Evaluation

Introduction

- L‘évaluation dépend des **phénomènes** a évaluer
- l'appréciation des résultats dépend de l'**application visée**
- l‘évaluation doit pouvoir être **reproduite**
- l‘évaluation doit être **compréhensible**
- Pour une évaluation nous devons mesurer et fournir une *baseline* et un gold standard

- Pour évaluer un système en TAL, il faut connaître la sortie souhaitée du système :
  - Ce n'est pas toujours facile. Par ex. pour une recherche d'informations sur le Web, on ne peut pas connaître tous les documents pertinents : s'il y avait un moyen de les retrouve r, on aurait alors un système parfait, et donc plus besoin de l'évaluer.
  - Pour un système de résumé automatique, il existe plusieurs réponses correctes, ou chaque résumé est plus ou moins correcte.
  - Pour un système de traduction automatique, plusieurs traductions sont possibles en sachant que la traduction 100% correcte n'existe pas (toute traduction entraîne une modification du sens).

Précision/rappel

- **Précision** : ratio de réponses correctes

Vrais Positifs VS Faux Positifs

- **Rappel** : ratio de réponses trouvées

Vrais Positifs VS Faux Négatifs

et a l'inverse :

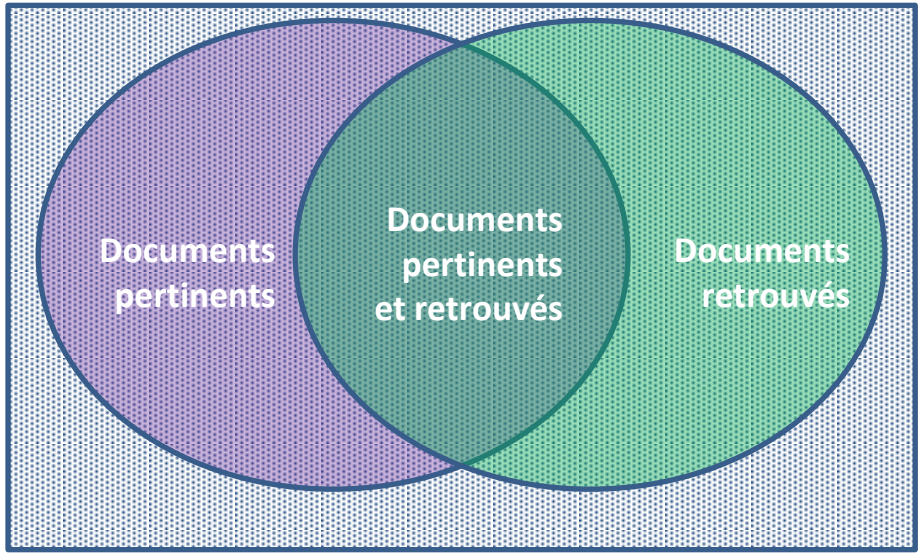
- **Bruit** : ratio de réponses incorrectes

  faux positifs / nb total d'alertes
- **Silence** : ratio de réponses oubliées

  faux négatifs / nb total erreurs

4 cas possibles

	Alerte	**Pas alerte**
Erreur	OK	Faux négatifs
Pas erreur	Faux positif	OK



Mesure d'évaluation

Précision et rappel :

Ce sont des nombres réels entre 0 et 1

- 0,5 de **précision** signifie que le système se trompe pour la moitié de documents qu'il a identifié.
- 0,75 de **précision** signifie que le système se trompe pour 25% des documents qu'il a identifié.
- 0,5 de **rappel** signifie que le système a identifié correctement la moitié de tous les documents et il n'a pas pu reconnaître l'autre moitié.

