Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 277-285 recibido 1-05-2009; aceptado 5-06-2009

Comparación de Enfoques para Evaluar la Validación de Respuestas*

Comparing Approaches for Evaluating Question Answering Validation

Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
C/ Juan del Rosal, 16, 28040 Madrid
{alvarory, anselmo, felisa}@lsi.uned.es

Resumen: La Validación de Respuestas ha sido vista recientemente como un problema de clasificación donde se puede introducir aprendizaje automático con el propósito de mejorar los resultados de los sistemas de Búsqueda de Respuestas. La naturaleza no balanceada de las colecciones ha llevado al uso de medidas de evaluación basadas en precisión y cobertura. Sin embargo, para este tipo de evaluaciones se suele usar más análisis ROC (Relative Operating Characteristic). En este artículo se comparan ambos enfoques de acuerdo a sus fundamentos, su estabilidad en función del tamaño de las colecciones, su poder de discriminación y su adecuación a las particularidades de la Validación de Respuestas.
Palabras clave: Búsqueda de Respuestas, Validación de Respuestas, Evaluación

Abstract: The Validation of Answers has been seen recently as a classification problem able to introduce Machine Learning for improving Question Answering results. The unbalanced nature of collections has led to the use of measures based on precision and recall. However, Relative Operating Characteristic (ROC) analysis is preferred sometimes in similar classification tasks. In this article we compare both approaches according to their rationale, their stability with respect to the size of the collection, their discriminative power and their adequacy to the particularities of the Answer Validation task.
Keywords: Question Answering, Answer Validation, Evaluation

1. *Introducción*

Los sistemas tradicionales de Búsqueda de Respuestas (en inglés Question Answering, QA) suelen tener una arquitectura en cadena (Hovy et al., 2001; Moldovan et al., 2000; Prager et al., 2000) que genera una alta dependencia entre módulos y que es muy sensible a la propagación de errores. Por ejemplo, un sistema de QA que haga uso de un módulo de recuperación de documentos y otro de extracción de respuestas en el que ambos tengan una precisión del 80 % obtendría como mucho un 64 % de precisión.

Una manera de superar los límites del procesamiento en cadena sería introducir más inteligencia. La Validación de Respuestas (en inglés Answer Validation, AV) contribuye a este tipo de mejora (Harabagiu y Hickl., 2006). Un sistema de AV recibe una *Pregunta* y una *Respuesta* y devuelve un valor indicando si la *Respuesta* es o no correcta y en qué grado.

La decisión acerca de la corrección de las respuestas es una tarea de clasificación donde las respuestas han de ser clasificadas como correctas o incorrectas. Esta tarea de clasificación binaria tiene la matriz de confusión mostrada en el Cuadro 1 (las fórmulas del resto del artículo serán dadas en términos de dicha matriz).

En aprendizaje automático se usa principalmente *accuracy* (1) para evaluar la clasificación binaria. Al aplicarlo a AV *accuracy* premia en la misma proporción la detección de respuestas correctas e incorrectas. Sin embargo la tarea de AV presenta una particu-

* Este trabajo ha sido subvencionado parcialmente por el proyecto QEAVis-Catiex (TIN2007-67581-C02-01) del Ministerio de Ciencia e Innovación, el proyecto europeo Trebe CLEF (ICT-1-4-1 215231), la Consejería de Educación de la Comunidad de Madrid y el Fondo Social Europeo (F.S.E.)

		True class		
		Correct	**Incorrect**	
Hypothesized class	**Correct**	**T**rue **P**ositives	**F**alse **P**ositives	**C**
	Incorrect	**F**alse **N**egatives	**T**rue **N**egatives	**I**
Column totals:		**P**	**N**	

Cuadro 1: Matriz de confusión

laridad que debe ser tenida en cuenta: dado que los sistemas actuales de QA producen una cantidad diferente de respuestas correctas que de respuestas incorrectas, la evaluación de AV debe de tenerlo en cuenta (Peñas et al., 2008) y considerar la naturaleza no balanceada de las colecciones. Por tanto, haciendo uso de *accuracy* un sistema de AV no informado podría obtener valores demasiado altos.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

En este tipo de evaluaciones se suelen utilizar dos enfoques distintos: *precisión-cobertura* y análisis ROC (Receiver Operating Characteristic). El objetivo de este trabajo es determinar cuál es el enfoque más apropiado para evaluar AV.

