



Sophie Rosset

LIMSI, CNRS

Mesures d'évaluation



Pourquoi ?

- Mesurer si le système de détection de concepts détecte bien les concepts
- Mesurer s'il tend à bien trouver quelques concepts ou beaucoup
- Mesurer s'il tend à se tromper sur les frontières, les types ou ...

Mesures

- Précision, rappel et F-mesure
- SER et C(V)ER
- BLEU (pour la génération)

Donnée par le ratio entre nombre de réponses correctes et toutes les réponses données par un système.

Permet d'estimer la fiabilité d'un système

$$P = \frac{C}{C + S + I} \quad (1)$$

Avec :

- C : nombre total d'objets annotés dans l'hypothèse qui sont corrects ;
- I : nombre total d'insertions opérées par le système c'est-à-dire d'éléments qui ne sont pas des entités mais que le système a considéré comme des entités ;
- S : nombre total de substitutions opérées par le système, c'est-à-dire d'entités bien détectées mais mal typées.

Avec donc :

- $C + S + I$: nombre total d'objets annotés dans l'hypothèse.

Donné par le ratio entre le nombre de réponses correctes et le nombre des réponses attendues (ie référence)

Permet d'estimer la capacité d'un système à couvrir l'ensemble des réponses se trouvant dans un corpus de test.

$$R = \frac{C}{C + S + D} \quad (2)$$

Avec :

- D : nombre total d'omission (*Deletions*) opérées par le système, c'est-à-dire d'entités non détectées ;
- $C + S + D$: nombre total d'objets à annoter dans la référence.

Systeme 1

deux concepts : pers et loc

REF : <pers> Bertrand Delanoë </pers> a été élu maire de
<loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers>
maire </pers> de <loc> Paris </loc>

$$P = \frac{2}{3} = 0,67 \quad (3)$$

$$R = \frac{2}{2} = 1 \quad (4)$$

Système 2

deux concepts : pers et loc

REF : <pers> Bertrand Delanoë </pers> a été élu maire de
<loc> Paris </loc>

HYP2 : <pers> Bertrand Delanoë </pers> a été élu maire de
Paris

$P = 1$ et $R = 1/2 = 0,5$

Précision tient compte des insertions produites par le systèmes ;

Rappel tient compte des omissions produites par le système.

Pour avoir une vision globale, on calcule la moyenne harmonique entre P et R

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 P + R} \quad (5)$$

β = poids qui permet d'ajuster l'importance qu'on accorde à la précision par rapport au rappel. (1 indique égale importance)

Pour Système 1 et Système 2, on a :

$$F_{Sys1} = (1 + 1^2) \times \frac{0,67 \times 1}{1^2 \times 0,67 + 1} = 0,80 \quad (6)$$

$$F_{Sys2} = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0,5} = 0,67 \quad (7)$$

Remarques

Inconvénient de la F-mesure :

- fusionner la précision et le rappel minimise le poids des erreurs d'insertion et d'omission par rapport aux erreurs de substitution
- besoin d'avoir des mesures pas seulement binaire (poids sur erreurs selon leur type)

Proposition : Slot Error Rate (SER)

SER

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \quad (8)$$

Avec :

- S_t et S_f le nombre total d'erreur de substitution de type et de frontières respectivement ;
- D et I le nombre total d'erreur respectivement d'omission et d'insertion d'entité ;
- α_1 α_2 β et γ les poids affectées à chaque catégorie d'erreur.

Concept (Value) Error Rate

- sur le modèle du WER
- compatibilise tous les concepts erronés et rapporte le chiffre au nb total de concepts dans l'énoncé

$$CER = \frac{S + D + I}{N} * 100 \quad (9)$$

- pour CVER on regarde les paires (C,V) et pas seulement les concepts

- comparer deux énoncés
- en comptant le nombre de 1/2/3/4-gram et en calculant la moyenne géométrique des précisions des 1-4gram
- on ajoute souvent une pénalité si l'hypothèse est trop courte
- très utilisé en traduction automatique

$$\text{Precision}_n = \frac{\sum_{h \in H} \sum_{n\text{-gram} \in h} \#(n\text{-gram}, \text{ref}(h))}{\sum_{h' \in H} \sum_{n\text{-gram}' \in h'} \#(n\text{-gram}', h)}, \quad (10)$$

$$\text{BP} = \left\{ \begin{array}{ll} 1 & \text{if } c > r \\ e^{(1-r)/c} & \text{if } c \leq r \end{array} \right\}, \quad (11)$$

$$\text{bleu} = \left(\prod_{o=1}^4 \text{Precision}_o \right)^{\frac{1}{4}} \times \text{BP} \quad (12)$$