Description d'un protocole d'évaluation automatique comme alternative à l'évaluation humaine. Application à la validation de relations syntaxiques induites

Nicolas Béchet

Description d'un protocole d'évaluation automatique comme alternative à l'évaluation humaine.

Application à la validation de relations syntaxiques induites

Nicolas Béchet*

*LIRMM - UMR 5506, CNRS - Univ. Montpellier 2 - 34392 Montpellier Cedex 5 - France
bechet@lirmm.fr

Résumé. Cet article propose un protocole d'évaluation afin de valider la qualité d'approches, visant à évaluer et à ordonner automatiquement des relations syntaxiques dites induites. Les approches évaluées se fondent sur l'interrogation d'un moteur de recherche sur le Web. Les résultats du moteur de recherche sont alors couplés avec diverses mesures statistiques : l'information mutuelle, l'information mutuelle au cube, le coefficient de Dice et la fréquence, ou popularité.
Le protocole d'évaluation propose d'utiliser deux corpus, le premier de test et le second de validation, appartenant tous deux au même domaine. Le principe est de retrouver dans le second corpus les relations syntaxiques induites, non présentes originalement dans le premier corpus. Il est alors étudié la taille minimale du corpus d'évaluation afin de permettre une évaluation pertinente.

1 Introduction

Le domaine du traitement automatique des langues (TAL) regroupe de nombreux champs de recherche, pouvant rendre difficile la tâche d'évaluation, de par leurs diversités. Les ouvrages de Dale et al. (2000) et Mitkov (2003) établissent un survol de ces différents champs de recherche parmi lesquelles l'acquisition d'informations lexicales, l'analyse syntaxique, l'extraction d'informations, la traduction automatique, le résumé automatique ou encore des systèmes d'aide à la rédaction. Pospescu-Belis (2007) propose de classer ces sous domaines dans quatre catégories : les systèmes d'analyse ou d'annotation (1), les systèmes de génération ou de synthèse (2), les systèmes combinant l'analyse et la génération (3) et enfin des systèmes interactifs (4), faisant intervenir un (ou plusieurs) utilisateur(s) humain(s). Ces catégories se fondent sur les entrées sorties linguistiques des systèmes de TAL ainsi que sur l'interaction humaine (ou non). Notons qu'une majorité des champs de recherche du TAL peut être classée dans la première catégorie impliquant du contenu linguistique en entrée du système. L'évaluation de tels systèmes consiste le plus souvent à se référer à un (ou plusieurs) intervenant(s) humain(s). Celui-ci se voit attribuer la même tâche que le système TAL. Alors sont utilisées des métriques d'évaluation comme le coefficient *kappa* (Cohen (1960)), permettant de confondre les résultats de juges humains avec ceux du modèle. D'autres métriques assez largement utilisées en TAL sont le rappel et la précision (Salton et Mcgill (1986)), ainsi que leur moyenne harmonique, la f-mesure. Ces mesures permettent, parmi un ensemble de candidats, d'identifier les pertinents. Ces mesures ne sont pas toujours adaptées à une tâche comme l'évaluation de la qualité du classement d'un ensemble de candidat, auxquelles on préférera par exemple l'utilisation des courbes ROC.
Cet article présente un protocole d'évaluation, permettant de valider de manière automatique la qualité d'approches, afin de se passer d'une évaluation humaine. Ces approches permettent d'ordonner des relations syntaxiques dites induites. De telles relations ne sont pas présentes initialement dans un corpus et mais sont acquises à partir de celui-ci. Nous proposons dans un premier temps de définir les relations syntaxiques induites en montrant de quelles manières elles sont acquises, et pour quelles applications elles peuvent être utiles (section 2). Après avoir montré pourquoi il était nécessaire d'effectuer une validation qualitative de ces relations, nous décrierons les différents processus de validations utilisés (section 3). Nous obtenons alors en sortie d'une validation, une liste ordonnée de relations syntaxiques par qualité. Une manière de mesurer la qualité du classement proposé serait alors de faire valider ce classement par un expert. Une telle validation serait très longue et fastidieuse. Il est de plus difficile de trouver des experts d'un domaine, acceptant d'effectuer ce type d'évaluation. Il est alors proposé un protocole d'évaluation automatique se fondant sur l'utilisation de deux corpus d'un même domaine (4). Ce protocole va être ensuite discuté en étant notamment comparé à une validation manuelle triviale.

