Annotation collaborative de corpus : Évaluer la qualité de l'annotation manuelle

Karën Fort

karen.fort@sorbonne-universite.fr / https://members.loria.fr/KFort/

LETTRES SORBONNE UNIVERSITÉ

Quelques sources d'inspiration

par ordre d'importance décroissant

- Les articles de référence :
  - *Inter-Coder Agreement for Computational Linguistics* [Artstein and Poesio, 2008]
  - *The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment* [Mathet et al., 2015]
- Présentation de Massimo Poesio à LREC sur le sujet (avec son accord)
- Le cours de Gemma Boleda et Stefan Evert sur le sujet à ESSLLI 2009 (avec leur accord) [http://esslli2009.labri.fr/course.php?id=103]
- Yann Mathet

Introduction

Question fondamentale : **les annotations sont-elles correctes ?**

- les systèmes apprennent les erreurs des annotateurs humains (bruit $\neq$ régularités dans les erreurs [Reidsma and Carletta, 2008])
- l'évaluation peut-être faussée
- les résultats d'analyse linguistique ou de systèmes symboliques peuvent être faussés et non concluant

Rappel : le consensus, au cœur de l'annotation

Il faut «convenir pour mesurer »[Desrosières, 2008]

L'annotation est de l'ordre de la quantification

Mesurer *vs* quantifier [Desrosières, 2008] :

- **mesurer** : implique une forme mesurable (par ex. la hauteur du Mont Blanc)
- **quantifier** : suppose des conventions d'équivalences préalables

Outiller le consensus :

- guide d'annotation (12 p. pour le football)
- réunions avec les annotateurs et le gestionnaire de la campagne
- évaluer le consensus (la cohérence)

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte
- nous ne pouvons mesurer que la **fiabilité** de l'annotation

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte
- nous ne pouvons mesurer que la **fiabilité** de l'annotation
  - *i.e.* si les annotateurs humains prennent les mêmes décisions de manière **cohérente** $\Rightarrow$ ils ont internalisé le schéma d'annotation

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte
- nous ne pouvons mesurer que la **fiabilité** de l'annotation
  - *i.e.* si les annotateurs humains prennent les mêmes décisions de manière **cohérente** $\Rightarrow$ ils ont internalisé le schéma d'annotation
  - hypothèse sous-jacente : une fiabilité élevée implique la validité de l'annotation

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte
- nous ne pouvons mesurer que la **fiabilité** de l'annotation
  - *i.e.* si les annotateurs humains prennent les mêmes décisions de manière **cohérente** $\Rightarrow$ ils ont internalisé le schéma d'annotation
  - hypothèse sous-jacente : une fiabilité élevée implique la validité de l'annotation
- Comment déterminer cette fiabilité ?

Mesurer la fiabilité (cohérence)

- chaque item est annoté par un seul annotateur, avec des vérifications aléatoires ($\approx$ seconde annotation)
- certains items sont annotés par deux annotateurs ou plus
- chaque item est annoté par deux annotateurs ou plus - suivi d'une phase de conciliation
- chaque item est annoté par deux annotateurs ou plus - suivi d'une décision finale prise par un superannotateur (expert)

Dans tous les cas, la mesure de la fiabilité est un coefficient d'accord (inter-annotateurs)

Cas particulier : existence d'un *gold-standard*

Dans certains cas (rare et souvent artificiels), il existe une "référence" :
le corpus a été annoté, au moins partiellement, et cette annotation est considérée comme "parfaite", une référence
[Fort and Sagot, 2010].

Dans ces cas, une autre mesure, complémentaire, peut être utilisée :

Laquelle ?

Cas particulier : existence d'un *gold-standard*

Dans certains cas (rare et souvent artificiels), il existe une "référence" :
le corpus a été annoté, au moins partiellement, et cette annotation est considérée comme "parfaite", une référence
[Fort and Sagot, 2010].

