Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 59-66 recibido 30-04-2005; aceptado 01-06-2005

Evaluación de resúmenes automáticos mediante QARLA *

Enrique Amigó, Julio Gonzalo, Víctor Peinado, Anselmo Peñas, Felisa Verdejo
Dept. de Lenguajes y Sistemas Informáticos - UNED
c/Juan del Rosal, 16 - 28040 Madrid - Spain
{enrique, julio, victor, anselmo, felisa}@lsi.uned.es

Resumen: Este artículo muestra la aplicación del marco de evaluación QARLA sobre los resúmenes evaluados en el foro DUC-2004, para las tareas 2 y 5. El marco QARLA permite evaluar de forma automática los sistemas según diferentes aspectos (métricas de similitud) en relación a un conjunto de resúmenes modelo, identificando así los aspectos más deficitarios de las estrategias de resumen existentes. Por otro lado, el marco QARLA permite combinar y meta-evaluar diferentes métricas de similitud, otorgando más peso a los aspectos que caracterizan a los modelos en relación a los resúmenes automáticos.
Palabras clave: Resumen automático, marco de evaluación, QARLA, DUC

Abstract: This article shows an application of the QARLA evaluation framework on DUC-2004 (tasks 2 and 5). The QARLA framework allows to evaluate summaries with regard to different features. Second, it allows to combine and meta-evaluate different similarity metrics, giving more weigh to metrics which characterize models (manual summaries) regarding automatic summaries.
Keywords: automatic summarization, evaluation framework, QARLA, DUC

1. *Introducción*

A lo largo de los últimos años se han evaluado en el foro DUC, sobre diferentes tareas, distintas aproximaciones automáticas al problema del resumen. Existen diversos trabajos orientados a la evaluación automática dentro de la tarea de resumen o traducción automática (Culy y Riehemann, 2003; Coughlin, 2003; Joseph P. Turian y Melamed, 2003; Lin y Hovy, 2003). Sin embargo, en el DUC, la evaluación se ha realizado vía juicios humanos. Por un lado, los jueces responden a un cuestionario en relación a la calidad del resumen. Por otro lado, analizan la cobertura de los resúmenes automáticos en relación a un conjunto de resúmenes modelo, mediante un interfaz y siguiendo ciertas directrices. Esto permite generar un ranking de resúmenes automáticos. Este ranking se puede contrastar con los resultados obtenidos por ciertas métricas automáticas, como ROUGE (Lin y Hovy, 2003), por medio del coeficiente de correlación de Pearson.

Esta estrategia de evaluación tiene la limitación de que una evaluación manual es subjetiva, en cuanto que depende de los jueces y del protocolo de evaluación seguido, por ejemplo, en el caso del DUC, evaluación manual orientada a cobertura mediante el interfaz SEE (http://www.isi.edu/cyl/SEE/).

El modelo QARLA (sección 2), propuesto en (Amigó et al., 2005), aborda este problema:

1. QARLA dispone de un criterio automático (QUEEN) de evaluación de resúmenes sobre un conjunto de modelos y un conjunto de métricas de similitud. El resultado de la aplicación (QUEEN) de métricas de similitud es independiente de la escala de las métricas. Esto implica que podemos comparar los valores de calidad de un mismo resumen según distintos criterios de similitud. (sección 2.1)
2. QARLA dispone de un criterio (KING) de meta-evaluación automática de métricas de similitud, alternativa a la correlación con rankings manuales. Estas medidas se apoyan en criterios objetivos, otorgando más peso a aquellas métricas de similitud que discriminan a los modelos respecto de los resúmenes automáticos. Es decir, otorga más peso a los rasgos característicos de los resúmenes modelo. (sección 2.2)

En este artículo aplicamos el modelo QARLA sobre dos tareas del DUC 2004. La tarea 2 consiste en generar resúmenes cortos (100 palabras) a partir de un conjunto de doc-

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología a través del proyecto HERMES (TIC2000-0335-C03-1).

umentos, y la tarea 5 consiste en generar un resumen corto en base a una cuestión del tipo ¿Quién es...?

La sección 2 describe el marco de evaluación QARLA: QUEEN como estimación de la calidad de un resumen automático, KING como estimación de la calidad de una métrica de similitud, y JACK como estimación de la fiabilidad del proceso de evaluación en relación al conjunto de resúmenes automáticos disponibles. La sección 3 describe las métricas de similitud empleadas en este trabajo, así como el método de selección de las mismas. En la sección 4 se evalúan y analizan los resúmenes automáticos generados por los sistemas participantes en DUC 2004.

