarrolló un test estadístico exacto. Una dama (Muriel Bristol) se jactaba de ser capaz de distinguir si una taza de té con leche había sido servida primero con la leche o primero con el té. Para examinar su pretensión, Fisher le pidió probar 8 tazas de té con leche. En 4 de ellas se sirvió el té sobre la leche y en las otras 4 se sirvió la leche sobre el té. El test estadístico propuesto por Fisher se basa sobre el conteo del número de buenas y malas respuestas mediante una tabla de contingencia como la del Cuadro 3.

	Respuesta correcta	
Respuesta dama	*primero leche*	*primero té*
primero leche	a =3	b =1
primero té	c =1	d =3

Cuadro 3: Tabla de contingencia de las respuestas de la catadora de té.

Fisher mostró que la probabilidad de obtener una tabla de contingencia como la del del Cuadro 3 esta dada por la ley hipergeométrica (ecuación 5). Donde $\binom{l}{k}$ es el coeficiente binomial y n es la suma de todas las celdas de la tabla. Usando las respuestas del Cuadro 3, se tiene que $p = 0{,}229$ es la probabilidad de obtener los resultados de esa tabla por mera coincidencia. Se sigue que las

respuestas de la catadora no establecen prueba de su supuesta habilidad.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \quad (5)$$

7 Validación de resultados del juego de la imitación con el test exacto de Fisher

En nuestra evaluación, le pedimos a los 54 jueces que distinguieran para cada uno de los 12 resúmenes si había sido creado por un humano o por una máquina y se crearon tablas de contingencia, como la mostrada en el Cuadro 4. Finalmente se aplicó el test de Fisher a las respuestas obtenidas.

	Origen del resumen	
Respuesta juez	*Humano*	*Máquina*
Humano	a	b
Máquina	c	d

Cuadro 4: Tabla de contingencia para la evaluación de resúmenes con el test exacto de Fisher.

La hipótesis nula de nuestros tests (H_0) es que no existe asociación entre las respuestas y el origen del resumen. La hipótesis alternativa (H_1) es que sí existe una asociación positiva. Utilizamos la función *fisher.test* del lenguaje de programación R para calcular los valores de p. En nuestros experimentos utilizamos la configuración estándar del test: dos cola con un intervalo de confianza del 95 %.

El Cuadro 9 muestra las tablas de contingencia así como los valores de p de la evaluación del origen de los resúmenes para los 54 jueces del experimento.

Una sola persona (el Juez 1 en el Cuadro 9) presenta un resultado estadísticamente significativo, de haber distinguido entre los dos tipos de resúmenes. Respecto a los 53 jueces restantes, no podemos decir que el resultado sea significativo en la distinción del origen verdadero de los resúmenes. Nos inclinamos a afirmar entonces que los 53 jueces restantes encontraron la misma calidad en los resúmenes manuales que en los automáticos.

8 Evaluación de resúmenes según el tipo de segmentación y el tamaño

También utilizamos el test exacto de Fisher para verificar si se observan diferentes resultados según el segmentador automático empleado. El Cuadro 5 muestra el número de veces que los jueces identificaron correcta o incorrectamente los resúmenes según el segmentador. Para afirmar con significación estadística que una segmentación en particular permite identificar más facilmente el origen de los resúmenes, la hipótesis nula es, en este caso, que el grado de identificación es independiente del tipo de segmentación. Los resultados dan un valor $p = 0{,}4965$ al 95 % con un intervalo de confianza de $[0{,}63; 2{,}76]$. Se sigue que como $p > 0{,}05$, entonces aceptamos H_0: el hecho de que un resumen haya sido segmentado con DiSeg o CoSeg no influye en la identificación realizada por los jueces.

	Origen correctamente identificado	**Origen erroneamente identificado**
DiSeg	45	63
CoSeg	19	35

Cuadro 5: Evaluación de la influencia del tipo de segmentación para la identificación de los resúmenes.

	Correctamente identificado (*observado*)	**Correctamente identificado** (*esperado*)
$\tau < 50\,\%$	27	25
$\tau \approx 50\,\%$	30	25
$\tau > 50\,\%$	18	25

Cuadro 6: Valores esperados y observados en la identificación correcta del origen de los resúmenes.