2 Les relations syntaxiques induites

2.1 Acquisition des relations syntaxiques induites

La génération de relations syntaxiques induites à partir d'un corpus requière une extraction des relations syntaxiques classiques. Ainsi, l'analyseur SYGFRAN (Chauché (1984)) est utilisé afin d'extraire les relations classiques d'un corpus (dans cet article, nous nous intéresserons uniquement aux relations syntaxiques Verbe-Objet). La proximité sémantique des verbes extrais est ensuite étudiée avec la mesure d'Asium (Faure (2000)). Cette mesure, dont la formule est donnée ci-dessous, considère comme proche deux verbes possédant un certain nombre d'objets en commun.
Soit p et q, deux verbes avec leurs objets respectifs $p_1, ..., p_n$ et $q_1, ..., q_n$ illustrés sur la figure 1. $NbOccCom_p(q_i)$

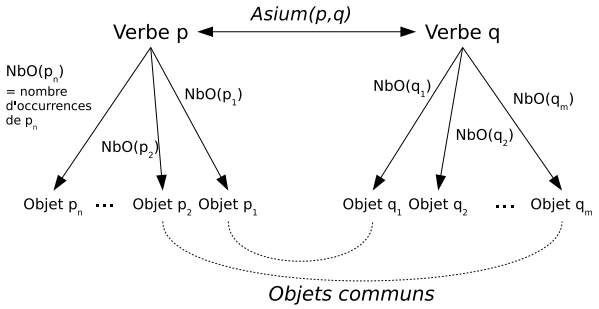

FIG. 1 – *Mesure d'Asium entre les verbes p et q*

représente le nombre d'occurrences des objets q_i en relation avec le verbe q qui sont aussi des objets du verbe p (objets communs). $NbOcc(q_i)$ représente le nombre d'occurrences des objets q_i. La mesure d'Asium est alors définie de la manière suivante :

$$Asium(p,q) = \frac{log_{Asium}(\sum NbOccCom_q(p_i)) + log_{Asium}(\sum NbOccCom_p(q_i))}{log_{Asium}(\sum NbOcc(p_i)) + log_{Asium}(\sum NbOcc(q_i))}$$

Avec $log_{Asium}(x)$ valant :
- pour $x = 0$, $log_{Asium}(x) = 0$
- sinon $log_{Asium}(x) = log(x) + 1$

Un score proche de 1 obtenu avec la mesure d'Asium implique une importante proximité sémantique.
Une fois l'ensemble de la proximité des verbes du corpus mesuré, le rassemblement des objets jugés proches est alors effectué. Dès lors, deux types d'objets peuvent être utilisés afin de définir une relation syntaxique : les objets **communs**, objets étant originalement présents dans le corpus (*argent* et *vêtement* sur la figure 2), et les objets **complémentaires**, objets induits de relations syntaxiques existantes décrits dans Faure et Nédellec (1998); Béchet et al. (2009) (objet *avertissement* pour le verbe *offrir* et *cadeau* pour le verbe *donner* sur la figure 2). La relation syntaxique ainsi formée avec un objet complémentaire (*former bateau* sur la figure 2) est alors appelée une **relation syntaxique induite** par opposition à une **relation syntaxique classique**.