2. *Descripción del marco de evaluación QARLA*

El marco QARLA está descrito en detalle en (Amigó et al., 2005). En este apartado hacemos una breve descripción.

El marco QARLA se define a partir de un conjunto A de resúmenes automáticos a evaluar, un conjunto M de resúmenes modelo generados manualmente, y un conjunto X de métricas de similitud entre pares de resúmenes.

2.1. $QUEEN$: Estimación de la calidad de un resumen automático

$QUEEN_{x,M}$ es una medida probabilística que estima la calidad de un resumen automático $a \in A$, a partir de un conjunto de resúmenes modelo M y de una medida de similitud x. La medida $QUEEN$ define la calidad de un resumen automático a como la probabilidad calculada sobre tripletes de resúmenes manuales m, m', m'' de que a esté más próximo a un modelo que los otros dos modelos entre sí. Formalmente:

$$QUEEN_{x,M}(a) \equiv P(x(a,m) \geq x(m',m''))$$

Podemos generalizar esta medida QUEEN para conjuntos de métricas si imponemos que el resumen evaluado cumpla la condición QUEEN para todas las métricas de similitud del conjunto X. Es decir:

$$QUEEN_{X,M}(a) \equiv P(\forall x \in X.x(a,m) \geq x(m',m''))$$

Esta medida cumple algunas propiedades interesantes. El valor QUEEN aumenta a medida que el resumen automático se asemeja a los modelos según las métricas X, pero es a la vez independiente de la escala de las métricas de similitud, dado que cualquier métrica de similitud x' monótona y estrictamente creciente respecto de x devuelve los mismos valores en QUEEN. Es decir, depende únicamente de la topología de las métricas de similitud, por lo que podemos comparar los valores de QUEEN sobre un mismo resumen automático para distintas métricas de similitud.

QUEEN definida como probabilidad tiene un rango entre 0 y 1, en el que un valor de 0.5 implica que el resumen evaluado se asemeja tanto a un modelo como dos modelos entre sí. Un valor de $QUEEN = 0$ implica que el resumen evaluado no se asemeja a ningún modelo más que dos modelos cualesquiera entre si.

Por ejemplo, un valor de $QUEEN = 0{,}5$ para una métrica basada en co-selección de frases, y un valor de $QUEEN = 0{,}1$ sobre una métrica basada en concurrencia de términos clave, implica que el sistema es capaz de comportarse como un resumen manual en cuanto a las frases escogidas, pero no acierta con los conceptos clave que deben estar incluidos en el resumen.

Otra propiedad interesante es que, gracias al cuantificador universal sobre x, métricas redundantes en el conjunto X no sesga el valor de QUEEN.

2.2. $KING$: estimación de la calidad de una métrica de similitud

Un resumen automático puede asemejarse al conjunto de resúmenes modelo en relación a una determinada métrica de similitud. Sin embargo, es posible que dicha métrica de similitud no sea un rasgo característico de los resúmenes modelo (por ejemplo, el número de ocurrencias de la letra 'a' en el resumen). La medida $KING$ estima el peso de una métrica de similitud en la evaluación evitando el coste (y la subjetividad) de un juicio humano directo. Partimos de la hipótesis de que la métrica más pesada es aquella que mejor caracteriza resúmenes manuales en oposición a resúmenes automáticos. Esa métrica debería identificar los resúmenes manuales como más próximos entre sí y más distantes a los resúmenes automáticos. Esta hipótesis la

podemos expresar como:

$$KING_{M,A}(X) \equiv$$
$$P(\forall a \in A.QUEEN_{M,x}(m) > QUEEN_{M,x}(a))$$

que representa la probabilidad de que un modelo $m \in M$ tenga mayor valor $QUEEN_{X,M}$, que cualquier resumen automático $a \in A$. Expresado en términos de ranking, indica la probabilidad de que un modelo ocupe la primera posición respecto a todos los resúmenes automáticos.

Esta medida cumple ciertas propiedades de gran interés. En primer lugar, KING es independiente de la distribución de resúmenes automáticos. Es decir, el cuantificador universal sobre a evita que características redundantes en los sistemas de resumen automático afecten a la medida.

Por otro lado, una métrica de similitud no informativa, por ejemplo, de valor constante o aleatorio, produce por definición un valor $KING = 0$.

Por último, al igual que QUEEN, KING no depende de la distribución de las métricas del conjunto X. Es decir, métricas redundantes en X que evalúan las mismas características de los resúmenes, no sesga el valor de KING.

