**BLEU SCORE**

La metrica BLEU ci consente di stimare l’efficienza del modello o sistema di Machine Translation. Questa metrica è stata largamente impiegato per i modelli e le applicazioni di *NLP*.

La metrica BLEU si basa sulla comparazione tra la frase candidata e la frase di riferimento, stimandone la corrispondenza assegnando alla coppia un valore nell’intervallo a valori reali [0, 1].

La metrica BLEU è semplice da calcolare e comprendere, ed ha diversi interessanti vantaggi. Anche se ci sono numerose alternative, continua ad essere una delle metriche più frequentemente utilizzate. È basata sull’idea che più la frase predetta è vicina alla frase target, più e corretta.

Due concetti che caratterizzano fortemente la metrica sono gli *N-gram* e la *Precision*.

N-GRAM

Concetto impiegato nel regular text processing per descrivere un insieme di N parole consecutive in una frase.

Osserviamo un esempio: nella frase “marco ha un cane” possiamo individuare i seguenti N-gram:

* 1-gram (*unigram*): “marco”, “ha”, “un”, “cane”
* 2-gram (*bigram*): “marco ha”, “ha un”, “un cane”
* 3-gram (*trigram*): “marco ha un”, “ha un cane”
* 4-gram: “marco ha un cane”

PRECISION

La metrica misura il numero delle parole nella frase predetta che ricorrono anche nella frase target. Vediamo un esempio:

* Frase target: “marco ha un cane”
* Frase predetta: “marco ha un gatto”

Normalmente, calcoleremmo la precision usando la formula:

Quindi, la precision, nel nostro esempio, sarebbe pari a 3/4. Ma usare la precision in questo modo non è sufficiente. Ci sono due casi che abbiamo ancora bisogno di analizzare:

1. REPETITION

Il primo problema è che questa formula ci consente di imbrogliare, come nel seguente esempio:

* Frase target: “marco ha un cane”
* Frase predetta: “cane cane cane”

In questo caso applicando la metrica della precision potremmo ottenere una precisione perfetta: 3/3

1. MULTIPLE TARGET SENTENCES

Ci sono molti modi corretti di esprimere una stessa frase. In molti modelli di NLP, ci potrebbero essere assegnate multiple frasi target accettabili che considerano le molteplici variazioni.

CLIPPED PRECISION

Teniamo in considerazione questi due problemi usando una formula modificata della precisione, che chiamiamo *Clipped Precision*.

Vediamo un esempio per osservare il suo funzionamento, considerando due frasi target e una frase predetta:

* Target 1: “marco ha un cane nero”
* Target 2: “marco ha un cane scuro”
* Predetta: “marco marco ha animale nero”

Ora, eseguiamo le seguenti azioni:

* Confrontiamo ogni parola della frase predetta con tutte quelle delle frasi target. Se la parola corrisponde ad una di una frase target, sarà considerata corretta
* Limitiamo il conto per ogni parola corretta al numero massimo di volte che quella parola compare nella frase target. Questo aiuta ad evitare il problema della ripetizione.

|  |  |  |  |
| --- | --- | --- | --- |
| PAROLA | FRASE CORRISPONDENTE | CONTO DELLE CORRISPONDENZE | CLIPPED COUNT |
| Marco | Entrambe | 2 | 1 |
| Ha | Target 1 | 2 | 1 |
| Animale | Target 2 | 0 | 0 |
| nero | No | 1 | 1 |
| TOTALE |  | 5 | 3 |

Da qui in avanti, calcoliamo la Precision utilizzando la formula della Clipped Precision.

COME CALCOLIAMO IL PUNTEGGIO BLEU

Preso in considerazione un modello NLP che produce la predizione di una frase, calcoliamo la precision (1-gram, 2-gram, 3-gram, 4-gram) sulle frasi predette.

Successivamente, combiniamo i punteggi di precisione ottenuti usando la Geometric Average Precision:

Possiamo osservare che:

Il terzo step è il calcolo della *Brevity Penalty*. Se si osserva il modo in cui la precision è calcolata, possiamo ottenere in output una frase predetta che consiste di una singola parola come “ha”. Per questo, la precisione 1-gram sarebbe dovuta essere 1/1 = 1, indicando un punteggio perfetto. Questo è chiaramente fuorviante in quanto sprona il modello a dare in output frasi con poche parole per ottenere un punteggio alto.

Per compensare questo, il Brevity Penalty penalizza frasi che sono troppo brevi

Nella formula:

* c è la lunghezza predetta, ossia il numero di parole nella frase predetta
* r è la lunghezza target, ossia il numero di parole nella frase target

Questo assicura che la Brevity Penalty non possa essere maggiore di 1, anche se la frase predetta è molto più lunga della frase target. Inoltre, se la predizione è composta da poche parole, questo valore sarà ridotto.

Infine, per calcolare il punteggio BLEU, moltiplichiamo per il Brevity Penalty la Geometric Average Precision degli score:

Il punteggio Bleu può essere calcolato per diversi valori di N. Generalmente, si usa N = 4.

* BLEU-1 usa il punteggio di unigram Precision
* BLEU-2 usa la Geometric Average Precision di unigram e bigram
* BLEU-3 usa la Geometric Average Precision di unigram, bigram e trigram
* ...

IMPLEMENTAZIONE DI BLEU

Per implementare la metrica BLEU utilizziamo il modulo NLTK che contiene la funzione *sentence\_bleu()*. Ci consente di passare una frase di riferimento e una frase candidata. Le funzioni confrontano la frase candidata e la frase di riferimento. Se viene individuato un match perfetto, il punteggio BLEU restituito è 1. Se non viene individuata alcun tipo di corrispondenza, il punteggio BLEU restituito è 0. Quindi, per un match parziale il punteggio BLEU sarà compreso tra 0 e 1.

VANTAGGI E SVANTAGGI DEL BLEU SCORE

Vantaggi:

* è veloce da calcolare e facile da comprendere;
* corrisponde al modo in cui un uomo valuterebbe lo stesso testo;
* è indipendente dal linguaggio, rendendone semplice l’applicazione ad un modello di NLP;
* può essere usato quanto vi è più di una frase vera (ground truth);
* è ampiamente usata, ciò consente di comparare facilmente i risultati ottenuti con quelli degli altri.

Svantaggi:

* non considera il significato delle parole. È perfettamente accettabile per un umano l’impiego di parole diverse con lo stesso significato;
* cerca unicamente corrispondenze esatte tra le parole. Qualche volta può essere impiegata una variante della stessa parola;
* ignora l’importanza delle parole. Non considera parole che sono più o meno significative nel dare significato ad una frase;
* non considera l’ordine delle parole.