El resto de este artículo está estructurado de la siguiente manera: en la Sección 2 se explican los fundamentos del análisis ROC mientras que el enfoque *precisión-cobertura* de explica en la Sección 3. En la Sección 4 se comparan ambos enfoques de acuerdo a sus fundamentos, su estabilidad en función del tamaño de las colecciones, su poder de discriminación y su adecuación a las particularidades de la AV. Finalmente se dan algunas conclusiones en la Sección 5.

2. Análisis ROC

El análisis ROC (Receiver Operating Characteristic) es una metodología que ha comenzado a ser utilizada para evaluar clasificadores en Inteligencia Artificial (Beck y Shultz, 1986; Friedman y Wyatt, 1997). En un problema de clasificación binaria el espacio ROC es una representación en dos dimensiones con la proporción de ejemplos positivos detectados (en inglés true positive rate, *tp rate* (2)) en el eje Y y la proporción de falsos positivos (en inglés false positive rate, *fp rate* (3)) en el eje X. De este modo, cada matriz de confusión genera un punto en el espacio ROC.

$$tp\ rate = \frac{TP}{TP + FN} \quad (2)$$

$$fp\ rate = \frac{FP}{FP + TN} \quad (3)$$

Son varios los puntos a destacar en el espacio ROC (ver Figura 1). El punto (0,0) representa a un clasificador que predice todas las instancias como negativas mientras que el punto (1,1) corresponde a un clasificador que predice todas las instancias como positivas. Por otro lado, el punto (0,1) representa una clasificación perfecta mientras que el punto (1,0) corresponde al clasificador que falla en todas sus predicciones. Por último, la línea diagonal $y{=}x$ representa la estrategia de clasificar aleatoriamente.

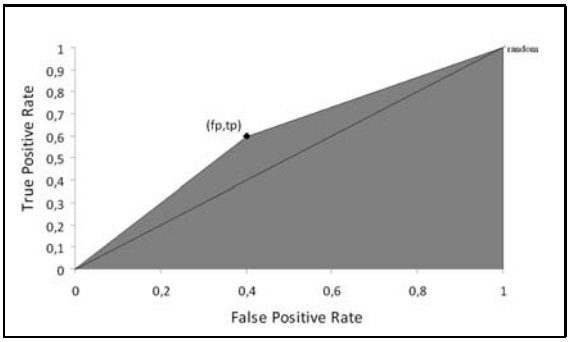

Figura 1: Área bajo la curva (AUC) del punto (0.4 , 0.6)

Además de esta representación gráfica se puede obtener también un valor escalar haciendo uso del concepto de curva ROC. La curva ROC de un clasificador está formada por una secuencia de puntos ROC del clasificador, incluyendo los punto (0,0) y (1,1), conectados por segmentos. Las curvas ROC presentan la ventaja de no verse afectadas por cambios en la distribución de las clases. El método usado para generar la secuencia

de puntos depende del clasificador. En caso de que no haya ningún método para generar la secuencia de puntos, un clasificador puede formar una curva ROC conectando el único punto ROC que genera junto con los puntos (0,0) y (1,1) (Drummond y Holte, 2004). Dada una curva ROC se puede utilizar el área bajo la curva (en inglés Area Under the ROC Curve, AUC) (Bradley, 1997; Hanley y McNeil, 1982) como indicador del rendimiento del clasificador. Sin embargo el cálculo tradicional de AUC no permite variar la importancia que se le da a *tp rate* o *fp rate*. Sería interesante poder contar con una modificación de AUC que permitiese dicho ajuste.

Dado que AUC es una porción del área del cuadrado unidad, su valor está comprendido entre 0 y 1. En la Figura 1 se puede ver un ejemplo. Cuando se utiliza la estrategia de clasificar aleatoriamente se obtiene un valor de AUC de 0.5, por lo que siempre es deseable obtener un valor superior a 0.5 ya que un valor inferior indica tener un rendimiento peor que el de la clasificación aleatoria.