	Erróneamente identificado (*observado*)	**Erróneamente identificado** (*esperado*)
$\tau < 50\,\%$	27	29
$\tau \approx 50\,\%$	24	29
$\tau > 50\,\%$	36	29

Cuadro 7: Valores esperados y observados en la identificación errónea del origen de los resúmenes.

	Correctamente identificado	**Erróneamente identificado**
$\tau < 50\,\%$	0.668	-0.668
$\tau \approx 50\,\%$	1.671	-1.671
$\tau > 50\,\%$	-2.339	2.339

Cuadro 8: Desviación estándar de la varianza residual con respecto a la tasa de compresión τ en la identificación de resúmenes automáticos.

Juez id	Contingencia	p	◦	Juez id	Contingencia	p	◦	Juez id	Contingencia	p	◦
Juez 1	4	0.030	falso	Juez 2	3	0.500	verdadero	Juez 3	5	0.772	verdadero
	0				2				5		
	2				3				1		
	6				4				1		
Juez 4	1	0.998	verdadero	Juez 5	3	0.716	verdadero	Juez 6	4	0.727	verdadero
	5				3				4		
	5				3				2		
	1				3				2		
Juez 7	5	0.772	verdadero	Juez 8	4	0.283	verdadero	Juez 9	5	0.272	verdadero
	5				2				3		
	1				4				1		
	1				4				3		
Juez 10	3	0.969	verdadero	Juez 11	3	0.878	verdadero	Juez 12	4	0.283	verdadero
	1				4				2		
	3				3				2		
	5				2				4		
Juez 13	3	0.716	verdadero	Juez 14	3	0.716	verdadero	Juez 15	2	0.878	verdadero
	3				3				3		
	3				3				4		
	3				3				3		
Juez 16	3	0.969	verdadero	Juez 17	1	0.969	verdadero	Juez 18	3	0.500	verdadero
	5				3				2		
	3				5				3		
	1				3				4		
Juez 19	3	0.500	verdadero	Juez 20	2	0.960	verdadero	Juez 21	3	0.716	verdadero
	2				4				3		
	3				2				3		
	4				4				3		
Juez 22	3	0.878	verdadero	Juez 23	2	1.000	verdadero	Juez 24	1	0.992	verdadero
	4				6				4		
	3				4				5		
	2				0				2		
Juez 25	2	0.727	verdadero	Juez 26	4	0.283	verdadero	Juez 27	2	0.960	verdadero
	2				2				4		
	4				4				2		
	4				4				4		
Juez 28	4	0.727	verdadero	Juez 29	1	0.969	verdadero	Juez 30	3	0.878	verdadero
	4				3				4		
	2				5				3		
	2				3				2		
Juez 31	5	0.772	verdadero	Juez 32	2	0.960	verdadero	Juez 33	3	0.878	verdadero
	5				4				4		
	1				2				3		
	1				4				2		
Juez 34	2	0.960	verdadero	Juez 35	3	0.878	verdadero	Juez 36	4	0.960	verdadero
	4				4				2		
	4				2				4		
	2				2				2		
Juez 37	4	0.909	verdadero	Juez 38	3	0.878	verdadero	Juez 39	4	0.727	verdadero
	5				4				4		
	1				2				2		
	2				3				2		
Juez 40	4	0.283	verdadero	Juez 41	4	0.727	verdadero	Juez 42	1	0.998	verdadero
	2				4				5		
	4				2				5		
	2				2				1		
Juez 43	3	0.878	verdadero	Juez 44	2	1.000	verdadero	Juez 45	5	1.000	verdadero
	3				6				6		
	3				4				1		
	3				0				0		
Juez 46	4	0.500	verdadero	Juez 47	2	0.878	verdadero	Juez 48	4	0.283	verdadero
	3				3				2		
	3				4				2		
	3				3				4		
Juez 49	3	0.716	verdadero	Juez 50	2	0.878	verdadero	Juez 51	4	0.500	verdadero
	3				3				3		
	3				4				2		
	3				3				3		
Juez 52	2	0.992	verdadero	Juez 53	2	0.992	verdadero	Juez 54	3	0.716	verdadero
	5				5				3		
	2				2				3		
	1				1				3		

Cuadro 9: Resultados del test de Turing en evaluación de resumen automático aplicado a 54 jueces.