Dans ces cas, une autre mesure, complémentaire, peut être utilisée :

F-mesure

Précision / Rappel : retour à la base

- Rappel :

- Silence :
- Précision :

- Bruit :

Précision / Rappel : retour à la base

- Rappel : mesure la quantité d'annotations trouvées

$$\text{Rappel} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb d'annotations correctes attendues}}$$

- Silence :
- Précision :

- Bruit :

Précision / Rappel : retour à la base

- Rappel : mesure la quantité d'annotations trouvées

$$\text{Rappel} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb d'annotations correctes attendues}}$$

- Silence : *complément* du rappel (annotations correctes non trouvées)
- Précision :

- Bruit :

Précision / Rappel : retour à la base

- Rappel : mesure la quantité d'annotations trouvées

$$\text{Rappel} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb d'annotations correctes attendues}}$$

- Silence : *complément* du rappel (annotations correctes non trouvées)
- Précision : mesure la qualité des annotations trouvées

$$\text{Précision} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb total d'annotations trouvées}}$$

- Bruit :

Précision / Rappel : retour à la base

- Rappel : mesure la quantité d'annotations trouvées

$$\text{Rappel} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb d'annotations correctes attendues}}$$

- Silence : *complément* du rappel (annotations correctes non trouvées)
- Précision : mesure la qualité des annotations trouvées

$$\text{Précision} = \frac{\text{Nb d'annotations correctes trouvées}}{\text{Nb total d'annotations trouvées}}$$

- Bruit : *complément* de la précision (annotations incorrectes trouvées)

F-mesure : retour à la base (Wikipedia 10 déc., 2010)

Moyenne harmonique de la précision et du rappel ou F-score équilibré

$$F = 2x\frac{\text{précision } x \text{ rappel}}{\text{précision} + \text{rappel}}$$

... ou la F1 mesure, le rappel et la précision ayant des poids équivalents.

Il s'agit d'un cas particulier de la Fβ mesure :

$$F\beta = (1+\beta^2)x\frac{\text{précision } x \text{ rappel}}{\beta^2 x\text{précision} + \text{rappel}}$$

La valeur de β permet de favoriser :

- le rappel ($\beta = 2$)
- la précision ($\beta = 0{,}5$)

“Gold-standard” ?

- il est rare qu’une référence existe déjà
- peut-elle être “parfaite” ? [Fort and Sagot, 2010]

→ peut-on utiliser la F-mesure dans d’autres cas ? Voir [Hripcsak and Rothschild, 2005]

⇒ Retour aux coefficients d’accord inter-annotateurs.

Exemple

Validation d'annotations sémantiques (contenu/contenant) :

Phrase	A	B	D'accord ?
Put tea in a heat-resistant jug and add the boiling water.	✓	✓	✓
Where are the batteries kept in a phone ?	✗	✓	✗
Vinegar's usefulness doesn't stop inside the house.	✗	✗	✓
How do I recognize a room that contains radioactive materials ?	✓	✓	✓
A letterbox is a plastic, screw-top bottle that contains a small notebook and a unique rubber stamp.	✓	✗	✗

→ **Accord inter-annotateurs ?**

Représentation synthétique (matrice de confusion)

		A		
		✓	✗	Total
B	✓	**4**	2	6
	✗	2	**2**	4
	Total	6	4	**10**

Accord observé (A_o)

proportion de réponses sur lesquelles les annotateurs sont du même avis.

Ici :

Représentation synthétique (matrice de confusion)

		A		
		✓	✗	Total
B	✓	**4**	2	6
	✗	2	**2**	4
	Total	6	4	**10**

Accord observé (A_o)

proportion de réponses sur lesquelles les annotateurs sont du même avis.

Ici : $A_o = \frac{4+2}{10} = \mathbf{0{,}6}$

Et si...

... une partie de l'accord était due au **hasard** :
dans notre exemple, quelle proportion d'accord peut être due au hasard ?

Et si...

... une partie de l'accord était due au **hasard** :

- Deux annotateurs annotant au hasard seront d'accord **la moitié du temps** (0,5).
- L'accord qui peut être dû au hasard varie selon le schéma d'annotation et les données annotées.

L'accord significatif est celui qui se situe au-dessus du hasard.
$\rightarrow$ rejoint le concept de *baseline* (résultat plancher).