3. Enfoque Precisión-Cobertura

En AV una respuesta debería de ser validada si hay suficientes evidencias acerca de su corrección. Bajo esta perspectiva, la tarea de AV consiste en detectar respuestas correctas y asegurarse de que sólo las respuestas correctas son validadas. De acuerdo con esta visión, al evaluar AV hay que cuantificar la habilidad de un sistema para detectar si hay suficientes evidencias como para validar una respuesta. Es por ello que se propone el uso de *precisión* y *cobertura* sobre las respuestas correctas:

- *Precisión*: proporción de respuestas validadas por el sistema y que realmente son correctas (4).
- *Cobertura*: proporción de las respuestas correctas que han sido detectadas por el sistema (5).

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$cobertura = \frac{TP}{TP + FN} \quad (5)$$

Precisión mide lo bueno que es un sistema validando respuestas mientras que *cobertura* evalúa la habilidad de un sistema para detectar las respuestas correctas que hay.

Estas dos medidas se pueden representar gráficamente con *cobertura* en el eje X y *precisión* en el eje Y. En esta representación gráfica el punto (1,1) corresponde al mejor rendimiento posible mientras que el punto (0,0) corresponde al peor posible. La validación aleatoria está representada por una línea paralela al eje X. El valor de *precisión* de dicha línea depende de la naturaleza de las colecciones utilizadas, siendo el cociente entre el número de respuestas correctas y el número total de respuestas. La única excepción es el caso de un sistema que no valida ninguna respuesta, sistema que obtiene el valor 0 tanto de *precisión* como de *cobertura*.

La medida más utilizada para combinar *precisión* y *cobertura* es la *medida-F* (6). En esta medida β representa un valor positivo usado para dar más importancia a la *cobertura* o a la *precisión*. Si $\beta < 1$ se da más importancia a la *precisión* mientras que si $\beta > 1$ cobra más importancia la *cobertura*. Cuando se usa el valor $\beta = 1$ se evalúan ambos aspectos de un sistema dándole a cada uno la misma importancia. Al ser el valor $\beta = 1$ el más comúnmente utilizado, será el que usemos para nuestro estudio.

$$F = \frac{(\beta^2 + 1) * cobertura * precision}{\beta^2 * cobertura + precision} \quad (6)$$

4. Comparación de ROC con Precisión-Cobertura

Una crítica que reciben *precisión* y *cobertura* es que sus valores cambian al cambiar la distribución en clases de las colecciones aunque no haya cambios significativos en el rendimiento de los sistemas. Este es un argumento utilizado en algunas ocasiones para decantarse por el uso de análisis ROC cuando las colecciones son no balanceadas (Provost y Fawcett, 2001).

Hay que destacar que *cobertura* y *tp rate* son diferentes nombres del mismo concepto. Por tanto, la principal diferencia entre los dos enfoques es el uso de *precisión* y *fp rate*.

En las siguientes secciones se procede a comparar los enfoques expuestos anteriormente de acuerdo a sus fundamentos, su estabilidad en función del tamaño de las colecciones, su poder de discriminación y su adecuación a la tarea de AV. Antes se informa sobre los datos que han sido utilizados para realizar el estudio.

4.1. Datos para el Análisis

Para realizar la comparación se hizo uso de los recursos públicos de la tarea Answer Validation Exercise[1] (AVE) 2008 (Rodrigo, Peñas, y Verdejo, 2009) celebrada en el Cross Language Evaluation Forum[2] (CLEF).

Las colecciones del AVE fueron creadas a partir de la salida real de sistemas de QA. Estas colecciones contienen conjuntos de pares {*Respuesta, Texto Soporte*} que están agrupados por *Pregunta*. Los sistemas participantes deben de considerar cada *Pregunta* y clasificar cada uno de los pares como correcto o incorrecto. El número de pares {*Respuesta, Texto Soporte*} por idioma así como la distribución en correctos e incorrectos depende de la salida de los sistemas participantes en la tarea de QA. El Cuadro 2, tomado de Rodrigo, Peñas, y Verdejo (2009), muestra el número de preguntas y respuestas por idioma en las colecciones de evaluación del AVE 2008 así como el número de runs participantes.

Se decidió realizar el estudio usando la colección de inglés ya que tanto su tamaño como el número de runs participantes con el que contó dicho idioma eran suficientes para los propósitos del estudio. Además de los runs participantes se han utilizado dos sistemas adicionales: un sistema que valida todas las respuestas (100 % YES) y otro que valida aleatoriamente la mitad de las respuestas (50 % YES).

