TALN 2010, Montréal, 19–23 juillet 2010

Évaluer des annotations manuelles dispersées : les coefficients sont-ils suffisants pour estimer l'accord inter-annotateurs ?

Karën Fort[1,2] Claire François[1] Maha Ghribi[1]
(1) INIST / CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy
(2) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430 Villetaneuse
{karen.fort,claire.francois,maha.ghribi}@inist.fr

Résumé. L'objectif des travaux présentés dans cet article est l'évaluation de la qualité d'annotations manuelles de relations de renommage de gènes dans des résumés scientifiques, annotations qui présentent la caractéristique d'être très dispersées. Pour cela, nous avons calculé et comparé les coefficients les plus communément utilisés, entre autres κ (Cohen, 1960) et π (Scott, 1955), et avons analysé dans quelle mesure ils sont adaptés à nos données. Nous avons également étudié les différentes pondérations applicables à ces coefficients permettant de calculer le κ pondéré (Cohen, 1968) et l'α (Krippendorff, 1980, 2004). Nous avons ainsi étudié le biais induit par la grande prévalence d'une catégorie et défini un mode de calcul des distances entre catégories reposant sur les annotations réalisées.

Abstract. This article details work aiming at evaluating the quality of the manual annotation of gene renaming relations in scientific abstracts, which generates sparse annotations. To evaluate these annotations, we computed and compared the results obtained using the commonly advocated inter-annotator agreement coefficients such as κ (Cohen, 1960) or π (Scott, 1955) and analyzed to which extent they are relevant for our data. We also studied the different weighting computations applicable to κ_ω (Cohen, 1968) and α (Krippendorff, 1980, 2004) and estimated the bias introduced by prevalence. We then define a way to compute distances between categories based on the produced annotations.

Mots-clés : Annotation manuelle, évaluation, accord inter-annotateurs.

Keywords: Manual annotation, evaluation, inter-annotator agreement.

1 Introduction

De nombreuses tâches de traitement automatique des langues (TAL) nécessitent une annotation manuelle en amont, afin, non seulement d'entraîner des outils automatiques, mais également de créer une référence pour l'évaluation. Or, s'il a été démontré qu'une annotation incohérente limite les capacités des moteurs entraînés à partir de celle-ci (Alex *et al.*, 2006; Reidsma & Carletta, 2008), la qualité de cette référence est rarement justifiée. En effet, peu de campagnes détaillent la manière dont celle-ci a été constituée. Lorsque des mesures d'accord inter-annotateurs sont données, elles le sont sous forme d'un coefficient qui est devenu un standard de fait : le "kappa" de Cohen (1960) ou de Carletta (1996), sans plus de précision.[1]

[1]Pour plus de détails sur les problèmes de terminologie liés aux "kappa", voir l'introduction de (Artstein & Poesio, 2008).

Di Eugenio & Glass (2004) ont montré la sensibilité de ces coefficients au biais entre annotateurs et au problème de prévalence. La discussion reste très ouverte concernant la représentativité de ces différents coefficients et la nécessité d'en présenter plusieurs. Artstein & Poesio (2008) ont réalisé un inventaire très intéressant des différents modes de calcul de l'accord inter-annotateurs et ont discuté l'utilisation de ces mesures dans les tâches d'annotation en TAL. Cependant, il reste difficile de savoir quel coefficient utiliser en fonction des caractéristiques des données. Nous présentons dans cet article l'évaluation que nous avons réalisé d'une campagne d'annotations manuelles en appliquant et comparant les différentes méthodes proposées par ces auteurs.

Nous décrivons brièvement la campagne d'annotation que nous avons menée, puis nous détaillons et analysons les résultats des accords inter-annotateurs obtenus en utilisant les coefficients simples S, π et κ. Nous appliquons ensuite les coefficients pondérés α et κ_ω, pour lesquels nous étudions le calcul des distances entre catégories. Enfin, nous discutons des résultats qui nous semblent nécessaires à présenter, en particulier dans des cas comme le nôtre de répartition non homogène des phénomènes langagiers.