Et si ?

Exercice

- chaque unité doit être annotée
- 2 catégories  et 
- 3 annotateurs : A_1, A_2 et A_3

Quelles sont les différentes possibilités d'annotations (sur une unité) ?

Correction et suite de la réflexion

Dans ce cas précis, il est impossible d'avoir un accord (par paire d'annotateurs) nul :

A_1	A_2	A_3	Paires en accord
🟦	🟦	🟦	?
🟦	🟦	🟨	?
🟦	🟨	🟨	?
🟨	🟨	🟨	?
🟨	🟨	🟦	?
🟨	🟦	🟦	?
🟨	🟦	🟨	?
🟦	🟨	🟦	?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
🟦	🟦	🟦	3
🟦	🟦	🟨	?
🟦	🟨	🟨	?
🟨	🟨	🟨	?
🟨	🟨	🟦	?
🟨	🟦	🟦	?
🟨	🟦	🟨	?
🟦	🟨	🟦	?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
🟦	🟦	🟦	3
🟦	🟦	🟨	1
🟦	🟨	🟨	?
🟨	🟨	🟨	?
🟨	🟨	🟦	?
🟨	🟦	🟦	?
🟨	🟦	🟨	?
🟦	🟨	🟦	?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			?
			?
			?
			?
			?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			?
			?
			?
			?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			1
			?
			?
			?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			1
			1
			?
			?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			1
			1
			1
			?

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			1
			1
			1
			1

Correction et suite de la réflexion

A_1	A_2	A_3	Paires en accord
			3
			1
			1
			3
			1
			1
			1
			1

Dans le cas le pire, on aurait bien $8x1/8x3 = 0,333$

Et si ?

Exercice (suite)

- chaque unité doit être annotée
- 2 catégories
- ~~3~~ 2 annotateurs

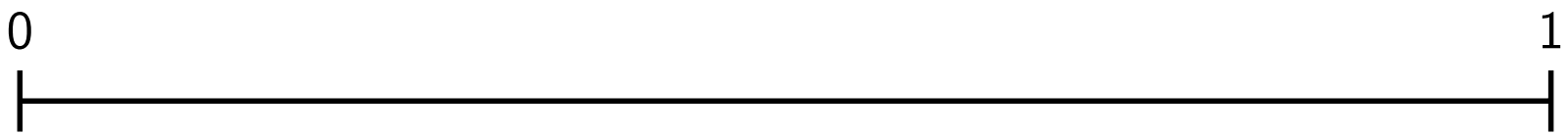
Quelles sont les différentes possibilités d'annotations (sur une unité) ?

Échelles d'accord

L'accord inter-annotateurs ne va pas être calculé sur la même échelle selon les cas :

- Cas 1 : 3 annotateurs pour 2 catégories

- Cas 2 : 2 annotateurs pour 2 catégories



Échelles d'accord

L'accord inter-annotateurs ne va pas être calculé sur la même échelle selon les cas :

- Cas 1 : 3 annotateurs pour 2 catégories

- Cas 2 : 2 annotateurs pour 2 catégories

$\rightarrow$ nécessité d'une certaine correction des résultats observés pour pouvoir interpréter les résultats

Prendre en compte le hasard

Accord attendu (*expected agreement* A_e)

valeur attendue de l'accord observé.

Montant d'accord au-dessus du hasard : $A_o - A_e$
Maximum possible d'accord au-dessus du hasard : $1 - A_e$

Proportion d'accord au-dessus du hasard atteinte : $\frac{A_o - A_e}{1 - A_e}$

Accord parfait : $\frac{1 - A_e}{1 - A_e}$
Désaccord parfait : $\frac{-A_e}{1 - A_e}$

Accord attendu

Comment calculer le montant d'accord attendu dû au hasard (A_e) ?

S [Bennett et al., 1954]

S

Même probabilité pour tous les annotateurs et toutes les catégories.