2.3. JACK: fiabilidad del conjunto de resúmenes automáticos

La medida JACK determina si el conjunto A de resúmenes automáticos es disperso respecto del conjunto M de resúmenes modelo. Esta medida es útil, en primer lugar, para determinar la fiabilidad de los resultados dados por KING, dado que se trata de un valor estadístico dependiente del conjunto A. Por tanto, si A es más heterogéneo, entonces KING es más fiable. En segundo lugar, JACK nos dice si los sistemas de resumen automático siguen estrategias redundantes.

Esta medida JACK se define como:

$$JACK(X, M, A) \equiv$$
$$\equiv P(\exists a, a' \in A.QUEEN(a) > 0 \wedge QUEEN(a') > 0$$
$$\wedge \forall x \in X.x(a, a') \leq x(a, m))$$

es decir, la probabilidad sobre todos los resúmenes modelo m de encontrar un par de resúmenes automáticos a, a' menos cercanos entre sí que a m según todas las métricas. Las condiciones relativas a QUEEN(a) hacen que JACK sea robusto ante resúmenes automáticos dispersos pero muy distantes de los resúmenes modelo.

3. Selección de métricas de similitud

Cada métrica de similitud permite caracterizar los resúmenes sobre la base de criterios diferentes. Nuestro objetivo es seleccionar las métricas que permiten caracterizar mejor los resúmenes manuales y, a la vez, obtener la mayor información posible sobre el comportamiento de los resúmenes automáticos.

En primer lugar, se describen las métricas disponibles como punto de partida (59 en total). A continuación se van a seleccionar 12 métricas de forma que se minimice la información redundante que puede aportarnos cada una de ellas y, a la vez, se maximice la capacidad de caracterizar resúmenes.

Finalmente, se analiza brevemente las características de las métricas seleccionadas. Estas métricas serán las que se utilicen en el siguiente apartado de evaluación de resúmenes del DUC 2004.

3.1. Descripción de las medidas de similitud

Para aplicar el modelo QARLA necesitamos definir el conjunto X de métricas de similitud. Las métricas de similitud empleadas en este trabajo son:

Métricas basadas en ROUGE (R): La métrica de evaluación ROUGE (Lin y Hovy, 2003) estima la calidad, en términos de cobertura de n-gramas, de un resumen automático en relación a un conjunto de resúmenes modelo. Podemos transformar esta métrica de evaluación en una métrica de similitud entre pares de resúmenes si consideramos un único modelo en el cálculo. Existen diferentes tipos de métricas ROUGE como ROUGE-W, ROUGE-L, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 etc. (Lin, 2004). Cada una de estas métricas puede calcularse sobre resúmenes preprocesados de tres formas: con stemming y sin términos de parada (c), con stemming (b) o sin ningún pre-procesado (a). En total, obtenemos 24 métricas de similitud derivadas de ROUGE.

Métricas basadas en ROUGE invertida (Rpre) Las métricas ROUGE están orientadas a cobertura. Por tanto, si invertimos la dirección del cálculo de la similitud, obtendremos métricas orientadas a precisión. Es decir $Rpre(a, b) = R(b, a)$. Generamos así otras 24 métricas de similitud derivadas de ROUGE invertida

TruncatedVectModel (TVM_n): Estas métricas de similitud comparan la distribución en los resúmenes de n de los términos más frecuentes en los documentos originales. El proceso de cálculo consiste en: (1) obtener una lista con los n términos lematizados, ignorando los términos de parada, más frecuentes en los documentos originales; (2) generar un vector con la frecuencia relativa de cada término en el resumen; (3) la medida de similitud se calcula como la inversa de la distancia euclídea entre los vectores asociados a dos resúmenes. Hemos usado nueve variantes de esta medida, correspondientes a $n = 1, 4, 8, 16, 32, 64, 128, 256, 512$.

AveragedSentenceLengthSim (AVLS): Está métrica de similitud compara la longitud promedio de las frases contenidas en dos resúmenes, y puede ser útil para comparar el grado de abstracción de los resúmenes.

GRAMSIM: Está métrica de similitud compara la distribución de tipos de palabras (verbos, adjetivos, etc.) de dos resúmenes. El proceso de cálculo es el siguiente: (1) etiquetamos los resúmenes mediante la herramienta tree-tagger (Schmid, 1994); (2) por cada resumen generamos un vector con la frecuencia de cada una de las posibles etiquetas gramaticales; (3) medimos la distancia euclídea entre los vectores correspondientes a los dos resúmenes. Esta métrica de similitud no está orientada a contenidos, y está relacionada con el estilo lingüístico de los resúmenes.