2 Présentation de la campagne d'annotation

L'INIST a été chargé, dans le cadre du programme Quaero[2], de faire annoter par ses experts les relations de renommages de gènes de Bacillus Subtilis présentes dans un corpus de 1 843 textes courts (résumés), soit plus de 400 000 tokens (ici, chaînes de caractères séparées par des blancs), sélectionnés dans Medline par l'équipe MIG de l'INRA de Jouy en Josas[3], à l'aide de nomenclatures de noms de gènes et d'un ensemble de mots-clefs dénotant des renommages.

Cette annotation avait pour but, d'une part, de construire une base de données de couples de renommage de gènes de Bacillus Subtilis, et d'autre part d'entraîner et d'évaluer les outils d'extraction automatique des partenaires du programme. Au final, cette campagne aura permis de mettre au jour manuellement environ 200 couples de renommage, tel que : "*Inactivation of a previously unknown gene,* ***yqzB*** *(renamed* ***ccpN*** *for control catabolite protein of gluconeogenic genes [..]*".

Nous avons appliqué pour cela la méthodologie proposée par Fort *et al.* (2009), et avons calculé l'accord inter-annotateurs dès le début de la campagne, afin de mettre au jour les désaccords et de modifier le guide d'annotation en conséquence. Nous avons donc fait annoter par deux annotateurs experts (que nous noterons ici, A1 et A2) un même échantillon de 93 fichiers, correspondant à plus de 15 000 tokens, à partir duquel nous avons ensuite calculé l'accord inter et intra-annotateurs, tel que recommandé par Gut & Bayerl (2004). Les relations de renommage sont annotées ici très simplement, grâce à l'outil Cadixe[4], en sélectionnant le nom d'origine du gène (annoté *Former*), puis son nouveau nom (annoté *New*). Le reste du texte n'est pas annoté, mais doit être pris en compte dans le calcul de l'accord inter-annotateurs. Nous avons décidé de nommer cette "pseudo" catégorie *Rien*, en français, pour la différencier des catégories signifiantes *Former* et *New*.

Certains fichiers (plus d'un tiers d'entre eux) ne comportent pas de renommage du tout. Nous obtenons ainsi, en moyenne sur l'échantillon, 1 renommage par fichier. Les accords et désaccords ont été analysés qualitativement, ce qui nous a permis d'ajouter les cas non traités dans le guide d'annotation. Les résultats

[2] `http://quaero.org`

[3] Merci à cette équipe : `http://genome.jouy.inra.fr/bibliome/renommage/`

[4] `http://caderige.imag.fr/Cadixe/`

quantitatifs de cette annotation sont présentés sous forme de matrice de confusion dans le tableau 1. Les résultats en diagonale présentent le nombre d'éléments sur lesquels les deux annotateurs sont d'accord, pour chaque catégorie. Les autres cellules représentent les éléments pour lesquels les annotateurs ont choisi des catégories différentes (par exemple, 13 éléments ont été annotés *New* par A1 et *Former* par A2). Cette matrice révèle la prédominance de la catégorie *Rien* (plus de 99% du corpus) et montre ainsi que les éléments annotés sont très dispersés.

		A1			
		Former	New	Rien	Total
A2	Former	71	13	23	107
	New	8	69	15	92
	Rien	7	8	18 840	18 855
	Total	86	90	18 878	19 054

TAB. 1 – Matrice de confusion calculée à partir de l'ensemble des tokens

Pour construire cette matrice de confusion, nous avons choisi de prendre en compte le nombre total de tokens, soit 19 054. Partant du principe que les noms de gènes correspondent à un sous-ensemble bien spécifique de tokens dans les textes, nous pourrions également considérer l'ensemble total des occurrences de noms de gènes, soit 1 165, selon les résultats obtenus par MIG après application d'un dictionnaire de noms de gènes de l'INRA. Néanmoins, ce choix nous semble discutable, d'une part parce que la fiabilité des résultats dépend de la complétude du dictionnaire, qui, étant donné la forte évolutivité du domaine, ne peut être totale, et d'autre part, parce que cela revient à négliger le fait que les annotateurs doivent souvent lire tout le texte pour prendre des décisions, le renommage n'étant parfois attesté qu'à la fin du texte.

Nous verrons dans la section 3.2 que cette décision a un impact non nul sur les résultats. Il est donc fondamental de justifier ce type de choix lorsque l'on donne des résultats d'accord inter-annotateurs.