4.2. Interpretación Gráfica

Como se indicó en las Secciones 2 y 3, ambos enfoques permiten representar el rendimiento de un sistema en un gráfico en dos dimensiones. Dichas representaciones gráficas permiten realizar una visión general del rendimiento de sistemas. Sin embargo hay que tener cuidado a la hora de sacar conclusiones de una representació gráfica ya que si se quiere ser realmente objetivo es mejor utilizar una medida numérica.

El Cuadro 3 muestra los resultados (ordenados por AUC) de los runs utilizados en el estudio mientras que las Figuras 2 y 3 muestran la representación gráfica para el espacio ROC y para *precisión-cobertura* respectivamente. Dado que *tp rate* es igual que *cobertura*, se ha modificado la representación del espacio ROC (Figura 2) con el propósito de asemejarla a la de *cobertura-precisión*. De este modo se representa *tp rate* en el eje X (al igual que cobertura en el gráfico *cobertura-precisión*) y *fp rate* en el eje Y. Además se ha modificado el gráfico ROC (Figura 2) para representar en la esquina inferior izquierda el valor 1 de *fp rate* mientras que en la esquina superior izquierda se representa el valor 0. Se ha hecho esta transformación con el propósito de representar el mejor rendimiento posible en análisis ROC (*fp rate* = 0 y *tp rate* = 1) en la esquina superior derecha al igual que en el gráfico *cobertura-precisión*.

Si se observa el gráfico ROC (Figura 2), da la sensación de que los mejores sistemas están mucho más cerca del punto óptimo que en el gráfico *cobertura-precisión* (Figura 3). Dado que el valor en el eje X es el mismo en ambos gráficos, esta sensación se produce debido a que según *fp rate* los sistemas tienen mejor rendimiento que según *precisión*. Sin embargo tener un valor bajo de *fp rate* no siempre implica tener un buen rendimiento en términos de un sistema de AV como se discute en la Sección 4.5. Por tanto la representación gráfica en el espacio ROC puede ser demasiado optimista, dando una imagen errónea del rendimiento real de un sistema de AV.

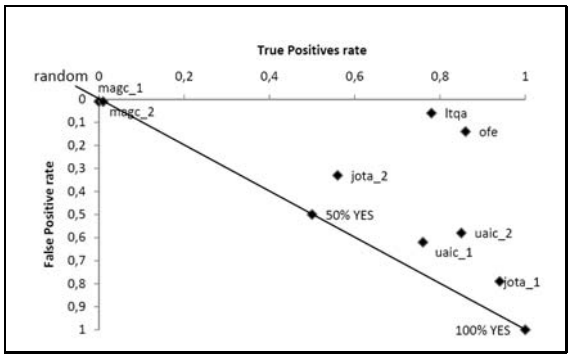

Figura 2: Representación en el espacio ROC

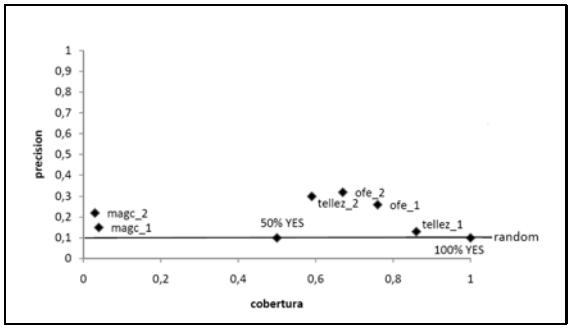

Figura 3: Representación en el espacio Cobertura-Precisión

[1] http://nlp.uned.es/clef-qa/ave/
[2] http://www.clef-campaign.org/

	Alemán	Inglés	Español	Francés	Búlgaro	Holandés	Portugués	Rumano	Vasco
Preguntas	119	160	136	108	27	128	149	119	104
Respuestas	965	1019	1507	178	21	221	955	458	522
CORRECTAS	111	79	153	52	12	44	208	52	39
INCORRECTAS	854	940	1354	126	9	177	747	406	483
Runs participantes	3	8	6	5	0	0	0	2	0

Cuadro 2: Número de preguntas, respuestas y runs participantes por idioma en las colecciones de evaluación del AVE 2008

