Statistiques pour l'évaluation

Damien Nouvel

Plan

Tâche et données

- Mesurer la **performance** pour une tâche donnée
- L'**évaluation** quantitative en TAL met en jeu
  - Une **tâche** à accomplir automatiquement
  - Un jeu de **données** pour
    - **Entraîner** (paramétrer) les modèles
    - **Développer** (améliorer) des modèles
    - **Évaluer** les modèles

  ⇒ **Efforts** pour préparer les données

  ⇒ Éviter les **biais**
- **Corrélations** entre **modélisation** et **évaluation**
  - Pour une tâche, les **performances** varient fortement selon
    - **Types** de données (domaine, modalité, qualité, etc.)
    - **Mesure** d'évaluation utilisée
    - ...

  ⇒ Bien tenir compte des **conditions** de l'évaluation

⇒ Recherche d'**objectivité** et de **fiabilité**

Plan

Mesures binaires

- Sur N données, cacul d'un **taux d'erreur** binaire
- Exactitude (en anglais **accuracy**)
- Exemple : "Je suis parti en vacances avec Marie" (7 mots)
  - Combien de données ont été **correctement traitées** ?

  $\Rightarrow$ Etiquetage morpho-syntaxique
    - `PRO VER VER PREP NC PREP NP` : $7/7 = 100\%$
    - `PRO VER` **`NC`** `PREP NC PREP NP` : $6/7 = 86\%$

  $\Rightarrow$ Extraction de noms propres
    - `NA NA NA NA NA NA NP` : $7/7 = 100\%$
    - `NA NA NA NA NA NA` **`NA`** : $6/7 = 86\%$
- Terminologie de l'évaluation :
  - Tests statistiques (maladies) : positif, négatif, vrai, faux
  - Recherche d'information (documents) : trouvé, correct

Table de contingence

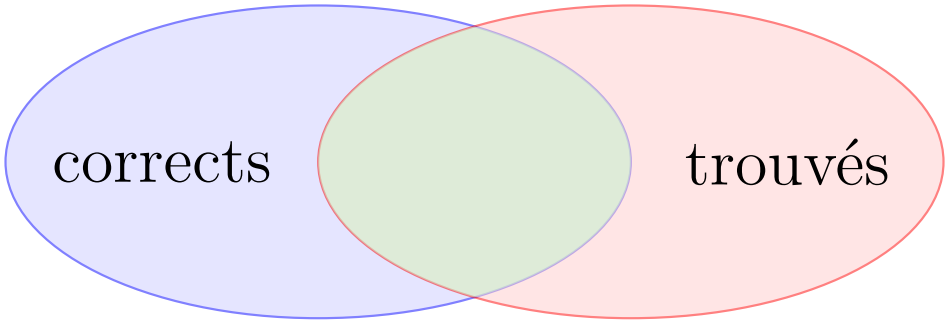
	malade	*non-malade*	*totaux*
positifs	30 (VP)	10 (FP)	40
négatifs	20 (FN)	50 (VN)	70
totaux	50	60	110

- Valeurs extraites de la table
  - **Vrais** (VP, VN) : test en accord avec la maladie
  - **Faux** (FP, FN) : test en désaccord avec la maladie
- Mesures pour **évaluer** les tests (rappels)
  - **Sensibilité** : $\frac{|VP|}{|VP|+|FN|} = \frac{|VP|}{|malade|} = \frac{30}{50} = 60\%$

  $\Rightarrow$ Taux de détection de la **présence** de maladie ?
  - **Spécificité** : $\frac{|VN|}{|VN|+|FP|} = \frac{|VN|}{|non-malade|} = \frac{50}{60} = 83\%$

  $\Rightarrow$ Taux de détection de l'**absence** de maladie ?