Nombre d'étiquettes : q
Probabilité qu'un annotateur choisisse une catégorie q_a : $\frac{1}{q}$
Probabilité que deux annotateurs choisissent une catégorie q_a : $(\frac{1}{q})^2$

Probabilité que deux annotateurs choisissent la même catégorie :

$$A_e^S = q.(\frac{1}{q})^2 = \frac{1}{q}$$

Toutes les catégories sont équiprobables : conséquences

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

Toutes les catégories sont équiprobables : conséquences

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = \frac{30+20}{60} = 0,8333$
$A_e^S = \frac{1}{2} = 0,5$

$S = \frac{0,8333-0,5}{1-0,5} = \mathbf{0,6666}$

Toutes les catégories sont équiprobables : conséquences

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = \frac{30+20}{60} = 0,8333$
$A_e^S = \frac{1}{2} = 0,5$

$S = \frac{0,8333-0,5}{1-0,5} = \mathbf{0,6666}$

Toutes les catégories sont équiprobables : conséquences

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = \frac{30+20}{60} = 0,8333$
$A_e^S = \frac{1}{2} = 0,5$

$S = \frac{0,8333-0,5}{1-0,5} = \mathbf{0,6666}$

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = \frac{30+20}{60} = 0,8333$
$A_e^S = \frac{1}{4} = 0,25$

$S = \frac{0,8333-0,25}{1-0,25} = \mathbf{0,7777}$

π [Scott, 1955]

π

Différentes probabilités d'apparition pour différentes catégories.

Nombre total de jugements : N
Probabilité qu'un annotateur choisisse une catégorie q_a : $\frac{n_{q_a}}{N}$
Probabilité que deux annotateurs choisissent une catégorie q_a : $(\frac{n_{q_a}}{N})^2$

Probabilité que deux annotateurs choisissent la même catégorie :

$$A_e^{\pi} = \sum_q (\frac{n_q}{N})^2 = \frac{1}{N^2} \sum_q n_q^2$$

Comparer S et π

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$
$S = \mathbf{0,6666}$

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$
$S = \mathbf{0,7777}$

Comparer S et π

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$

$S = \mathbf{0,6666}$

$A_e^{\pi} = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$

$S = \mathbf{0,7777}$

Comparer S et π

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$

$S = \mathbf{0,6666}$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$

$S = \mathbf{0,7777}$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

κ [Cohen, 1960]

κ

Différents annotateurs peuvent avoir différentes interprétations des instructions (biais/préjugés). κ prend en compte le biais individuel.

Nombre total d'items : i
Probabilité qu'un annotateur A_x choisisse une catégorie q_a : $\frac{n_{A_x q_a}}{i}$
Probabilité que deux annotateurs choisissent une catégorie q_a : $\frac{n_{A_1 q_a}}{i} \cdot \frac{n_{A_2 q_a}}{i}$

Probabilité que deux annotateurs choisissent la même catégorie :

$$A_e^\kappa = \sum_q \frac{n_{A_1 q}}{i} \cdot \frac{n_{A_2 q}}{i} = \frac{1}{i^2} \sum_q n_{A_1 q} n_{A_2 q}$$

Comparer π et κ

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_o = 0,8333$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

Comparer π et κ

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$

$A_e^{\pi} = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_e^{\kappa} = \frac{(\frac{35x35}{60})+(\frac{25x25}{60})}{60} = 0,5139$

$\kappa = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_o = 0,8333$

$A_e^{\pi} = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

Comparer π et κ

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_e^\kappa = \frac{(\frac{35x35}{60})+(\frac{25x25}{60})}{60} = 0,5139$

$\kappa = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

	Oui	Non	C	D	Total
Oui	**30**	5	0	0	35
Non	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	35	25	0	0	**60**

$A_o = 0,8333$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_e^\kappa = \frac{(\frac{35x35}{60})+(\frac{25x25}{60})}{60} = 0,5139$

$\kappa = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

Comparer π et κ (matrice non symétrique)

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$

$A_e^\pi = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

	Oui	Non	Total
Oui	**24**	8	32
Non	14	**24**	38
Total	38	32	**70**

$A_o = 0,68$

$A_e^\pi = \frac{((\frac{38+32}{2})^2+(\frac{32+38}{2})^2)}{70^2} = 0,5$

$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$

Comparer π et κ (matrice non symétrique)