Sistema	fp rate	cobertura/ tp rate	precisión	AUC	F
ltqa	0.06	0.78	0.54	0.86	0.64
ofe	0.14	0.86	0.35	0.86	0.49
uaic_2	0.58	0.85	0.11	0.64	0.19
jota_2	0.33	0.56	0.13	0.62	0.21
jota_1	0.79	0.94	0.09	0.58	0.17
uaic_1	0.62	0.76	0.09	0.57	0.17
magc_2	0.01	0.01	0.17	0.50	0.02
magc_1	0.01	0	0	0.50	0
100 % YES	1	1	0.08	0.50	0.14
50 % YES	0.5	0.5	0.08	0.50	0.13

Cuadro 3: Valores de fp rate, cobertura (tp rate), precisión, AUC y medida-F para los runs del estudio

4.3. Poder de Discriminación y Estabilidad

Una característica importante en una medida de evaluación es su estabilidad, que es el error asociado a la conclusión *el sistema A es mejor que el sistema B* (Buckley y Voorhees, 2000). Otra característica a tener en cuenta es el número de veces que dos sistemas son considerados equivalentes. Esta característica se denomina poder de discriminación y con que más discriminatoria es una medida, menos empates habrá entre sistemas y menor será la diferencia necesaria para concluir qué sistema es mejor (Buckley y Voorhees, 2000).

Se ha realizado un estudio acerca de la estabilidad y el poder de discriminación de la *medida-F* y de *AUC* basándose en el método descrito en Buckley y Voorhees (2000). Una vez adaptado a nuestro estudio funciona de la siguiente manera: sea S un conjunto de runs y sean x e y un par de runs de S. Sea Q la colección de evaluación entera. Sea f el valor de fuzziness, tal que si la diferencia entre dos sistemas es menor que f se considera que el rendimiento de los sistemas es equivalente.

Con el fin de incrementar el conjunto de datos empleado se partió aleatoriamente la colección de evaluación original en subcolecciones de un determinado tamaño c, realizando 200 ejecuciones distintas. De este modo, $\mathrm{M}(x,Q_{trial})$ representa la media de la medida M para el run x sobre las subcolecciones de Q_{trial}.

Para aplicar el método de Buckley y Voorhees (2000) se hizo uso del algoritmo de la Figura 4 para obtener los datos necesarios para calcular la tasa de error (7) (que se utiliza para medir la estabilidad) y el porcentaje de empates (8) (que se usa para estudiar el poder de discriminación). La intuición usada para calcular la tasa de error es la siguiente: se asume que para una determinada medida la decisión correcta sobre si el run x es mejor que el run y ocurre cuando hay más casos en los que el run x es mejor que el run y. Por tanto el número de veces que y es mejor que x se considera el número de ocasiones en las cuáles el test es erróneo.

$n = |Q|$ / c;
for each trial from 1 to 200
  Q_{trial} = select at random n different subcols of size c from Q;
  for each pair of runs x,y ϵ S
    $margin = f$ * max (M(x,Q_{trial}),M(y,Q_{trial}));
    if(|M(x,Q_{trial}) - M(y,Q_{trial})| < $margin$)
      $EQ_M(x,y)$++;
    else if(|M(x,Q_{trial}) > M(y,Q_{trial})|)
      $GT_M(x,y)$++;
    else
      $GT_M(y,x)$++;

Figura 4: Algoritmo para calcular $EQ_M(x,y)$, $GT_M(x,y)$ y $GT_M(y,x)$

$$Tasa\ de\ error_M = \frac{\sum_{x,y \epsilon S} min(GT_M(x,y), GT_M(y,x))}{\sum_{x,y \epsilon S}(GT_M(x,y) + GT_M(y,x) + EQ_M(x,y))} \quad (7)$$

$$\%\ Empates_M = \frac{\sum_{x,y \epsilon S} EQ_M(x,y)}{\sum_{x,y \epsilon S}(GT_M(x,y) + GT_M(y,x) + EQ_M(x,y))} \quad (8)$$

Aumentando el valor de fuzziness se consigue disminuir la tasa de error pero también disminuir el poder de discriminación. Dado que usar un único valor fijo de fuzziness no sería lo suficientemente informativo y de acuerdo con Sakai (2007), se decidió variar el valor de fuzziness de 0.01 a 0.10 y dibujar las curvas *porcentaje de empates - tasa de error* para comparar las medidas propuestas. La Figura 5 muestra dichas curvas para la *medida-F* y AUC siendo $c = 150$.