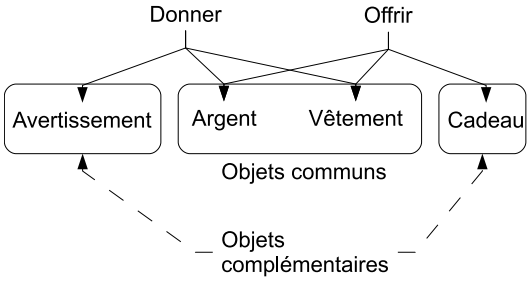

FIG. 2 – *Objets communs et complémentaires des verbes "Donner" et "Offrir".*

2.2 Pourquoi utiliser des relations syntaxiques induites ?

Les relations syntaxiques classiques peuvent être utilisées afin d'enrichir un corpus comme avec l'approche ExpLSA (Béchet et al. (2008)). Celle-ci propose d'enrichir un corpus afin d'améliorer les résultats de classification automatique de textes. Après l'extraction des relations syntaxiques induites et classiques, le principe est de compléter les termes du corpus par les objets provenant des relations syntaxiques (classiques ou induites). En effet, nous pouvons enrichir le corpus avec uniquement les objets communs, ou alors en utilisant les communs et les complémentaires. L'intérêt des relations syntaxiques induites, réside dans le fait qu'elles ne sont pas originalement présentes dans le corpus. Ainsi, le corpus enrichi avec ExpLSA se voit apporter des informations nouvelles pouvant permettre de faciliter une tâche de classification ou de clustering.

Outre l'expansion de corpus, les relations syntaxiques induites peuvent également contribuer à la représentation d'un ensemble de connaissances d'un domaine donné. Montrons par exemple comment des relations syntaxiques induites peuvent permettre de contribuer à l'acquisition ou l'enrichissement d'ontologies qui peuvent être définies comme la représentation de connaissances d'un domaine sous forme de concepts hiérarchisés. Un réseau de dépendance se fondant sur des ressources syntaxiques partagées, notamment des relations de dépendance syntaxique autours des verbes (dans notre cas les relations Verbe-Objet) peuvent servir de support dans un processus d'acquisition ontologique comme dans Bourigault (2002). Ainsi, un concept peut être constitué d'objets provenant de relations syntaxiques Verbe-Objet. Les relations induites peuvent alors apporter une information supplémentaire, impossible à acquérir avec des relations syntaxiques classiques. En considérant les objets de la figure 2, nous pourrions constituer le concept comme suit :
- Avec les objets communs (*argent* et *vêtement*)
- Avec les objets communs et l'objet complémentaire du premier verbe (*argent*, *vêtement* et *cadeau*)
- Avec les objets communs et l'objet complémentaire du second verbe (*argent*, *vêtement* et *avertissement*)
- Avec les objets communs et les objets complémentaires (*argent*, *vêtement* et *avertissement*, *cadeau*)

3 Validation des relations syntaxiques induites par le Web

La qualité des relations syntaxiques induites extraient à partir d'un corpus tel que présenté dans la section précédente est discutable. En effet, si deux verbes partagent un même objet et que l'on considère notre corpus comme bien écrit, il va de soit que les relations syntaxiques produites avec chacun des verbes sont cohérentes, car elles proviennent directement du corpus. Néanmoins, l'idée d'une relation syntaxique induite est qu'elle n'existe pas originalement dans le corpus. Alors, rien ne garantit la cohérence d'une telle relation syntaxique, même si les deux verbes partagent par ailleurs un nombre important d'objets communs. Prenons par exemple la relation *offrir avertissement* issue de la figure 2. Celle-ci n'est manifestent pas cohérente contrairement à l'autre relations induite issue de cette même figure (*donner cadeau*).
Il apparaît alors nécessaire de mesurer la cohérence des relations syntaxiques induites. On pourra par exemple utiliser les relations syntaxiques validées afin de créer des classes conceptuelles, ou ontologies en ne sélectionnant que les premières relations une fois ordonnées, ou utiliser les meilleures relations induites afin d'enrichir un corpus en évitant ainsi l'ajout important de bruit.
Ainsi, il est présenté dans cette section une manière de valider la cohérence des relations syntaxiques induites en se fondant sur le Web. La validation automatique des relations syntaxiques propose de mesurer la dépendance entre verbe et objet d'une relation induite. Il est employé pour cela un moteur de recherche en utilisant une API (*http ://api.search.yahoo.com*). Une requête est ainsi soumise au moteur de recherche. Des mesures statistiques sont finalement employées afin de proposer un classement des relations syntaxiques. Diverses définitions préalables sont nécessaires afin de permettre d'adapter les mesures statistiques à l'étude de la cohérence des relations syntaxiques.