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$A_o = 0,8333$

$A_e^{\pi} = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$

$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

$A_e^{\kappa} = \frac{(\frac{35x35}{60})+(\frac{25x25}{60})}{60} = 0,5139$

$\kappa = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$

	Oui	Non	Total
Oui	**24**	8	32
Non	14	**24**	38
Total	38	32	**70**

$A_o = 0,68$

$A_e^{\pi} = \frac{((\frac{38+32}{2})^2+(\frac{32+38}{2})^2)}{70^2} = 0,5$

$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$

Comparer π et κ (matrice non symétrique)

	Oui	Non	Total
Oui	**30**	5	35
Non	5	**20**	25
Total	35	25	**60**

$$A_o = 0,8333$$
$$A_e^{\pi} = \frac{((\frac{35+35}{2})^2+(\frac{25+25}{2})^2)}{60^2} = 0,5139$$
$$\pi = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$$
$$A_e^{\kappa} = \frac{(\frac{35 \times 35}{60})+(\frac{25 \times 25}{60})}{60} = 0,5139$$
$$\kappa = \frac{0,8333-0,5139}{1-0,5139} = \mathbf{0,6571}$$

	Oui	Non	Total
Oui	**24**	8	32
Non	14	**24**	38
Total	38	32	**70**

$$A_o = 0,68$$
$$A_e^{\pi} = \frac{((\frac{38+32}{2})^2+(\frac{32+38}{2})^2)}{70^2} = 0,5$$
$$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$$
$$A_e^{\kappa} = \frac{(\frac{38 \times 32}{70})+(\frac{32 \times 38}{70})}{70} = 0,49$$
$$\kappa = \frac{0,68-0,49}{1-0,49} = \mathbf{0,37}$$

S, π et κ

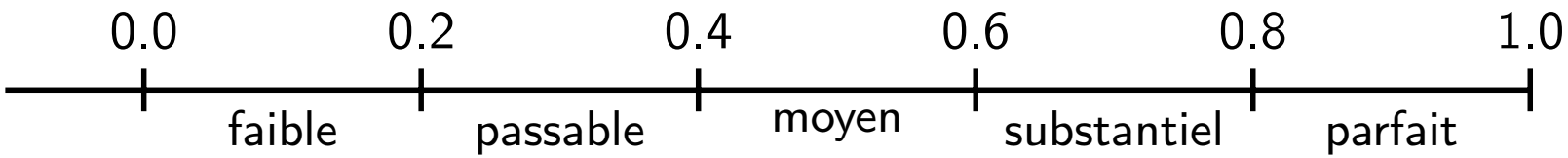
Pour n'importe quel échantillon :

$$A_e^\pi \geqslant A_e^S \qquad \pi \leqslant S$$
$$A_e^\pi \geqslant A_e^\kappa \qquad \pi \leqslant \kappa$$

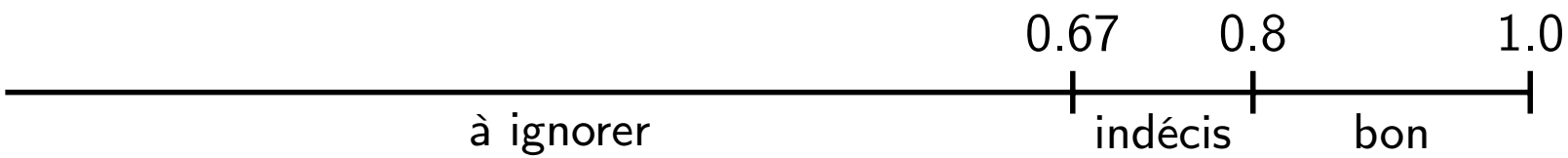
Qu'est-ce qu'un "bon" κ (ou π ou S) ?