3 Évaluation à l'aide de coefficients simples

Dans la suite de l'article, nous utiliserons les notations et les formules de Artstein & Poesio (2008) concernant les mesures d'accord inter-annotateurs. Les calculs seront réalisés par défaut à partir du tableau 1.

3.1 Calcul des coefficients simples

La mesure la plus évidente d'accord inter-annotateurs est l'accord observé (A_o), qui correspond à la proportion d'éléments sur lesquels les annotateurs sont d'accord, autrement dit, le nombre total d'éléments pour lesquels il y a accord, divisé par le nombre total d'éléments, ici :

$$A_o = \frac{71 + 69 + 18840}{19054} = 0,996116$$

Le résultat est extrêmement élevé, mais il ne prend pas en compte l'accord attendu (*expected agreement*, A_e), c'est-à-dire la possibilité que les annotateurs classent un élément quelconque dans une même catégorie par hasard.

Pour analyser nos résultats nous utilisons donc ici les coefficients permettant de prendre en compte le hasard, décrits par Artstein & Poesio (2008) : S (Bennett *et al.*, 1954), κ (Cohen, 1960) et π (Scott, 1955), qui sont tous les trois obtenus à partir de la formule suivante, dans laquelle seul l'accord attendu (A_e) diffère selon le coefficient :

$$S, \kappa, \pi = \frac{A_o - A_e}{1 - A_e}$$

La différence entre ces coefficients réside dans la manière de calculer l'accord attendu en fonction des hypothèses concernant le comportement des annotateurs dans le cas d'une annotation des éléments au hasard. S suppose que les annotations réalisées au hasard suivent une distribution uniforme dans les différentes catégories (ici, trois), l'accord attendu est donc calculé de la façon suivante :

$$A_e^S = \frac{1}{3} = 0,333333$$

$$S = 0,99417$$

Le biais le plus important de cette mesure est qu'elle est directement corrélée au nombre de catégories. Par conséquent, plus le nombre de catégories est élevé, plus l'accord attendu est faible, ce qu'il est en général, sa valeur maximale étant de $0,5$ ($\frac{1}{2}$) pour deux catégories.

Le coefficient π (Scott, 1955), appelé également K dans (Siegel & Castellan, 1988) ou $kappa$ dans (Carletta, 1996), considère lui aussi que les distributions réalisées par les annotateurs par hasard sont équivalentes, mais il suppose que la répartition des éléments entre catégories n'est pas homogène et qu'elle peut être estimée par la répartition moyenne réalisée par les annotateurs. L'accord attendu est donc calculé de la façon suivante :

$$A_e^\pi = \frac{((\frac{86+107}{2})^2 + (\frac{90+92}{2})^2 + (\frac{18878+18855}{2})^2)}{19054^2} = 0,980464$$

$$\pi = 0,8012$$

Le coefficient κ (Cohen, 1960) suppose lui dans sa modélisation du hasard que la répartition des éléments entre catégories peut être différente pour chaque annotateur. Dans ce cas, la probabilité pour qu'un élément soit assigné dans une catégorie est le produit de la probabilité que chaque annotateur l'assigne dans cette catégorie. L'accord attendu est donc calculé de la façon suivante :

$$A_e^\kappa = \frac{(86 \times 107) + (90 \times 92) + (18878 \times 18855)}{19054^2} = 0,980463$$

$$\kappa = 0,80121$$

3.2 Analyse des résultats

En comparant les 3 coefficients obtenus, nous observons que S (0,99417) est à peine plus faible que l'accord observé (0,996116), tandis que π (0,8012) et κ (0,80121) sont très proches, tout en étant plus faibles que A_o et S. La valeur relative de ces coefficients est conforme à l'ordre $S \geqslant \pi$ et $\pi \leqslant \kappa$ décrit par Artstein & Poesio (2008). La valeur élevée de S montre que les éléments sont annotés selon une certaine logique. Pour un accord observé constant, le coefficient S ne dépend que du nombre de catégories, il n'est donc pas sensible à la répartition des éléments dans les catégories, au contraire de π et κ (Di Eugenio & Glass,

2004). Ces auteurs montrent que lorsque les catégories sont disproportionnées, en dépit d'un fort accord sur la catégorie prédominante, les coefficients π et κ sont très sensibles aux désaccords sur les catégories minoritaires. Les coefficients de type κ sont interprétés comme étant corrects à partir de 0,67 (Krippendorff, 1980), κ et π sont donc ici très satisfaisants, ce qui nous rassure quant à l'accord obtenu dans les deux catégories minoritaires. Par ailleurs, κ et π sont très proches, ce qui, selon Di Eugenio & Glass (2004) est très courant, et signifie que nos données montrent peu de biais dû aux annotateurs, puisque, dans le cas de deux annotateurs, cela reflète des distributions marginales similaires (Artstein & Poesio, 2008).