Para verificar la influencia de la tasa de compresión de un resumen en la elección de los jueces, utilizamos el test de χ^2. En este caso, no podemos utilizar el test exacto de Fisher porque la tabla de contingencia asociada es de 3×2 (Cuadros 6 y 7). Los resultados del test de χ^2 dan un valor $p = 0{,}0547$, apenas superior al valor crítico lo que nos llevó a comparar los valores esperados bajo la hipótesis nula en ambos cuadros, confirmandose que, para los resúmenes con $\tau > 50\,\%$, resultó más difícil identificar el origen artificial.

Este hecho puede confirmarse en el Cuadro 8 a partir de las varianzas residuales donde la desviación estándar para los resúmenes fijando $\tau > 50\,\%$ es más de dos veces superior a la media. Para los jueces resultó muchos más complicado identificar correctamente un resumen automático cuando había sido menos comprimido.

9 Conclusiones

En este trabajo hemos abordado la evaluación de resúmenes de documentos textuales producidos con métodos automáticos. La motivacion principal de este trabajo es que, a pesar que existen métodos efectivos de evaluación para resúmenes por extracción, estos resultan inadecuados para evaluar resúmenes por compresión de frases, porque que no toman en cuenta la gramática.

Ante este panorama, hemos propuesto un método basado en el test de Turing en el que los jueces humanos deben develar el origen (automático o manual) de varios resúmenes, y por medio del test exacto de Fisher, se calcula la fiabilidad de las respuestas de dichos jueces.

Aunque hemos aplicado la evaluación al área de resumen automático, encontramos que la metodología resulta lo bastante general para ser aplicada a cualquier otra área del procesamiento del lenguaje natural.

Agradecimentos

A todos los voluntarios que realizaron el test. A Mariana Tello Signoret por su ayuda con las traducciones. A Eric SanJuan, Gerardo Sierra y Carlos Mendez por la asesoría para la realización de este trabajo.

Referencias

Agresti, Alan. 2002. *Categorical data analysis*, vol. 359. Wiley interscience.

Amigó, Enrique, Julio Gonzalo, Anselmo Peñas & Felisa Verdejo. 2005. Qarla: a framework for the evaluation of text summarization systems. En *43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 280–289. Ann Arbor, MI, Etats-Unis: ACL.

da Cunha, Iria, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes & Irene Castellón. 2012. Diseg 1.0: The first system for spanish discourse segmentation. *Expert Systems with Applications* 39(2). 1671–1678.

Edmundson, H. P. 1969. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery* 16(2). 264–285.

Grefenstette, G. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. En *AAAI Spring Symposium on Intelligent Text summarization (Working notes)*, 111–118. Stanford University, CA, Etats-Unis.

Harnad, Stevan. 2000. Minds, machines and turing. *Journal of Logic, Language and Information* 9(4). 425–445.

Knight, Kevin & Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. En *17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 703–710. Austin, TX, Etats-Unis.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. En Marie-Francine Moens & Stan Szpakowicz (eds.), *Workshop Text Summarization Branches Out (ACL'04)*, 74–81. Barcelone, Espagne: ACL.

Louis, Annie & Ani Nenkova. 2008. Automatic Summary Evaluation without Human Models. En *First Text Analysis Conference (TAC'08)*, Gaithersburg, MD, Etats-Unis.

Mani, Inderjeet, David House, Gary Klein, Lynette Hirschman, Therese Firmin & Beth Sundheim. 1999. The tipster summac text summarization evaluation. En *ninth conference on European chapter of the Association for Computational Linguistics*, 77–85. ACL.

Molina, Alejandro. 2013. Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. *Inteligencia Artificial* 16(51). 41–62.

Molina, Alejandro, Iria da Cunha, Juan-Manuel Torres-Moreno & Patricia Velazquez-Morales. 2010. La compresión de frases: un recurso para la optimización de resumen automático de documentos. *Linguamática* 2(3). 13–27.