3.2. Agrupación de métricas de similitud

Dado el conjunto de métricas de similitud descrito anteriormente, tenemos un total de 57 (24+24+9) métricas orientadas a contenidos, y 2 métricas dependientes de aspectos estilísticos. El análisis de 48 aspectos relativos a contenidos puede ser redundante. Sería interesante por tanto poder agrupar conjuntos de métricas de similitud que se comporten de manera similar.

Para ello, calculamos la similitud entre dos conjuntos de métricas como:

$$sim(X, X') \equiv Prob[H_X \leftrightarrow H_{X'}]$$
$$H_X \equiv \forall x \in X x(a, m) \geq x(m', m'')$$

Consideramos que dos conjuntos de métricas son similares si se comportan de la misma forma respecto a la condición $QUEEN$ (predicado H de la fórmula) ante una muestra $\{a, m, m', m''\}$. Es decir, la probabilidad de que ambos conjuntos de métricas discriminen los mismos resúmenes automáticos frente a los mismos pares de modelos.

La tabla de la figura 1 muestra, sobre un número fijado en 10, los grupos de métricas de similitud obtenidos para la tarea 2 de DUC 2003. En cuanto a las métricas de tipo ROUGE, el proceso agrupa las 48 métricas en siete grupos, dependiendo de la longitud de los n-gramas y del tipo de pre-procesado de los textos. En cuanto a las 9 métricas tipo TVM, se distinguen tres grupos: tomando solo el término más frecuente, 4 u 8 términos más frecuentes y otro conjunto con el resto de configuraciones TVM.

3.3. Calidad de las métricas de similitud: valores de KING

Dentro de cada grupo de la figura 1, la métrica de similitud marcada en negrita se corresponde con la métrica de mayor KING. Como puede verse, en los conjuntos basados en n-gramas con stemming, las métricas de mayor KING son de tipo R, basadas en cobertura (grupos 2,3 y 4), mientras que en los conjuntos de métricas sobre n-gramas sin stemming (grupos 5,6 y 7) las métricas de mayor KING son de tipo Rpre, basadas en precisión. En cuanto a métricas de tipo TVM, las de mayor KING son siempre aquellas que se basan en un número mayor de términos frecuentes (grupos 8,9 y 10).

Finalmente, tomamos como métricas representativas, la métrica de mayor $KING$ en cada grupo, en total 10 métricas orientadas a contenidos.

La figura 2 muestra el valor KING de las distintas métricas de similitud seleccionadas, es decir, la capacidad de las métricas de caracterizar a los resúmenes modelo frente a los resúmenes automáticos.

cluster ID	DESCRIPTION	SIMILARITY METRICS
Cluster 1	ROUGE based metrics	R-S.b R-SU.b R-S.a R-SU.a R-1.a R-1.b R-L.b R-L.a R-W-1.2.b R-W-1.2.a R-W-1.2.c R-S.c R-SU.c R-1.c R-L.c **Rpre-W-1.2.b** Rpre-W-1.2.a Rpre-W-1.2.c Rpre-L.c Rpre-1.c Rpre-S.c Rpre-SU.c Rpre-1.a Rpre-S.a Rpre-SU.a Rpre-1.b Rpre-S.b Rpre-SU.b Rpre-L.b Rpre-L.a
Cluster 2	ROUGE (Stemmed and non-stopwords 2-grams)	**R-2.c** Rpre-2.c
Cluster 3	ROUGE (Stemmed and non-stopwords 3-grams)	Rpre-3.c **R-3.c**
Cluster 4	ROUGE (Stemmed and non-stopwords 4-grams)	Rpre-4.c **R-4.c**
Cluster 5	ROUGE (Non-stemmed 2-grams)	R-2.b R-2.a **Rpre-2.b** Rpre-2.a
Cluster 6	ROUGE (Non-stemmed 3-grams)	R-3.b R-3.a **Rpre-3.b** Rpre-3.a
Cluster 7	ROUGE (Non-stemmed 4-grams)	Rpre-4.a **Rpre-4.b** R-4.b R-4.a
Cluster 8	TMV.Most salient term	**TVM.1**
Cluster 9	TMV.4 and 8 salient terms	TVM.4 **TVM.8**
Cluster 10	TMV.>8 Salient terms	TVM.16 TVM.32 TVM.64 TVM.128 TVM.256 **TVM.512**