Nos résultats sont donc élevés et montrent peu de biais. Ils nous semblent pourtant peu sûrs, puisqu'ils mettent sur le même plan des catégories très hétérogènes, deux minoritaires mais signifiantes (*Former* et *New*), et une non signifiante (*Rien*) très majoritaire. Notre problème est donc de nous assurer que ces coefficients calculés sur les trois catégories reflètent un accord significatif sur les deux catégories signifiantes *Former* et *New*.

Une première preuve de l'influence de ce déséquilibre apparaît en examinant l'évolution de ces coefficients en fonction de la référence choisie pour définir la catégorie *Rien*. En effet, si nous choisissons non plus le nombre total de tokens, mais le nombre d'occurrences de noms de gènes (1 165), à partir de la matrice de confusion du tableau 2, nous obtenons $S = 0,90472$, $\pi = 0,77557$ et $\kappa = 0,77571$. Ces trois coefficients ont des valeurs inférieures aux précédentes et présentent un écart constant entre eux. Même si la répartition des éléments et le comportement des annotateurs semblent constants, la taille de la catégorie *Rien* influe donc sur la valeur définitive des accord inter-annotateurs.

		A1			
		Former	New	Rien	Total Noms gènes
A2	Former	71	13	23	107
	New	8	69	15	92
	Rien	7	8	951	966
	Total Noms Gènes	86	90	989	1 165

TAB. 2 – Matrice de confusion calculée à partir des noms de gènes

Une seconde méthode pour estimer dans quelle mesure sa très forte prévalence entraîne un biais dans le calcul des accords inter-annotateurs est de considérer uniquement les catégories *Former* et *New* et calculer la matrice de confusion correspondante (tableau 3).

		A1		
		Former	New	Total
A2	Former	71	13	84
	New	8	69	77
	Total	79	82	161

TAB. 3 – Matrice de confusion sans la catégorie *Rien*

A partir de ce tableau et des formules présentées en section 3.1, nous obtenons $\pi = 0,7390$ et $\kappa = 0,73934$. Ces valeurs sont inférieures aux coefficients π et κ obtenus à partir de la matrice de confusion complète (tableau 1), ou de la matrice de confusion obtenue à partir de l'ensemble des occurrences de noms de

gènes (tableau 2). La catégorie *Rien* fait l'objet d'un accord très important et présente une très forte prévalence, ce qui semble avoir induit une surestimation de ces coefficients, surestimation dont l'importance dépend de la taille de cette catégorie.

Laignelet & Rioult (2009), confrontés à la même disproportion entre catégories dans leur campagne d'annotation, se sont appuyés sur une suggestion de Hripcsak & Heitjan (2002) et ont utilisé le coefficient R (Finn, 1970) proposé dans le logiciel R[5]. Le coefficient R est calculé selon la formule suivante :

$$R = 1 - \frac{Variance\ observee}{Variance\ attendue}$$

la variance observée étant la moyenne des variances sur les éléments annotés et la variance attendue étant la variance de la distribution uniforme discrète à n catégories (ci-dessous $nb\ categories$), soit[6] :

$$Variance\ attendue = \frac{(nb\ categories)^2 - 1}{12}$$

Dans notre cas, nous obtenons $R = 0,994$. Cette valeur proche de S (0,99417) peut être expliquée par le fait que ce coefficient modélise le hasard comme S, en considérant une distribution uniforme des catégories et n'est donc pas plus sensible que S à la répartition des éléments dans les catégories. Notre conclusion rejoint de ce point de vue l'opinion de Ron Artstein, lorsqu'il dit : "*R is similar to Krippendorff's alpha except that it assumes a uniform distribution as its model of chance annotation ; R is to alpha like S is to Scott's pi, and the same criticisms apply*." (R. Artstein, communication personnelle, 4 décembre 2009). Le coefficient R de Finn n'apporte pas plus que S dans des cas de dispersion des annotations et donc de dissymétrie des catégories.