Molina, Alejandro, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan & Gerardo Sierra. 2012. Sentence compression in spanish driven by discourse segmentation and language models. *Cornell University, Computation and Language (cs.CL), Information Retrieval (cs.IR)* arXiv:1212.3493.

Orasan, Constantin & Laura Hasler. 2007. Computer-aided summarisation: how much does it really help. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)* 437–444.

Papineni, K., S. Roukos, T. Ward, & W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 311–318. Philadelphie, PA, Etats-Unis: ACL.

Saggion, H. & G. Lapalme. 2000. Concept identification and presentation in the context of technical text summarization. En *ANLP/NAACL Workshop on Automatic Summarization*, 1–10. Seattle, WA, Etats-Unis: ACL.

Saggion, Horacio, Juan-Manuel Torres-Moreno, Iria da Cunha & Eric SanJuan. 2010. Multilingual summarization evaluation without human models. En *23rd International Conference on Computational Linguistics: Posters (COLING'10)*, 1059–1067. Beijing, Chine: ACL.

Torres-Moreno, Juan-Manuel. 2014. *Automatic text summarization*. Wiley and Sons.

Torres-Moreno, Juan-Manuel, Horacio Saggion, Iria da Cunha & Eric SanJuan. 2010. Summary Evaluation With and Without References. *Polibits: Research journal on Computer science and computer engineering with applications* 42. 13–19.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460.

A Anexo 1: Test de evaluación

Algunos de los siguientes doce resúmenes que se muestran a continuación han sido creados de manera automática por un programa y otros han sido creados por humanos. Determine cuáles.

La persona con el cociente intelectual más alto del mundo
Su nombre es Marilyn vos Savant y nació en San Louis (Missouri) el 11 de agosto de 1946. Marilyn vos Savant está considerada como la persona con el cociente intelectual más alto del mundo. Hoy en día es una más que reputada columnista, escritora, conferenciante y dramaturga. En 1986 comenzó una columna dominical llamada Pregunta a Marilyn (Ask Marilyn) en la revista Parade, donde responde preguntas de los lectores acerca de diversos temas. Su mayor aspiración era el convertirse en escritora. Durante su juventud trabajó en la tienda de ultramarinos de su padre. Cursó varios seminarios de filosofía en la universidad. En la actualidad está casada con el prestigioso cardiólogo Robert Jarvik. A Marilyn se le asocia con el famoso problema de Monty Hall, o bien le fue planteado a ella a través de una consulta en su columna Ask Marilyn.

Cuadro 10: Resumen de *La persona con el cociente intelectual más alto del mundo* (tipo de segmentación: DiSeg, origen del resumen : Humano, τ =51.83 % del contenido original).

El Pulque
El Pulque o Neutle se obtiene de la fermentación de la savia azucarada o aguamiel, concentrados en el corazón de el maguey, antes de que salga el pedúnculo de la inflorescencia del maguey por el proceso conocido como raspado, que consiste en quitar el centro de la planta donde crecen las hojas tiernas dejando una oquedad que se tapa con una penca del maguey. El interior es entonces raspado con una especie de cuchara, lo que provoca que el maguey suelte un jugo el cual se concentra en el hueco. Este es, luego, a intervalos de uno o dos días absorbido hacia un cuenco hueco (llamado acococte, fruto de una cucurbitácea) y depositado en un recipiente llamado odre. Este proceso lo lleva a cabo el Tlachiquero o raspador, y el jugo se recolecta durante dos meses como máximo. Después es depositado en barriles de pino o, en cubas de acero inoxidable, donde se fermenta con la bacteria Zymomonas mobilis durante uno o dos días obteniendose un líquido blanco de aspecto lechoso con un 5 % de alcohol. Se debe beber inmediatamente ya que al seguirse fermentando adquiere un gusto muy fuerte.

Cuadro 11: Resumen de *El Pulque* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =90.90 % del contenido original).

La música en el antiguo Egipto
La Música en el antiguo Egipto se empleaba en varias actividades, pero su desarrollo principal fue en los templos, donde era usada durante los ritos dedicados a los diferentes dioses y era utilizada como remedio terapéutico. Como en otros pueblos, también se consideraba un medio de comunicación con los difuntos y los músicos alcanzaban una categoría tal que algunos están enterrados en las necrópolis reales. No se conoce cómo era realmente ya que no desarrollaron un sistema para representarla, se transmitía de maestro a alumno. También arrojan luz sobre este tema los instrumentos conservados en los museos y la representación en bajorrelieves y pinturas de instrumentos y bailarines, además de lo conservado por tradición oral por los cantores coptos.