Figura 1: Agrupaciones de métricas de similitud

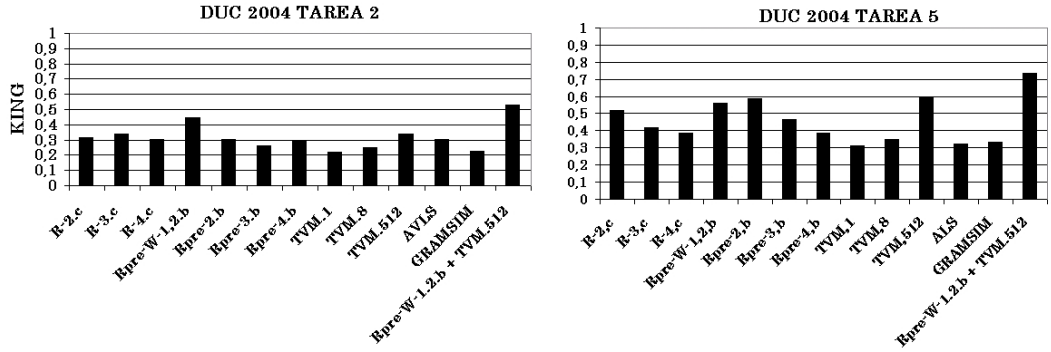

Figura 2: Calidad de las métricas de similitud

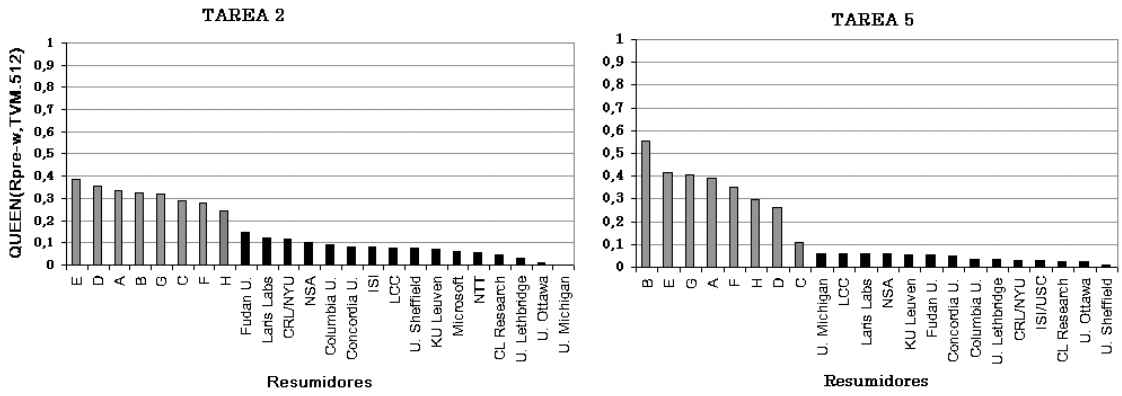

Figura 3: Calidad de los resúmenes automáticos

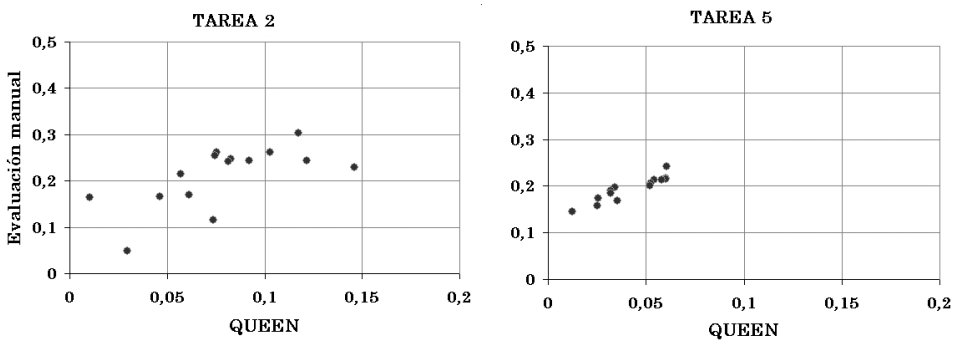

Figura 4: Correlación entre evaluación en DUC y QARLA

En cuanto a las métricas aisladas, Rpre-W obtiene un alto KING en ambas tareas. Esta métrica de similitud se basa en precisión de secuencias no contiguas de términos en los resúmenes (ROUGE-W-1.2). Para las métricas de tipo TVM a medida que consideramos más términos frecuentes, la métrica asciende en valores KING, es decir, caracteriza mejor a los resúmenes modelo.