4 Évaluation utilisant des coefficients pondérés

Selon Artstein & Poesio (2008), π et κ ont pour défaut de traiter tous les désaccords de la même manière et seuls des coefficients pondérés permettent de donner plus d'importance à certains désaccords.

4.1 Calcul des coefficients κ_ω et α

Artstein & Poesio (2008) détaillent deux coefficients pondérés : la version pondérée de κ, κ_ω (Cohen, 1968) et l'α (Krippendorff, 1980, 2004). Ces deux coefficients prennent pour base le désaccord entre annotateurs et utilisent une distance entre les catégories décrivant à quel point deux catégories sont distinctes l'une de l'autre. On trouve dans (Artstein & Poesio, 2008) une discussion sur la définition de cette distance en fonction du type d'annotation. Cette distance permet entre autres de traiter des annotations de structures complexes en introduisant plusieurs valeurs de distance entre annotations. Cette méthode présente l'inconvénient de complexifier l'interprétation des résultats.

[5]http ://www.r-project.org/

[6]Finn (1970) ne détaille pas le calcul de cette variance attendue, mais on le trouve dans les sources de la librairie *irr* du logiciel R. Pour une explication plus approfondie, voir : `http://mathworld.wolfram.com/DiscreteUniformDistribution.html`.

Dans notre cas, nous avons 2 catégories signifiantes *Former* et *New* et une non signifiante, *Rien*. Nous considérons donc qu'il est plus important d'identifier les couples de noms de gènes que de déterminer l'antériorité d'un nom par rapport à l'autre. Par conséquent, la distance entre *Former* et *New* devrait être moindre que celle entre ceux-ci et *Rien*. Si nous faisons l'hypothèse qu'elle est deux fois moindre, nous obtenons le tableau de distances entre catégories suivant (dans l'intervalle [0,1]) :

	Former	New	Rien
Former	0	0,5	1
New	0,5	0	1
Rien	1	1	0

TAB. 4 – Tableau de distances estimées entre catégories

Les coefficients pondérés κ_ω et α sont calculés à partir de la formule suivante :

$$\kappa_\omega, \alpha = 1 - \frac{D_0}{D_e}$$

où D_0 représente le désaccord observé entre les annotateurs et D_e représente le désaccord attendu (*expected*), autrement dit, si l'affectation est réalisée au hasard. Le désaccord attendu de κ_ω et d'α suit la même logique que κ et π respectivement, et inclut la notion de distance entre catégories. Il est à noter que si toutes les catégories sont parfaitement distinctes, nous obtenons $\alpha = \pi$ et $\kappa_\omega = \kappa$. Nous obtenons, à partir des distances du tableau 4, $\alpha = 0,8292$ et $\kappa_\omega = 0,8291$. Ces valeurs plus élevées que π et κ montrent que la pondération a fait diminuer le désaccord et augmenter légèrement l'accord inter-annotateurs.

4.2 Calcul des distances entre catégories à partir de la matrice de confusion

Pour pondérer l'accord inter-annotateurs, les distances entre catégories sont définies à partir de connaissances préalables sur la tâche d'annotation. En parallèle, il nous semble utile de les évaluer également en fonction de la difficulté qu'ont les annotateurs à répartir les éléments entre les catégories et de confronter ces deux approches. Pour ce calcul, nous utilisons la matrice de confusion du tableau 1.

Nous considérons que deux catégories sont distinctes s'il y a peu de chance d'erreur de classement entre elles. Plus précisément, soient deux catégories C_1 et C_2 appartenant à l'ensemble des catégories considérées, $P(C_2|C_1)$ représente la probabilité qu'un annotateur affecte un élément à la catégorie C_2 sachant que le deuxième annotateur l'affecte à la catégorie C_1 et elle se calcule de la façon suivante :

$$P(C_2|C_1) = \frac{n_{1C_1,2C_2} + n_{2C_1,1C_2}}{n_{C_1}}$$

avec $n_{1C_1,2C_2}$ représentant le nombre d'éléments classés par l'annotateur 1 dans la catégorie C_1 alors que l'annotateur 2 les a classés dans la catégorie C_2 ; n_{C_1} représente la somme des éléments classés dans la catégorie C_1 par les deux annotateurs.