Cuadro 12: Resumen de *La música en el antiguo Egipto* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =76.28 % del contenido original).

Confirman en Veracruz caso de influenza en niño de 5 años
El gobierno de Veracruz confirmó este domingo un caso de influenza porcina de la cepa H1N1 en un niño de cinco años originario de el poblado La Gloria. El subdirector de prevención y control de enfermedades de la Secretaría de Salud estatal dijo que el menor de nombre Edgar Hernández Hernández superó el cuadro de infección pulmonar.

Cuadro 13: Resumen de *Confirman en Veracruz caso de influenza en niño de 5 años* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =26.26 % del contenido original).

Efectos de la LSD
Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el entorno en que se use la droga, la pureza de ésta, la personalidad, el estado de ánimo y las expectativas del usuario. Algunos consumidores de LSD experimentan una sensación de euforia, mientras que otros viven la experiencia en clave terrorífica. Cuando la experiencia tienen un tono general desagradable, suele hablarse de mal viaje. Cuando la sustancia se administra por vía oral, los efectos tardan en manifestar se entre 30 minutos y una hora y, según la dosis, pueden durar entre 8 y 10 horas. Entre los efectos fisiológicos recurrentes están los siguientes: contracciones uterinas, fiebre, erizamiento del vello, aumento de la frecuencia cardíaca, transpiración, pupilas dilatadas, insomnio, hiperreflexia y temblores.

Cuadro 14: Resumen de *Efectos de la LSD* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =90.30 % del contenido original).

Introducción a las matemáticas
Cada vez que vas a la tienda, juegas en la computadora o en la consola de video juegos; cuando sigues las incidencias de un juego de béisbol o fútbol americano, cuando llevas el ritmo de una canción, estás utilizando relaciones numéricas y en tu mente realizas una serie de operaciones que tienen que ver con el lenguaje matemático. En este sentido, podemos afirmar que el pensamiento matemático está presente en la mayoría de nuestras actividades, desde las más sencillas hasta las más especializadas. Sin embargo, no siempre estamos conscientes de los conceptos, reglas, modelos, procedimientos y operaciones matemáticas que realizamos mentalmente a diario. A lo largo de esta unidad, mediante la adquisición de distintos conocimientos y la resolución de una serie de problemas y ejercicios, descubriremos cómo representar y formalizar algunas de las operaciones que mencionamos. Los cursos de matemáticas que llevaste con anterioridad, te han familiarizado con la utilización de ciertas operaciones básicas. Con ello podríamos decir que posees los conocimientos básicos para manejar algoritmos elementales. Así que reconocerás diferentes tipos de números como los naturales, los enteros, los fraccionarios (racionales) y los irracionales que son temas de esta unidad. Gracias al conocimiento de los distintos tipos de números construirás y aplicarás modelos matemáticos, los cuales trabajarás con razones y proporciones, así como con series y sucesiones, que te ayudarán a resolver diferentes situaciones de la vida cotidiana. Todos estos aprendizajes te servirán en las siguientes unidades para identificar, resolver, plantear, interpretar y aplicar diferentes procedimientos (algoritmos) con un distinto nivel de complejidad, en variedad de situaciones.

Cuadro 15: Resumen de *Introducción a las matemáticas* (tipo de segmentación: DiSeg, origen del resumen: Humano, τ =84.31 % del contenido original).

Ética de robots
Existe la preocupación de que los robots puedan desplazar o competir con los humanos. Las leyes o reglas que pudieran o debieran ser aplicadas a los robots u otros entes autónomos en cooperación o competencia con humanos han estimulado las investigaciones macroeconómicas de este tipo de competencia, notablemente por Alessandro Acquisti basandose en un trabajo anterior de John von Neumann. Actualmente, no es posible aplicar las Tres leyes de la robótica, dado que los robots no tienen capacidad para comprender su significado. Entender y aplicar las Tres leyes de la robótica, requeriría verdadera inteligencia y consciencia del medio circundante, así como de sí mismo, por parte del robot.