Échelles d'interprétation de Kappa

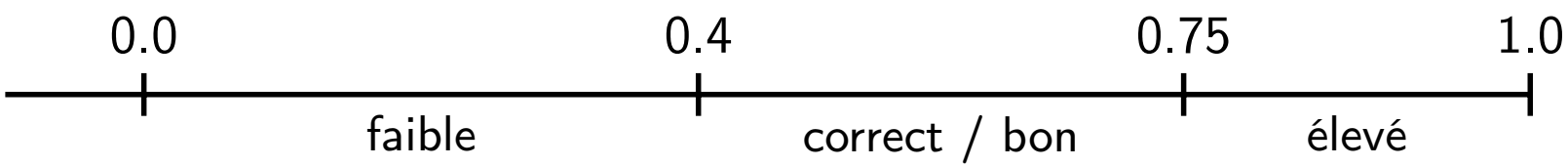
[Landis and Koch, 1977]



[Krippendorff, 1980]



[Green, 1997]



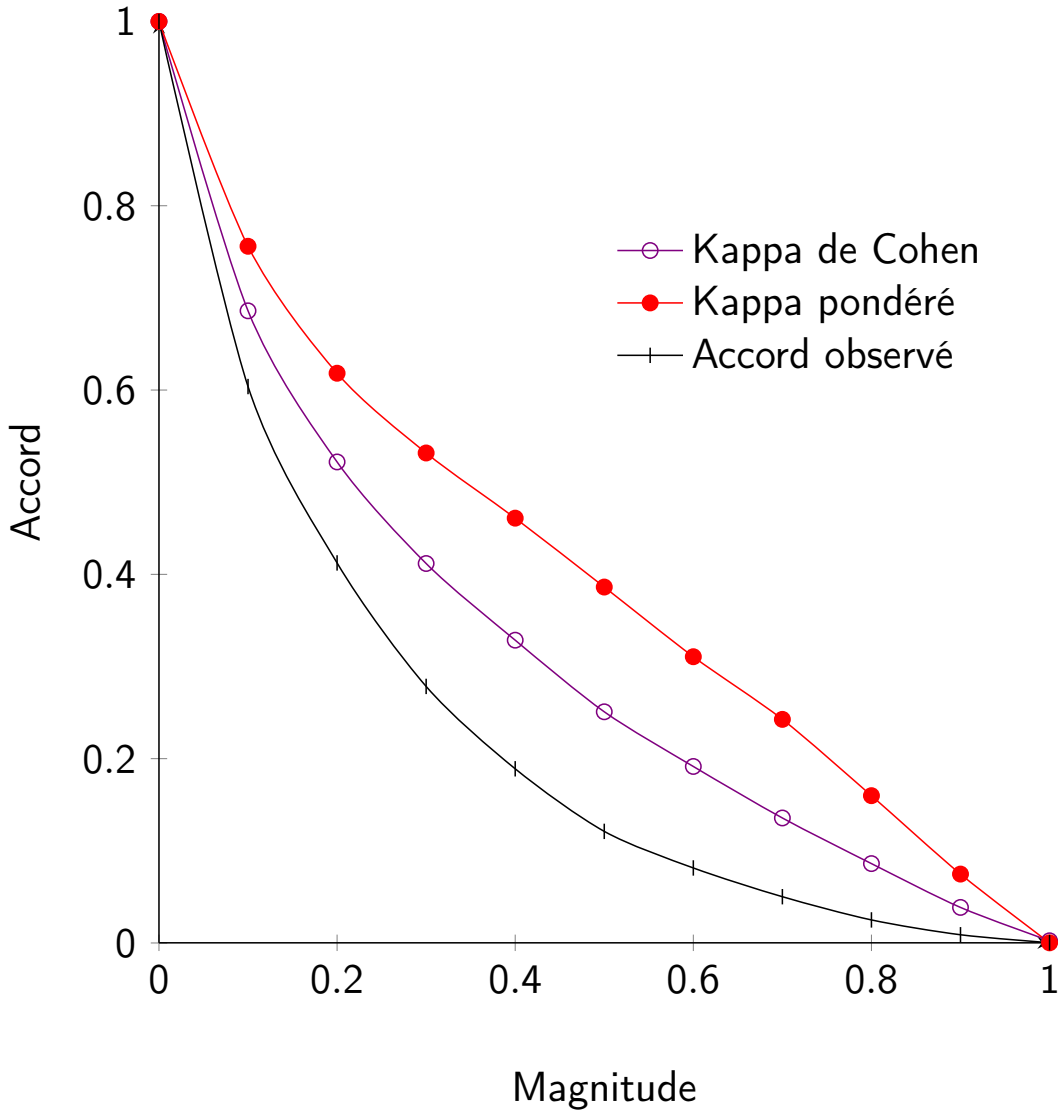
Donner du sens au résultat obtenu [COLING 2012a]