Quand cette probabilité est faible, la catégorie C_2 est peu similaire à la catégorie C_1 et le risque d'obtenir une classification différente est faible. Nous avons ainsi, selon les données du tableau 1 :

$$P(New|Former) = \frac{13 + 8}{107 + 86} = 0,108808$$

↙	Former	New	Rien
Former	0,735751	0,108808	0,155440
New	0,115385	0,758242	0,126374
Rien	0,000795	0,000609	0,998595

TAB. 5 – Tableau de probabilités

Dans le tableau 5, qui présente les valeurs de probabilité calculées pour notre cas d'application, la diagonale permet d'estimer l'accord entre annotateurs pour chaque catégorie. Il est très important pour la catégorie *Rien* (99 % d'accord) et plus faible pour les catégories *Former* et *New* (73 et 75 % d'accord respectivement). Les autres cellules du tableau permettent d'estimer le désaccord entre annotateurs, catégorie par catégorie. Ces probabilités sont très faibles, les catégories sont donc peu similaires. Nous observons également que ces probabilités sont asymétriques. Les valeurs $P(Former|Rien)$ et $P(New|Rien)$ sont très faibles (<1‰), le risque d'affecter un élément à la catégorie *Former* ou à la catégorie *New* sachant qu'il est déjà affecté à *Rien* est donc quasiment nul. Inversement, le risque d'affecter un élément à la catégorie *Rien* alors qu'il est déjà dans la catégorie *Former* ou dans la catégorie *New* est plus élevé (15 % et 12 % respectivement).

Nous utilisons ces probabilités dans le calcul des distances entre catégories selon le principe suivant : $d(C1, C1) = 0$, et pour tout C_1 différent de C_2, $d(C_1, C_2) = 1 - P(C_1|C_2)$.

Les probabilités n'étant pas symétriques, cette formule ne peut pas être utilisée telle quelle. En utilisant les mêmes hypothèses sur les distributions des annotations que pour la définition des coefficients, nous proposons deux transformations. Premièrement, le coefficient α suppose que, dans le cas d'une annotation au hasard, les annotateurs réalisent des distributions équivalentes. Nous définissons donc la distance associée comme étant la moyenne des distances orientées (calculées à partir du tableau 5) selon la formule suivante :

$$d_\alpha(C_1, C_2) = \frac{(1 - P(C_2|C_1)) + (1 - P(C_1|C_2))}{2}$$

Ce qui donne, dans notre cas :

$$d_\alpha(Former, New) = \frac{(1 - \frac{13+8}{107+86}) + (1 - \frac{13+8}{90+92})}{2} = 0,887904$$

Le κ_ω, quant à lui, suppose que les annotateurs procèdent de manière différente, aussi nous calculons la distance associée comme étant le produit des distances orientées (calculées à partir du tableau 5) selon la formule suivante :

$$d_{\kappa_\omega}(C_1, C_2) = (1 - P(C_2|C_1)) \times (1 - P(C_1|C_2))$$

Dans notre cas, nous obtenons donc :

$$d_{\kappa_\omega}(Former, New) = (1 - \frac{13+8}{107+86}) \times (1 - \frac{13+8}{90+92}) = 0,788362$$

Dans le tableau 6, nous observons que $d(Former, New)$ est inférieure dans les deux cas à $d(Former, Rien)$ et $d(New, Rien)$, ce qui semblerait confirmer que *Former* et *New* sont plus proches entre elles que de *Rien*. Cependant, ces valeurs sont nettement supérieures aux valeurs estimées dans le tableau 4, la distinction entre ces deux catégories s'avère donc dans les faits moins problématique que ce que nous avions craint.

	d_α	d_{κ_ω}
d(Former,New)	0,887904	0,788362
d(Former,Rien)	0,921882	0,843888
d(New,Rien)	0,936508	0,873094