La última columna representa la combinación de mayor KING de entre las posibles combinaciones de métricas de similitud. En ambas tareas esta combinación es $Rpre - W$ y TMV,512, y supera en valor KING a cualquiera de las métricas individuales. Este hecho muestra que la capacidad de caracterización de los modelos puede aumentar por medio de la estrategia de combinación de métricas que ofrece el marco QARLA.

4. Evaluación de resúmenes automáticos en DUC 2004

En esta sección, a partir de las métricas seleccionadas en la sección anterior, se abordan tres cuestiones: qué resúmenes automáticos tienen un contenido más similar a los resúmenes modelo, si siguen aproximaciones similares, y qué características tienen.

4.1. Ranking de resúmenes automáticos

La combinación de mayor KING en ambas tareas es Rpre-W y TVM.512. La figura 3 representa el ranking de resúmenes según dicha combinación, de acuerdo a los valores de $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ obtenidos. Los resúmenes modelo obtienen los mayores valores de QUEEN en ambas tareas con una diferencia significativa respecto a los resúmenes automáticos.

El ranking manual generado en DUC representa un criterio concreto de evaluación, mientras que el modelo QARLA da más peso a aquellos aspectos que son más discriminativos entre modelos y resúmenes automáticos. Los resultados por tanto no tienen por qué coincidir necesariamente. Sin embargo, sí debería de existir cierta correlación. La figura 4 muestra la correlación existente entre ambas estrategias de evaluación, es decir valores de $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ (eje horizontal) frente a los valores de la evaluación manual (eje vertical). Esta correlación es más patente en la tarea 5 (orientada a pregunta) donde los contenidos de los resúmenes son más precisos. Por otro lado, en la tarea 2 (resúmenes genéricos) puede verse que los sistemas con mayor puntuación en QUEEN también tienen una buena valoració0n por parte de los jueces humanos en DUC.

4.2. Heterogeneidad de los resúmenes automáticos

Mediante la medida JACK podemos analizar la heterogeneidad de los resúmenes automáticos. La figura 5 muestra como crece el valor JACK, tomando como métricas la combinación de máximo KING, Rpre-W y TVM.512, a medida que aumentamos el número de resúmenes automáticos. En ambas tareas los resultados tienden a estabilizarse a partir de cierto número de resúmenes, lo que indica cierta redundancia en las aproximaciones seguidas por los sistemas participantes.

Por otro lado, en la tarea 2 (resúmenes genéricos) el conjunto de resúmenes automáticos es más heterogéneo que en la tarea 5 (resumen orientado a pregunta). Es decir, se han empleado estrategias más diversas en la generación automática de resúmenes genéricos, mientras que en la tarea 5, probablemente, la pregunta haya centrado los sistemas en un mismo tipo de estrategia.

4.3. Caracterización de resúmenes automáticos

El modelo QARLA permite evaluar los resúmenes automáticos en relación a diferentes métricas de similitud. Si la métrica o el conjunto de métricas aplicado tiene un mayor KING, entonces representa un rasgo que caracteriza a los resúmenes manuales frente a resúmenes automáticos, considerándose más fiable como criterio de evaluación.

Sin embargo además de evaluar un sistema de resumen, QARLA permite analizar las propiedades de los resúmenes evaluados. Dado que QUEEN es independiente de la escala de las métricas, esto se puede hacer comparando los valores de QUEEN obtenidos por los sistemas en relación a distintas métricas de similitud.

Las figuras 6 y 7 muestran la calidad de los resúmenes automáticos en relación a las 12 métricas de similitud seleccionadas, extraídas mediante el proceso de agrupamiento descrito. Los valores de QUEEN más elevados aparecen marcados en negrita.