Création d'un outil "Richter" qui :

- prend en entrée une annotation de référence (réelle ou générée automatiquement)
- génère des dégradations d'une certaine magnitude (de 0 à 1)
- applique une ou des mesures d'accord inter-annotateurs sur chaque ensemble d'annotations (correspondant à une magnitude de dégradation)

Richter sur le corpus TCOF-POS

Pas de prévalence, mais proximité entre catégories prise en compte :



Biais

Les annotateurs bien formés sont moins sensibles aux biais :

- de la pré-annotation [Fort and Sagot, 2010]
- de l'outil d'aide à l'annotation [Dandapat et al., 2009]

et annotent moins "par hasard"

Utiliser un guide d'annotation permet d'obtenir de meilleures annotations [Nédellec et al., 2006]

Expert es-tu là ?

Experts :

- du domaine : annotation en microbiologie (renommage de gènes), en football, etc.
- de la tâche : annotation en entités nommées structurées

... des contradictions et des insuffisances :

→ pour des EN structurées dans de la presse ancienne, vaut-il mieux des spécialistes en EN structurées ou des historiens ?

- Précision, rappel, F-mesure
- Exactitude
- Accord observé
- S, κ, π
- Signification

À faire

Exercice

Calculez les accords inter-annotateurs sur la campagne de renommage de noms de gènes.

Que remarquez-vous ?


Artstein, R. and Poesio, M. (2008).
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4) :555–596.


Bennett, E. M., Alpert, R., and C.Goldstein, A. (1954).
Communications through limited questioning.
Public Opinion Quarterly, 18(3) :303–308.


Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1) :37–46.


Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
Complex linguistic annotation - no easy way out ! a case from bangla and hindi POS labeling tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.


Desrosières, A. (2008).

Pour une sociologie historique de la quantification : L'Argument statistique.
Presses de l'école des Mines de Paris.


Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63, Uppsala, Suède.


Green, A. M. (1997).
Kappa statistics for multiple raters using categorical classifications.
In Proceedings of the Twenty-Second Annual Conference of SAS Users Group, San Diego, USA.


Hripcsak, G. and Rothschild, A. S. (2005).
Agreement, the f measure, and reliability in information retrieval.
Journal of the American Medical Informatics Association (JAMIA), 12(3) :296–298.


Krippendorff, K. (1980).
Content Analysis : An Introduction to Its Methodology.
Sage, Beverly Hills, CA., USA.


Landis, J. R. and Koch, G. G. (1977).
The measurement of observer agreement for categorical data.
Biometrics, 33(1) :159–174.


Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012).
Manual corpus annotation : Evaluating the evaluation metrics.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 809–818, Mumbaï, Inde.
Poster.


Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015).
The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment.
Computational Linguistics, 41(3) :437–479.


Nédellec, C., Bessières, P., Bossy, R., Kotoujansky, A., and Manine, A.-P. (2006).
Annotation guidelines for machine learning-based named entity recognition in microbiology.
In et C. Nédellec, M. H., editor, Proceedings of the Data and text mining in integrative biology workshop, pages 40–54, Berlin, Allemagne.


Reidsma, D. and Carletta, J. (2008).
Reliability measurement without limits.
Computational Linguistics, 34(3) :319–326.


Scott, W. A. (1955).
Reliability of content analysis : The case of nominal scale coding.
Public Opinion Quaterly, 19(3) :321–325.