TAB. 6 – Tableau des distances entre catégories calculées à partir des données

5 Conclusion

A partir d'une campagne d'annotation, nous avons analysé différents modes de calcul pour estimer l'accord inter-annotateurs. La particularité de cette campagne est le caractère très dispersé des annotations dans les textes qui induit un biais lié à la grande prévalence des tokens non annotés. La matrice de confusion synthétise parfaitement cette information. Le tableau des distances calculées entre catégories montre que toutes les catégories sont bien distinctes, même les deux catégories minoritaires mais signifiantes *Former* et *New*. Bien que les coefficients π, κ, α et κ_w soient très sensibles à ce biais de prévalence, ils restent satisfaisants dans notre cas, indiquant ainsi un bon accord pour ces deux catégories. En première approximation, nous pouvons estimer ce biais en comparant les résultats obtenus avec les matrices de confusion suivantes : complète, réduite aux noms de gènes et réduite aux deux catégories signifiantes.

En complément des coefficients, et quand le mode opératoire de l'annotation le permet, le premier résultat à présenter est à notre avis la matrice de confusion accompagnée d'explications précises sur les choix effectués. Nous rejoignons Hripcsak & Heitjan (2002) lorsqu'ils écrivent "*showing the two-by-two contingency table with its marginal totals is probably as informative as any measure*". En effet, cette matrice résume les informations quantitatives obtenues dans une campagne d'annotation et permet entre autres d'avoir rapidement une idée des problèmes de prévalence et de biais entre annotateurs.

Le tableau des distances entre catégories calculées à partir des résultats d'annotation est également très riche en information car il permet d'analyser le risque réel d'erreur entre certaines catégories et de le confronter aux distances définies à priori en fonction des connaissances du domaine. Ces différents tableaux permettent de comprendre au mieux les caractéristiques de la campagne d'annotation et d'interpréter les différents coefficients obtenus selon leur mode de calcul.

De nouvelles campagnes d'annotation sont en cours et devraient nous permettre de tester les différents coefficients ainsi que la reproductibilité de nos propositions dans des cas aussi variés que l'annotation de brevets en pharmacologie (entités nommées, termes) ou des commentaires de matchs de football (entités nommées, relations diverses). Ces campagnes devraient également nous permettre d'élargir la réflexion à des annotations réalisées par plus de deux annotateurs. Dans ce dernier cas, le tableau des distances entre catégories calculées à partir des résultats d'annotation permettra de réaliser une bonne synthèse des problèmes existants pour distinguer les catégories.

Remerciements

Ce travail a été réalisé en partie dans le cadre du programme Quaero [7], financé par OSEO, agence nationale de valorisation de la recherche. Nous en remercions les participants, en particulier l'équipe MIG de

[7] http://www.quaero.org

l'INRA. Nous remercions également F. Tisserand et B. Taliercio, les annotateurs experts de l'INIST, ainsi que Ron Artstein, pour son intérêt et ses réponses détaillées.

Références

ALEX B., NISSIM M. & GROVER C. (2006). The impact of annotation on the performance of protein tagging in biomedical text. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)*, p. 595–600, Gène, Italie.

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.

BENNETT E. M., ALPERT R. & C.GOLDSTEIN A. (1954). Communications through limited questioning. *Public Opinion Quarterly*, **18(3)**, 303–308.

CARLETTA J. (1996). Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, **22**, 249–254.

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

COHEN J. (1968). Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**(4), 213–220.

DI EUGENIO B. & GLASS M. (2004). The kappa statistic : a second look. *Computational Linguistics*, **30**(1), 95–101.

FINN R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, **30**, 71–76.

FORT K., EHRMANN M. & NAZARENKO A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ? In *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles 2009 Traitement Automatique des Langues Naturelles 2009*, Senlis, France.

GUT U. & BAYERL P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Proceedings of Speech Prosody*, p. 565–568, Nara, Japon.

HRIPCSAK G. & HEITJAN D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, **35**(2), 99–110.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Its Methodology*, chapter 12. Sage : Beverly Hills, CA.

KRIPPENDORFF K. (2004). *Content Analysis : An Introduction to Its Methodology, second edition*, chapter 11. Sage : Thousand Oaks, CA.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

REIDSMA D. & 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.

SIEGEL S. & CASTELLAN N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York : McGraw-Hill, 2nd edition.