Cuadro 16: Resumen de *Ética de robots* (tipo de segmentación: CoSeg, origen del resumen: Humano, τ =63.52 % del contenido original).

Por qué el embarazo de las elefantas es tan largo
El período de gestación, que se prolonga por casi dos años, es una de esas rarezas de la biología que le permite al feto desarrollar suficientemente su cerebro. Los resultados de este estudio servirán para mejorar los programas de reproducción de elefantes en los zoológicos y podrían también contribuir al desarrollo de un anticonceptivo. Los elefantes son mamíferos muy sociales con un alto grado de inteligencia, similar a la de los homínidos y los delfines. Son, además, los que tienen el período de gestación más largo, que puede extenderse hasta por 680 días. Los elefantes nacen con un nivel avanzado de desarrollo cerebral, que utilizan para alimentarse mediante sus habilidosas trompas. Hasta ahora, los científicos no habían logrado entender en profundidad los procesos biológicos del maratónico embarazo de las elefantas. Pero gracias a los avances de las técnicas de ultrasonido, los veterinarios pudieron utilizar nuevas herramientas.

Cuadro 17: Resumen de *Por qué el embarazo de las elefantas es tan largo* (tipo de segmentación: DiSeg, origen del resumen: Humano, τ =69.85 % del contenido original).

Hallan genes asociados a migraña
Investigadores europeos y australianos indicaron el domingo que habían localizado cuatro nuevos genes asociados con la forma más común de la migraña. Las variantes genéticas fueron detectadas en el genoma de 4800 pacientes de migraña sin aura, la forma que asumen tres de cada cuatro ataques de migraña. Estas estas variantes genéticas no fueron halladas, sin embargo, en el grupo testigo de 7000 personas libres de la enfermedad, dijeron los investigadores. El estudio también confirmó la existencia de otros dos genes de predisposición, en un trío de genes ya identificados en un trabajo anterior. La migraña afecta a aproximadamente una de cada seis mujeres y a uno de cada ocho hombres. Los nuevos genes identificados en este estudio refuerzan el argumento según el cual la disfunción de las moléculas responsables de la transmisión de señales entre las células nerviosas, contribuye a la aparición de la migraña. Además, dos de estos genes refuerzan la hipótesis de un posible papel de las venas. La investigación, publicada en la revista especializada Nature Genetics, fue realizada por un consorcio internacional dedicado a la investigación sobre la genética de la migraña.

Cuadro 18: Resumen de *Hallan genes asociados a migraña* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =79.48 % del contenido original).

Problemas globales
Hoy se reconoce que existen problemas que denominamos globales. Estos problemas se presentan fundamentalmente por la carga de contaminantes liberados hacia la atmósfera terrestre. Por su magnitud y complejidad constituyen un grave problema que requiere medidas muy drásticas para su solución. La composición química de la atmósfera es muy inestable: cambia a través del tiempo y en función de diversas reacciones e interacciones de sus componentes. Hoy sabemos que además de los numerosos gases que la componen, existe una compleja interrelación de los gases. Esta interacción se manifiesta en el hecho de que la radiación solar aporta la energía necesaria para que se realicen las reacciones químicas que modifican la composición de la atmósfera. El diálogo entre la atmósfera y la radiación solar ha sido alterado por el hombre.

Cuadro 19: Resumen de *Problemas globales* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =59.44 % del contenido original).

Descubrimiento de mamut emociona a científicos
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. Al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses.

Cuadro 20: Resumen de *Descubrimiento de mamut emociona a científicos* (tipo de segmentación: CoSeg, origen del resumen: Máquina, τ =43.42 % del contenido original).

La Tundra
El ambiente de la tundra está caracterizado por una sequía prolongada. Las especies más típicas de la flora son los arbustos enanos, líquenes y musgos. Algunas especies, particularmente de aves, sólo pasan el verano en la tundra, época en la que anidan. Existen pocas especies de anfibios y reptiles.

Cuadro 21: Resumen de *La Tundra* (tipo de segmentación: DiSeg, origen del resumen: Máquina, τ =16.17 % del contenido original).