Métricas de evaluación automática

Víctor Peinado
victor@lsi.uned.es

Grupo de PLN y Recup. de Información
Dept. LSI - UNED

3 de febrero de 2004

Índice

- Evaluación.
- BLEU: evaluación automática para sistemas de MT.
- ROUGE: evaluación automática para sistemas de resúmenes automáticos.
- Referencias.

Índice

- Evaluación.
- BLEU: evaluación automática para sistemas de MT.
- ROUGE: evaluación automática para sistemas de resúmenes automáticos.
- Referencias.

Evaluación

- Evaluar sistemas es una tarea costosa: requiere mucho tiempo por parte de evaluadores humanos.
- Esto impide el avance de la investigación, especial-mente en campos en los que es recomendable realizar evaluaciones constantemente.
- Se hace, por lo tanto, imprescindible encontrar un método de evaluación automática que pondere aspectos como la adecuación, la fidelidad y la fuidez de la traducción.

Índice

- Evaluación.
- BLEU: evaluación automática para sistemas de MT.
- ROUGE: evaluación automática para sistemas de resúmenes automáticos.
- Referencias.

Sistemas de Traducción Automática

Cuanto más se aproxime una traducción automática a la de un traductor humano profesional, mejor traducción será.

Nuestro sistema de evaluación de MT necesita:

- Unidad numérica que mida la cercanía entre dos traducciones: una candidata y otra de referencia.
- Un corpus de referencia formado por traducciones humanas de calidad.

Precisión de una traducción

- Contabilizamos los ngramas comunes de la traducción candidata y la de referencia, sin tener en cuenta la posición en la que aparecen.
- Contabilizamos ngramas de distinta longitud tomando como unidad básica la oración.
- Cuantos más ngramas haya en común, mejor traducción será.

$$Precisión = \frac{C(ngramas\ comunes)}{C(ngramas\ candidato)}$$

Candidata 1: *A cat is on the mat.*
Candidata 2: *The cat exists in the board.*
Referencia: *The cat is on the table.*

Precisión(1) = 4/6 – Precisión(2) = 3/6

Precisión de ngramas modificada

Candidata 1: *the the the the the the.*
Referencia: *The cat is on the table.*

Precisión(1) = 6/6

- Tenemos en cuenta el número máximo de apariciones de cada ngrama en la traducción de referencia al contabilizar las apariciones en la traducción candidata.
- Dividimos por el total de ngramas de la traducción candidata.

$$p_n = \frac{C\_clip(ngramas\ comunes)}{C(ngramas\ cadidato)}$$

Candidata 1: *the the the the the the.*
Referencia: *The cat is on the table.*

$p_n(1) = 2/6$

Precisión de ngramas modificada

Este tipo de precisión modificada captura dos cualidades importantes de una buena traducción:

- Una traducción candidata que comparte palabras (unigramas) con la de referencia satisface la adecuación.
- La mayor longitud de los ngramas comunes satisface la fluidez.

Longitud de la frase

- La precisión de ngramas penaliza las palabras candidatas que no aparecen en la de referencia.
- La precisión modificada penaliza que una palabra candidata aprezca más frecuentemente que en la traducción de referencia.
- Necesitamos que la longitud de la traducción candidata sea similar a la de referencia.

Candidata 1: on *the*.
Referencia: *The cat is on the table.*

Precisión_m(1) = 2/2

Pensalización por brevedad

- Las traducciones candidatas más largas que las de referencia son penalizas por *Precisión_m*.
- Introducimos un *brevity penalty* que castiga las traducciones candidatas más breves.

$$BP = \begin{cases} 1 & si\ c > r \\ e^{(1-r/c)} & si\ c \leq r \end{cases}$$

Donde c es la longitud de la traducción candidata y r la longitud de la traducción de referencia.

Calculando BLEU

- Calculamos la media geométrica de la precisión modificada p_n, utilizando ngramas de varias longitudes (de 1 a N).
- Calculamos el factor de penalización por brevedad *BP*.
- Los pesos w_n han de ser positivos y sumar uno, normalmente $w_n = 1/N$.

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BLEU: Conclusiones

- Métrica tremendamente útil que acelera el avance del ciclo I+D en tareas de PLN: ahorra tiempo y recursos.
- Las puntuaciones BLEU se corresponden directamente con los juicios de evaluadores humanos y viceversa.
- Es posible diferenciar tanto sistemas de MT con diferencias substanciales como aquellos cuyas traducciones se diferencian sutílmente.
- Es independiente de las lenguas de origen y de destino.

Índice

- Evaluación.
- BLEU: evaluación automática para sistemas de MT.
- ROUGE: evaluación automática para sistemas de resúmenes automáticos.
- Referencias.

Evaluación de resúmenes

Cuanto más se aproxime un resumen automático al resumen realizado por un humano, mejor resumen será.

- ¿Podemos utilizar BLEU para evaluar la calidad de un resumen automático?
- BLEU es una métrica basada en la precisión. Quizá sea conveniente usar una medida basada en *recall*.
- Algunas características que hacen buena una traducción contradicen aquellas que hacen bueno un resumen.

ROUGE vs. BLEU

- Precisión C_n para ROUGE (Lin y Hovy, 2003), frente a precisión p_n para BLEU (Papineni *et al.*, 2002):

$$C_n = \frac{C\_clip(ngramas\ comunes)}{C(ngramas\ referencia)} \qquad p_n = \frac{C\_clip(ngramas\ comunes)}{C(ngramas\ cadidato)}$$

- Bonus de brevedad (*Brevity Bonus*), que premia los resúmenes breves que contienen la misma información que el resumen de referencia.

Calculando ROUGE

- Calculamos la media geométrica de la precisión modificada C_n, basada en recall y añadiendo el factor de bonificación por brevedad *BB*.

$$Ngram(i,j) = BB \cdot exp\left(\sum_{n=i}^{j} w_n \log C_n\right)$$

Donde j ≥ i, *i* y *j* van de 1 a 4 y $w_n = 1/(j-i+1)$.

Índice

- Evaluación.
- BLEU: evaluación automática para sistemas de MT.
- ROUGE: evaluación automática para sistemas de resúmenes automáticos.
- Referencias.

Referencias

- (Lin & Hovy, 2003a): "Automatic Evaluation of summaries using n-gram co-occurrence statistics".
- (Lin & Hovy, 2003b): "The Potential and Limitations of Automatic Sentence Extraction for Summarization".
- (NIST, 2002): "Automatic Evaluation of MT Quality using N-gram Co-Occurrence Statistics".
- (Papineni *et al.*, 2002): "BLEU: a method for Automatic Evaluation of Machine Translation".
- (Pastra & Saggion, 2002): "Colouring summaries BLEU".

Métricas de evaluación automática

Víctor Peinado
victor@lsi.uned.es

Grupo de PLN y Recup. de Información
Dept. LSI - UNED

3 de febrero de 2004