Système parfait / Gold standard :

- Précision = 1 & Rappel = 1

Mesures d'évaluation

- Supposons que l'on connaît la sortie souhaitée, pou r un système de classification de documents.
- Dans ce cas, l'évaluation s'effectue en utilisant l es mesures de précision et rappel

$$Précision = \frac{Nombre\ d'items\ pertinents\ trouvées\ (VP)}{nombres\ d'items\ trouvées\ (VP + FP)}$$

$$Rappel = \frac{Nombre\ d'items\ pertinents\ trouvées\ (VP)}{nombre\ d'items\ pertinents\ (VP + FN)}$$

- Mais il arrive qu'une augmentation de R se reflète positivement sur P
- P dépend indirectement de FN alors que R est indépendant de FP

F-mesure

- Evaluation combinée : la F-mesure :

- $F - mesure = (1 + \beta^2) * \frac{P*R}{(\beta^2*P)+R}$

- β^2= 1 : équilibrée, β < 1 favorise P et inversement

Précision et rappel : exemple

- Un système qui doit identifier parmi un ensemble de documents ceux qui parlent de sport.

- On le teste sur 9 documents :

Doc	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Référence	Sport	Autre	Sport	Sport	Autre	Sport	Sport	Autre	Sport
Hypothèse	Sport	Sport	Sport	Autre	Sport	Autre	Autre	Autre	Sport

1) Compter les correctes réponses

2) Calculer la précision et le rappel

Correction orthographique Word 10: Précision et rappel

Pharmacie : Actavis rachète Allergan et son botox

Le Monde.fr | 17.11.2014 à 08h24 • Mis à jour le 17.11.2014 à 16h15 | Par Chloé Hecketsweiler

Le laboratoire pharmaceutique américain Actavis a annoncée, lundi 17 novembre, qu'il allait racheter pour 66 milliards de dollars son compatriote Allergan, le fabricant de l'antirides Botox. Cette transaction, qui se faite en actions et en numéraire, vient contrecarrer les plans du groupe canadiennne Valeant Pharmaceuticals International, qui convoiter Allergan depuis Avril. Ce dernier a annoncé qu'il renonçait a surenchérir sur l'offre d'Actavis. Le mariage reste soumis à l'approbation des autorités de la concurrence. mais aussi et surtout à celle des actionnaires, dont ceux d'Allergan qui doivent se réunir en assemblé générale le 18 décembre. L'entité Allergan-Actavis aura un portefeuille diversifié, composée de médicaments sous protection et de génériques, notamment dans la dermatologie, la gynécologie et l'ophtalmologie.

Fautes correctement trouvées soulignées en jaune
Fautes qui ne sont pas des fautes soulignées en rouge
Fautes manquées soulignées en vert

A compléter à l'aide de la diapo 4, 5

	Alerte	Pas alerte	Total
Erreur			
Pas erreur			
Total			

Autres mesures : Bleu, NIST, WER score

BLEU score

- algorithme pour évaluer la précision du processus de traduction d'une langue en une autre
- nécessite préalablement un gold standard
- Le gold standard sera par la suite comparé avec la traduction produite automatiquement afin d'établir un score entre 0 et 1, où 1 signifie l'exactitude absolue entre les deux traductions et 0 que les deux traductions ne présentent aucune similarité
- Calcule la précision pour 1 à 4-grammes
- Pas compatible pour des phrases trop courtes ou trop longues

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

More info : https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

Autres mesures : Bleu, NIST, WER score

NIST score

- métrique alternative dérivée du BLEU score qui diffère sur le degré informatif qu'un n-gramme particulier peut avoir.
- Si un n-gramme rare a été correctement trouvé, il lui sera attribué plus de poids

WER (Word Error Rate)

- calcule la performance dans le domaine de la reconnaissance automatique de la parole et de la traduction automatique
- basée sur la mesure de distance de mots de Levenshtein et consiste à calculer le taux d'erreur entre une séquence de mots candidate séquence de mots de référence
- tient compte des erreurs de substitution (reconnaissance erronée d'un mot), des erreurs de suppression (absence de reconnaissance d'un mot) et des insertions (insertion de mots)

WER (Word Error Rate)

- $WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$

- *S* nombre de substitutions,
- *D* nombre de suppressions,
- *I* nombre insertions,
- *C* nombre de mots corrects,
- *N* nombre de mots en référence (N=S+D+C)