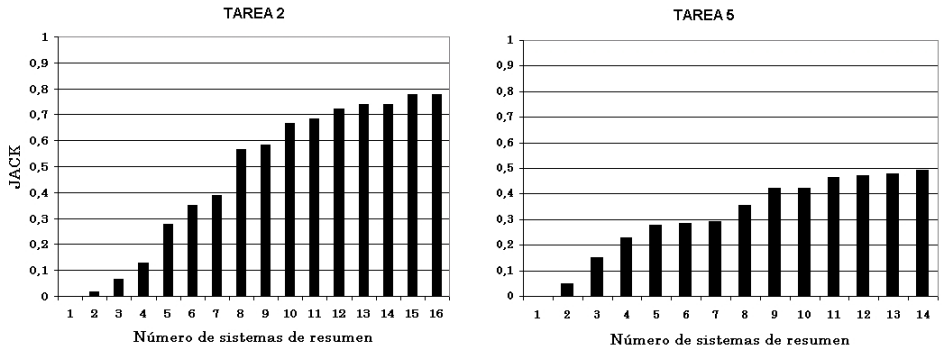

Figura 5: JACK vs. numero de resúmenes automáticos

	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM
NSA	**0,48**	**0,62**	**0,85**	0,19	**0,45**	**0,5**	**0,7**	**0,52**	**0,46**	0,35	0,28	**0,53**
U. Sheffield	**0,48**	**0,64**	**0,85**	0,16	0,4	**0,48**	**0,68**	**0,48**	0,33	0,26	0,28	**0,51**
Columbia U.	0,39	**0,6**	**0,85**	0,21	0,39	**0,49**	**0,69**	**0,52**	0,43	0,28	0,2	**0,55**
KU Leuven	0,39	**0,6**	**0,83**	0,2	0,4	**0,49**	**0,65**	**0,54**	0,44	0,33	0,19	0,33
ISI	0,4	**0,59**	**0,85**	0,15	0,37	**0,46**	**0,67**	**0,53**	0,4	0,29	0,27	**0,47**
CRL/NYU	**0,45**	**0,65**	**0,86**	0,41	**0,52**	**0,6**	**0,75**	**0,51**	0,39	0,19	0,2	0,38
Concordia U.	0,37	**0,58**	**0,84**	0,14	0,3	0,37	**0,62**	**0,58**	**0,49**	0,33	0,32	**0,51**
LCC	**0,48**	**0,66**	**0,85**	0,14	0,43	**0,49**	**0,67**	**0,46**	0,39	0,31	0,17	**0,48**
Laris Labs	0,43	**0,63**	**0,85**	0,22	0,38	**0,45**	**0,66**	**0,55**	**0,48**	0,35	0,32	**0,52**
Fudan U.	0,4	**0,59**	**0,84**	**0,26**	0,36	**0,44**	**0,67**	**0,59**	**0,53**	0,4	0,43	**0,55**
CL Research	0,41	**0,63**	**0,85**	0,21	0,41	**0,51**	**0,71**	0,44	0,3	0,15	0,1	0,34
NTT	0,23	**0,53**	**0,84**	0,22	0,25	0,35	**0,61**	**0,52**	0,36	0,16	0,31	0,31
U. Lethbridge	0,21	**0,52**	**0,83**	0,06	0,2	0,32	**0,61**	**0,51**	0,44	0,21	0,13	0,35
U. Michigan	0,36	**0,61**	**0,85**	0	0,3	**0,44**	**0,66**	**0,52**	0,4	0,3	0	0,3
Microsoft	0,25	**0,54**	**0,83**	0,15	0,24	0,34	**0,62**	**0,52**	0,36	0,2	0,44	0,39
U. Ottawa	0,05	0,46	**0,82**	0,02	0,03	0,21	**0,57**	0,38	0,24	0,21	**0,51**	0,13
Average	0,361	0,591	0,843	**0,171**	0,339	0,434	0,659	0,511	0,403	**0,27**	**0,26**	0,4156
	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM

Figura 6: Calidad de los resúmenes automáticos en la tarea 2

Los sistemas de resumen (eje vertical) están ordenados según el ranking generado manualmente en DUC. Como puede verse, en ambas tareas los primeros sistemas obtienen buena puntuación según la mayoría de los rasgos orientados a contenidos (métricas R, Rpre y TVM).

A partir de estos datos, si nos fijamos en los valores promedio (última fila) de QUEEN sobre distintas métricas de similitud, podemos extraer algunas conclusiones:

- Los resúmenes automáticos generados en el DUC no coindicen en longitud de frases con los resúmenes modelo. La puntuación promedio sobre la métrica *AVLS* (0.26 y 0.22) es bastante reducida en relación al valor obtenido por el resto de las métricas.
- Los resúmenes automáticos son capaces de identificar los términos más relevantes (valores de QUEEN más altos en TVM.1), mientras que no identifican elementos menos frecuentes de los documentos originales pero que sí son comunes en los resúmenes modelo (valores de QUEEN más bajos en TMV.512).
- El estilo lingüístico de los resúmenes automáticos es especialmente diferente al de los modelos en la tarea 5 (resumen orientado a pregunta) frente a la tarea 2 (resumen genérico). Presumiblemente los resúmenes manuales orientados a pregunta tienen unos rasgos de estilo más definidos.
- Los resúmenes automáticos se asemejan a los modelos en secuencias de palabras largas (valores altos en R-3,R-4) en la misma medida que los modelos entre sí, especialmente en los resúmenes de tipo genérico (tarea 2).