En la Figura 5 se puede ver que para ambas medidas la tasa de error disminuye a medida que aumenta el porcentaje de empates. Además, la tasa de error de la *medida-F* tiene valores bajos (entre el 5 y el 7 %) mientras que el porcentaje de empates no se incremente mucho (del 3 al 13 %). Sin embargo, la tasa de error de AUC es mayor que la de la *medida-F* cuando el porcentaje de empates es bajo. De hecho, AUC obtiene una tasa de error menor solo cuando el porcentaje de empates es mayor al 25 %.

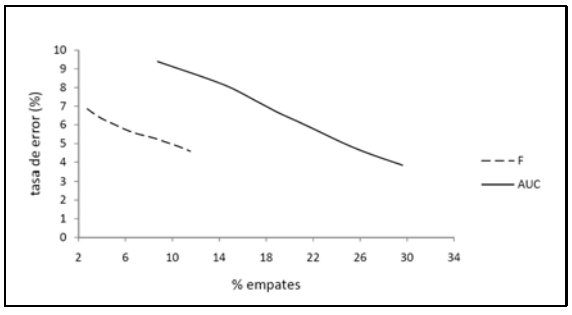

Figura 5: Curvas *porcentaje de empates - tasa de error* para medida-F y AUC

Por tanto se pueden obtener dos interpretaciones distintas del gráfico:

1. Si se fija el porcentaje de empates, la tasa de error obtenida para la *medida-F* es siempre menor que la obtenida para AUC. Por ejemplo, si se fija el porcentaje de empates al 10 % la tasa de error para la *medida-F* es del 5 % mientras que para AUC está entre el 9 y el 10 %. Por tanto a igual porcentaje de empates la *medida-F* es más estable que AUC.
2. Si se fija la tasa de error, el porcentaje de empates para AUC es mayor que para la *medida-F*. Por ejemplo, con una tasa de error del 6 % el porcentaje de empates para la medida-F es del 6 % mientras que para AUC es de casi el 22 %. Esto significa que si queremos asegurar que el sistema A es mejor que el sistema B con una determinada tasa de error usando la *medida-F*, tendremos menos empates entre sistemas y será necesario utilizar un valor menor de fuzziness que si se usa AUC.

Por tanto se observa que la *medida-F* es más estable que AUC además de tener un mayor poder de discriminación.

4.4. Cambios en el Ranking al Variar el Tamaño de las Colecciones

Otra característica importante en una medida de evaluación es cómo varía en función del

tamaño de la colección de evaluación utilizada. El comportamiento deseado es tener una medida que se mantenga estable sin importar el tamaño de la colección. Con este propósito se ha realizado un estudio creando colecciones de evaluación de distinto tamaño a partir de la colección original.

Se ha variado el tamaño de las colecciones desde 50 a 500 respuestas incrementando el tamaño de 50 en 50. Con el fin de aumentar el conjunto de datos se han realizado 200 ejecuciones distintas para cada tamaño de colección, calculándose para cada tamaño la media de la *medida-F* y de *AUC*.

Los resultados obtenidos se muestran en forma de gráfico en las Figuras 6 y 7. En ambas Figuras se puede comprobar que ambas medidas son bastante estables al variar el tamaño de la colección, habiendo menos fluctuaciones en el caso de *AUC*.

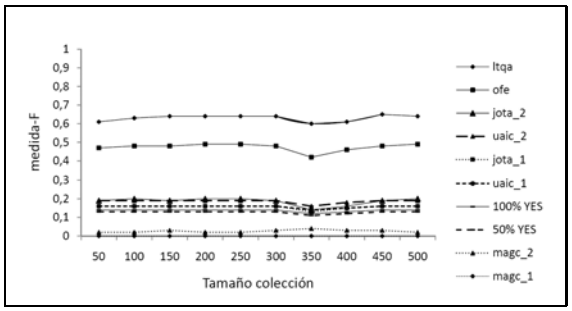

Figura 6: Variación de la media de la medida-F con distintos tamaños de colección