Précision, rappel, f-mesure



- Autre terminologie
  - **Trouvés** (positifs), **corrects** (malades)

  $\Rightarrow$ Pas d'équivalent pour **vrais** ou **faux**
- Mesures pour **évaluer** la **recherche d'information**
  - **Précision** : $P = \frac{|corrects \wedge trouvés|}{|trouvés|}$ (et **bruit**)
  - **Rappel** : $R = \frac{|corrects \wedge trouvés|}{|corrects|}$ (et **silence**)
  - **F-mesure** : $F_1 = \frac{2*P*R}{P+R}$

$\Rightarrow$ Par exemple : recherche de textes sportifs dans un corpus

- En complément : courbe ROC et mesure AUC

Moyennes de mesures

⇒ Lorsque l'on évalue **plusieurs catégories** (types)

type	*trouvés*	*corrects*	*trouvés et corrects*	**P**	**R**
personnes	90	100	80	0.89	0.8
organisations	15	20	10	0.67	0.5
macro-moyenne				0.78	0.65
micro-moyenne				0.86	0.75

- **Agrégation** de résultats
  - **Macro-moyenne**
    - Calcul des mesures, puis moyennes
    - Précision : $\frac{0.89+0.67}{2} = 0.78$ / Rappel : $\frac{0.8+0.5}{2} = 0.65$

    ⇒ Poids par catégorie
  - **Micro-moyenne**
    - Sommes sur les valeurs les plus fines, puis calcul de la mesure
    - Précision : $\frac{80+10}{90+15} = 0.86$ / Rappel : $\frac{80+10}{100+20} = 0.75$

    ⇒ Prise en compte de chaque donnée

    ⇒ Si toutes les données sont analysées, la précision, le rappel, l'exactitude sont égaux

Tables de confusion

- Visualisation des résultats réels vs prédits
- Matrice **carrée**
- **Diagonale** : résultats corrects
- Exemple :

		prédiction		
		sport	*politique*	*économie*
réel	*sport*	**15**	3	5
	politique	4	**18**	7
	économie	12	21	**19**

Plan

Séquences textuelles (énoncés)

- Reconnaissance d'entités nommées (**types**)
  - Classification de segments (**types** et **extensions**)

  ⇒ **SER** (Slot Error Rate)
- Reconnaissance de parole / écriture
  - Erreurs par mots (**insertions, substitutions**)

  ⇒ **WER** (Word Error Rate)
- Traduction (**présence de mots traduits**)
  - Qualité globale du texte, dont vocabulaire

  ⇒ **BLEU** (Bilingual Evaluation Understudy)
- Résumé automatique
  - Qualité globale du texte, dont vocabulaire
  - **ROUGE** (Recall-Oriented Understudy of Gisting Eval.)

Données ordonnées

- Nombreux résultats (moteurs de recherche)
  - **Pertinence** des premiers résultats
  - Moins d'importance en **fin de classement**
- Mesures selon les k premiers résultats sur N
  - Précision au rang k : $P(k)$ (et rappel $R(k)$)
  - R-Précision, selon le nombre de résultats attendus
  - Précision moyenne :

$$MoyP = \sum_{k=1}^{N} P(k) * (R(k) - R(k-1))$$

  - nDCG : normalized discounted cumulative gain

$\Rightarrow$ Prise en compte de l'ordre en **recherche d'information**

Plan

Fiabilité des annotations

- L'annotation est sujette à de la **variabilité**
  - Qualité des observables
  - Défnition des objets à annoter
  - Interprétations lors de l'annotation
  - ...
- Mesures d'**accord inter-annotateur**
  - Taux d'annotations identiques (T)
  - Kappa de Cohen :

$$\kappa = \frac{T - P}{1 - P}$$

P est la probabilité de l'accord pas hasard

$\Rightarrow$ 1 : accord parfait / 0 : désaccord ou hasard

Échantillonnage de données

- Pour des variables continues
  - Sélection à intervalles réguliers
  - Génération de valeurs par paramètres (moyenne, écart-type)
  - ...
- Échantillonnage d'un ensemble **très volumineux**
  - **Tirage aléatoire** des exemples
  - Sélection d'exemples **représentatifs** (similarité)
  - **Pondération** des exemples selon leur distance au centroïdes
  - ...
- Échantillonnage d'ensembles **infinis** ("Reservoir Sampling")
  - Sélection des k premiers éléments dans le réservoir
  - Pour le autres éléménts $i > k$:
    - Tirage au sort d'un chiffre $0 < p < i$
    - Si $p < k$ remplacer l'élément p du réservoir