5. Conclusiones

El marco de evaluación QARLA ha permitido nos ha permitido, por un lado, identificar métricas de evaluación fiables y comparar resúmenes automáticos entre sí. Los resultados muestran que existe una alta correlación entre la evaluación resultante de QARLA y la evaluación realizada por jueces humanos.

	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM
LCC	**0,25**	**0,44**	**0,67**	0,16	0,19	**0,32**	**0,45**	**0,47**	0,29	0,17	0,29	0,32
KU Leuven	0,23	0,4	0,65	0,15	0,21	0,3	0,43	**0,57**	0,4	**0,26**	0,36	0,21
U. Michigan	0,2	0,36	0,64	0,2	0,21	0,28	0,41	**0,54**	0,33	0,18	0,11	0,26
NSA	0,29	0,41	0,65	0,09	0,2	0,26	0,39	**0,54**	0,32	0,23	0,38	0,32
Laris Labs	0,24	0,42	0,66	0,13	0,17	0,27	0,41	**0,57**	0,44	0,22	0,35	**0,41**
Columbia U.	0,29	**0,44**	0,66	0,13	**0,23**	0,29	0,41	0,38	0,23	0,17	0,38	0,31
Concordia U.	**0,25**	**0,44**	0,66	0,14	0,18	0,28	0,42	**0,45**	0,29	0,18	0,34	0,31
U. Lethbridge	0,23	0,41	0,65	0,14	0,18	0,26	0,41	**0,46**	0,25	0,15	0,21	0,35
Fudan U.	0,22	0,38	0,64	0,12	0,17	0,24	0,37	0,41	0,27	0,2	0,42	0,33
U. Sheffield	0,23	0,39	0,65	0,03	0,16	0,23	0,39	0,5	0,31	0,18	0,17	0,37
CRL/NYU	0,22	0,4	0,63	**0,25**	**0,23**	0,32	0,42	**0,51**	0,29	0,08	0,26	0,18
U. Ottawa	0,13	0,3	0,62	0,13	0,09	0,16	0,3	0,24	0,11	0,08	**0,39**	0,19
CL Research	0,19	0,37	0,64	0,12	0,15	0,23	0,38	0,37	0,22	0,09	0,28	0,27
ISI/USC	0,18	0,31	0,62	0,07	0,12	0,17	0,33	**0,59**	**0,45**	**0,26**	0,37	0,31
Average	**0,13**	0,209	0,286	0,3229	**0,15**	0,227	0,3	0,336	0,37	**0,301**	**0,22**	**0,1797**
	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM

Figura 7: Calidad de los resúmenes automáticos en la tarea 5

Por otro lado, QARLA ha permitido extraer conclusiones acerca de qué aspectos están mejor o peor cubiertos por las aproximaciones de los sistemas existentes al problema del resumen automático, por medio de la aplicación de distintas métricas de similitud. Aplicado sobre los datos del DUC-2004, los resultados han resultado coherentes con las características reales de los sistemas, por lo general, extractivos y basados en estrategias superficiales. Los resultados son también coherentes con las diferencias entre la tarea de generación de un resumen genérico y un resumen orientado a pregunta.

Bibliografía

Amigó, E., J. Gonzalo, A. Peñas, y F. Verdejo. 2005. QARLA: a Framework for the Evaluation of Text Summarization Systems. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Coughlin, Deborah. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. En *In Proceedings of MT Summit IX*, New Orleans,LA.

Culy, Christopher y Susanne Riehemann. 2003. The Limits of N-Gram Translation Evaluation Metrics. En *Proceedings of MT Summit IX*, New Orleans,LA.

Joseph P. Turian, Luke Shen y I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. En *In Proceedings of MT Summit IX*, New Orleans,LA.

Lin, C. y E. H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-ocurrence Statistics. En *Proceeding of 2003 Language Technology Conference (HLT-NAACL 2003)*.

Lin, Chin-Yew. 2004. Rouge: A Package for Automatic Evaluation of Summaries. En Marie-Francine Moens y Stan Szpakowicz, editores, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, páginas 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lin, Chin-Yew y Eduard Hovy. 2003. The Potential and Limitations of Automatic Sentence Extraction for Summarization. En Dragomir Radev y Simone Teufel, editores, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, Edmonton, Alberta, Canada, May 31 - June 1. Association for Computational Linguistics.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. En *International Conference on New Methods in Language Processing*.