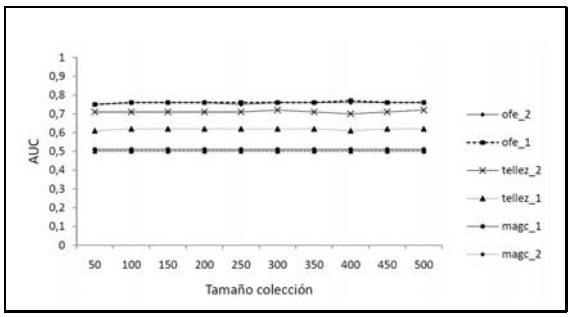

Figura 7: Variación de la media de AUC con distintos tamaños de colección

Por tanto, los resultados sugieren que ambas medidas son estables al variar el tamaño de la colección de evaluación.

4.5. Adecuación al Problema a Evaluar

El objetivo de un sistema de AV es mejorar los resultados del sistema de QA que lo esté utilizando. Para ello el sistema de AV debe colaborar en aumentar el número de respuestas correctas a la vez que se disminuye el número de respuestas incorrectas a la salida del sistema de QA.

Cuando un sistema de AV recibe una respuesta existen cuatro posibilidades dependiendo de si el sistema de AV valida o no la respuesta y si ésta es o no correcta. A cada una de estas posibilidades le corresponde una posición distinta en la matriz de confusión del Cuadro 1. Además, cada posibilidad contribuye de distinta manera al objetivo de mejorar los resultados en QA. De este modo, en este trabajo se ha establecido el siguiente orden de preferencias basándose en dicha contribución:

1. **Validar respuestas correctas**: esta es la salida con mayor valor dado que es la que contribuye en mayor medida a mejorar los resultados en QA.
2. **Rechazar respuestas incorrectas**: éste es también un comportamiento esperado. Detectar respuestas incorrectas permite que el sistema de QA considere otras respuestas que podrían ser correctas. Sin embargo y dado que el beneficio está subordinado a la habilidad de detectar respuestas correctas, a esta posibilidad se le da menos valor y ocupa el segundo lugar en el orden de preferencia.
3. **Rechazar respuestas correctas**: aunque este es un comportamiento erróneo, no provoca que el sistema de QA devuelva una respuesta incorrecta. Además el sistema se puede recuperar del error. Dado que esta recuperación está condicionada a encontrar una respuesta correcta, se considera a esta posibilidad en tercer lugar.
4. **Validar respuestas incorrectas**: esta es la posibilidad que más contribuye a empeorar los resultados de un sistema de QA. Dado que no hay oportunidad de recuperarse de este error, ésta es la peor salida que se puede obtener y por tanto es la menos deseada.

Se ha estudiado cómo asumen este orden de preferencias a la hora de evaluar sistemas de AV los dos enfoques estudiados en este trabajo.

Ambos enfoques tienen en cuenta la primera preferencia haciendo uso de *cobertura* (*tp rate*), de tal modo que se dan valores altos

		True class: Correct	True class: Incorrect	
Hypothesized class	**Correct**	68	129	197
	Incorrect	11	811	822
Column totals:		79	940	

Cuadro 4: Matriz de confusión del sistema *ofe*

a los sistemas que son capaces de validar correctamente una alta proporción de respuestas.

Las principales diferencias se encuentran en cuanto al control que se ejerce en la incorrecta validación de respuestas. Un valor bajo de *precisión* indica que se ha validado incorrectamente una cantidad alta de respuestas mientras que un valor alto significa lo contrario. Por tanto *precisión* premia a los sistemas que son capaces de validar una cantidad pequeña de respuestas incorrectas mientras que penaliza a los sistemas que validan grandes cantidades de respuestas incorrectas. De este modo *precisión* contribuye a controlar la incorrecta validación de respuestas.

Sin embargo este comportamiento no está tan bien controlado por *fp rate*. Esto es porque aunque un valor bajo de *fp rate* significa que se ha validado incorrectamente una baja proporción de respuestas, este valor es respecto al número total de respuestas incorrectas que hay y no respecto al número de respuestas que se han validado. Este concepto se entiende mejor con el siguiente ejemplo: obsérvese el Cuadro 4 que muestra la matriz de confusión del sistema *ofe* presente en las Figuras 2 y 3 y en el Cuadro 3. Según dicha matriz, *ofe* está validando incorrectamente 129 respuestas, obteniendo una *fp rate* de 0.14. Sin embargo, 129 respuestas representan casi el doble de respuestas de las que *ofe* está validando correctamente (68 respuestas). Esto es reflejado por *precisión*, cuyo valor es 0.35.

