Corpus, ressources et linguistique outillée

Évaluation et Validation des Corpus et Modèles NLP

Dr Wiem Takrouni
Sorbonne Université
Master Langue et Informatique
UFR Sociologie et Informatique pour les Sciences Humaines

April 11, 2025

Plan du cours

1. Pourquoi évaluer ?
2. Métriques standards
3. Évaluer un corpus annoté
4. Évaluer un modèle NLP
5. Limites & biais de l'évaluation
6. Ressources recommandées

Objectifs

- Comprendre les enjeux de l'évaluation dans le traitement des corpus et des modèles NLP.
- Savoir choisir et utiliser les bonnes métriques.
- Identifier les limites et biais d'évaluation.
- Appliquer les méthodes sur des cas concrets.

Pourquoi évaluer ? (1/2)

Évaluation des corpus

- Fiabilité des annotations.
- Reproductibilité des résultats linguistiques.
- Ajustement des consignes via divergences.

Pourquoi évaluer ? (2/2)

Évaluation des modèles

- Mesure de la généralisation.
- Comparaison équitable des approches.
- Détection des erreurs et biais.

Conséquences pratiques : utile en production (recherche, chatbots, etc.)

Métriques standards : Classification (1/3)

Définitions des termes clés

- **TP (True Positives)** : Le modèle prédit la classe positive et c'est correct.
- **FP (False Positives)** : Le modèle prédit la classe positive alors qu'elle est négative.
- **FN (False Negatives)** : Le modèle ne détecte pas une instance réellement positive.
- **TN (True Negatives)** : Le modèle prédit correctement la classe négative.

Métriques standards : Classification (2/3)

Métriques associées

- **Précision** = TP / (TP + FP)
  - Mesure la qualité des prédictions positives.
  - Important pour éviter les faux positifs (ex : filtre anti-spam).
- **Rappel** = TP / (TP + FN)
  - Mesure la capacité à retrouver tous les cas positifs.
  - Crucial en médecine, sécurité, etc.

Métriques standards : Classification (3/3)

- **F1-score** = 2 × (précision × rappel) / (précision + rappel)
  - Combine précision et rappel.
  - Idéal pour classes déséquilibrées.
- **Accuracy** = (TP + TN) / Total
  - Taux global de bonnes prédictions.
  - Moins pertinent si les classes sont déséquilibrées.

Métriques standards : Annotation (1/2)

Objectif de l'évaluation d'annotation

- Vérifier la qualité et la cohérence des annotations manuelles.
- Quantifier l'accord entre annotateurs.
- Identifier les divergences pour affiner les consignes.

Cohen's Kappa

- Utilisé pour deux annotateurs.
- Prend en compte les accords dus au hasard.
- Interprétation :
- < 0 : pire que le hasard.
- $= 0$: équivalent au hasard.
- > 0.8 : excellent accord.

Formule : $\kappa = \frac{P_o - P_e}{1 - P_e}$

Métriques standards : Annotation (2/2)

Krippendorff Alpha

- Gère plusieurs annotateurs, données manquantes.
- Compatible avec différentes échelles (nominale, ordinale...).
- Idéal pour les projets collaboratifs à grande échelle.

Exemples d'interprétation

- $\alpha = 0.70$: accord modéré $\rightarrow$ discussion recommandée.
- $\alpha > 0.80$: accord solide $\rightarrow$ données utilisables pour entraînement.

Métriques standards : Génération (1/3)

BLEU (Bilingual Evaluation Understudy)

- Mesure basée sur les **n-grammes** communs entre la sortie générée et la référence.
- Calculée avec une pénalité de longueur (brevity penalty).
- Plus la sortie contient des séquences similaires à la référence, plus le score est élevé.
- Très utilisée pour la **traduction automatique**.

Limites :

- Sensible aux petites variations de formulation.
- Ne prend pas en compte la synonymie.

Métriques standards : Génération (2/3)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Compare des **unités de texte** (n-grammes, mots, phrases) entre résumé généré et référence.
- **ROUGE-1** : chevauchement des mots.
  ROUGE-2 : chevauchement des bigrammes.
  ROUGE-L : plus longue sous-séquence commune (LCS).

Utilisation :

- Évaluation des **résumés automatiques**.
- S'intéresse à la couverture du contenu important.

Métriques standards : Génération (3/3)

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

- Combine précision et rappel avec pondération équilibrée.
- Tient compte de :
  - **Synonymes** (via WordNet)
  - **Stemmatisation** (formes canoniques)
  - **Ordre des mots**
- Fournit des scores plus corrélés avec les jugements humains que BLEU.

Utilisation :

- Traduction, résumé, génération libre.

Évaluer un corpus annoté

- Vérifier l'homogénéité des annotations.
- Distinguer erreurs humaines/linguistiques.
- Utiliser des outils : WebAnno, INCEpTION.
- Construire un consensus.

Évaluer un modèle NLP : Exemples

- Classification d'avis : IMDb, Twitter.
- NER : corpus CoNLL2003, médical.
- Résumé automatique : CNN-DailyMail.

Méthodologie d'évaluation

- Séparation train/dev/test.
- Choix des métriques adaptées.
- Courbes utiles : matrice de confusion, PR-curve.

Outils pratiques pour l'évaluation

- `sklearn.metrics` : classification
- `seqeval` : NER
- `evaluate (Hugging Face)` : BLEU, ROUGE, METEOR

Limites et biais (1/2)

- Corpus biaisé : classes sur/sous-représentées.
- Ambiguïtés : annotations multiples possibles.

Limites et biais (2/2)

- Surapprentissage sur corpus test.
- Mauvais choix de métriques : BLEU pour contenu créatif ?

Ressources recommandées

- Hugging Face Datasets + Evaluate
- Scikit-learn
- Rouge-score
- Papers with Code : `https://paperswithcode.com/`