Por tanto, cuando un sistema de QA que utilice el sistema *ofe* devuelve una respuesta, hay más opciones de que la respuesta sea incorrecta que de que sea correcta. Esto ocurre porque según los resultados sólo el 35 % de las respuestas que son validadas por *ofe* son realmente correctas. Este es un comportamiento que no se desea pero al que la medida *fp rate* le está dando un buen valor (0.14 siendo el mejor 0), mientras que la *precisión* es baja (0.35 siendo el mejor valor posible 1). De este modo *precisión* está indicando de manera más fiable que *fp rate* los resultados que se podrían obtener en QA usando un determinado sistema de AV.

5. Conclusiones

En este trabajo se han comparado dos enfoques distintos para evaluar la Validación de sistemas de Búsqueda de Respuestas: uno basado en análisis ROC (Relative Operating Characteristic) y otro basado en el uso de *precisión-cobertura*. Ambos enfoques han sido comparados de acuerdo a sus fundamentos, su estabilidad en función del tamaño de las colecciones, su poder de discriminación y su adecuación a las particularidades de la Validación de Respuestas (AV).

Al estudiar la estabilidad de los enfoques, el basado en *precisión-cobertura* se ha mostrado más estable cuando se fija el poder de discriminación. Por otro lado, cuando se fija la estabilidad el enfoque basado en *precisión-cobertura* ha mostrado tener un mayor poder de discriminación.

Al variar el tamaño de las colecciones de evaluación ambos enfoques se han mostrado bastante estables.

La principal conclusión que se ha obtenido es que el enfoque *precisión-cobertura* parece mostrarse más útil para evaluar AV. Ello se debe a que el enfoque *precisión-cobertura* contribuye en mayor medida a controlar la incorrecta validación de respuestas, que es la peor salida que se puede obtener de un sistema de AV. Con esta salida no hay posibilidad de recuperarse del error y no se pueden mejorar los resultados de los sistemas de Búsqueda de Respuestas, el cuál es el objetivo principal de la AV.

Bibliografía

Beck, J. R. y E. K. Shultz. 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology & Laboratory Medicine*, 110(1):13–20.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of

machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July.

Buckley, C. y E. M. Voorhees. 2000. Evaluating Evaluation Measure Stability. En *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 33–40. ACM.

Drummond, C. y R. C. Holte. 2004. What ROC Curves can't do (and Cost Curves can). En *Proceedings of the 1st Workshop on ROC Analysis in Artificial Intelligence (held in conjunction with ECAI 2004), pages 19–26.*

Friedman, C. P. y J. C. Wyatt. 1997. Evaluation Methods in Medical Informatics. En *Springer-Verlag, New York.*

Hanley, J. A. y B. J. McNeil. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. *Radiology*, 143(1):29–36, April.

Harabagiu, S. y A. Hickl. 2006. Methods for Using Textual Entailment in Open-Domain Question Answering. En *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 905-912, Sydney.*

Hovy, E., L. Gerber, U. Hermjakob, M. Junk, y C. Lin. 2001. Question Answering in Webclopedia. En *Proceedings of the Ninth Text REtrieval Conference, pages 655-664.*

Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, y V. Rus. 2000. The structure and performance of an open-domain question answering system. En *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, páginas 563–570.

Peñas, A., Á. Rodrigo, V. Sama, y F. Verdejo. 2008. Testing the Reasoning for Question Answering Validation. En *Journal of Logic and Computation. 18(3)*, páginas 459–474.

Prager, J., E. Brown, A. Coden, y D. R. Radev. 2000. Question-answering by predictive annotation. En *Proceedings of the 23rd SIGIR Conference*, páginas 184–191.

Provost, F. y T. Fawcett. 2001. Robust Classification for Imprecise Environments. *Machine Learning*, 42(3):203–231.

Rodrigo, Á., A. Peñas, y F. Verdejo. 2009. Overview of the Answer Validation Exercise 2008. En *LNCS. To appear.*

Sakai, T. 2007. On the reliability of information retrieval metrics based on graded relevance. *Inf. Process. Manage.*, 43(2):531–548.