3.1 Définitions

Nous définissons tout d'abord, la fonction $nb(X)$, comme étant le nombre de pages retournées par le moteur de recherche en réponse à la requête X. Définissons également o et v comme étant respectivement l'objet et le verbe évalué. Ainsi, $nb(o)$ va retourner le nombre de pages trouvées pour l'objet o, ceci reflétant la popularité de l'objet o sur le Web. Définissons enfin $nb_{max}(v, o)$ et $nb_{sum}(v, o)$ comme suit :

$nb_{max}(v, o) = max(nb(v$ **un** $o), nb(v$ **une** $o), nb(v$ **le** $o), nb(v$ **la** $o), nb(v$ **l'** $o))$
et
$nb_{sum}(v, o) = nb(v$ **un** $o) + nb(v$ **une** $o) + nb(v$ **le** $o) + nb(v$ **la** $o) + nb(v$ **l'** $o)$

avec $nb(v$ **un** $o)$ qui est le nombre de pages retournées par le moteur de recherche Yahoo pour la relation syntaxique 'v un o". Les différentes mesures statistiques utilisées pour ordonner les relations syntaxiques sont présentées ci-dessous.

3.2 L'information mutuelle

Une des mesures les plus couramment utilisées en recherche d'information afin d'établir un classement est l'Information Mutuelle (IM) (Church et Hanks (1990)) définie comme suit :

$$IM(x,y) \quad = \quad \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

$P(x,y)$ peut alors être vu comme la probabilité des réponses retournées par le moteur de recherche Yahoo pour la relation syntaxique v, o. Cette mesure vise à faire ressortir les co-occurrences les plus rares et les plus spécifiques (Daille (1996); Thanopoulos et al. (2002)). Appliquée au contexte de la validation des relations syntaxiques induites, la formule 1 va devenir [1] :

$$IM(v,o) \quad = \quad \frac{nb(v,o)}{nb(v)nb(o)} \tag{2}$$

Avec $nb(v,o)$, étant soit $nb_{max}(v,o)$ ou bien $nb_{sum}(v,o)$ suivant que l'on utilise le max ou la somme. Par ailleurs, le log_2 a été supprimé, ne modifiant pas le rang obtenu.

3.3 L'information mutuelle au cube

L'information mutuelle au cube est une information empirique fondée sur l'information mutuelle, qui accentue l'impacte des co-occurrences fréquentes, ce qui n'est pas le cas avec l'information mutuelle originale Daille (1994). Cette mesure est définie ainsi :

$$IM^3(x,y) \quad = \quad \log_2 \frac{P(x,y)^3}{P(x)P(y)} \tag{3}$$

Qui adaptée à la mesure de la cohérence des relations syntaxiques devient :

$$IM^3(v,o) \quad = \quad \frac{nb(v,o)^3}{nb(v)nb(o)} \tag{4}$$

3.4 Le coefficient de Dice

Une mesure également intéressante en terme d'évaluation de qualité est le coefficient de Dice (Smadja et al. (1996)) défini comme suit :

$$Dice(x,y) \quad = \quad \frac{2 \times P(x,y)}{P(x) + P(y)} \tag{5}$$

Qui devient :

$$Dice(x,y) \quad = \quad \frac{2 \times nb(x,y)}{nb(x) + nb(y)} \tag{6}$$

Les différentes mesures statistiques vont permettre d'obtenir un classement des relations syntaxiques en fonction de leur pertinence sur le Web. Huit listes de relations syntaxiques triées vont alors être retournées en utilisant les approches suivantes :
- La fréquence d'utilisation de la relation syntaxique sur le Web
- L'information mutuelle
- L'information mutuelle au cube
- Le coefficient de Dice

Ces quartes mesures peuvent être obtenues en utilisant le maximum $nb_{max}(v,o)$ ou la somme $nb_{sum}(v,o)$ nous donnant ainsi les huit listes triées. Il est maintenant nécessaire d'établir un protocole d'évaluation afin de déterminer la qualité des fonctions de rangs proposées, et par la même de la validation Web.

[1] en écrivant $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x,y) = \frac{nb(x,y)}{nb\_total}$ avec $x = v$ et $y = o$

4 Un protocole d'évaluation automatique

4.1 Définition du protocole

La validation manuelle semble être intuitivement la solution la plus adaptée afin de mesurer la qualité de la validation des relations syntaxiques induites par le Web. Néanmoins il est très difficile de trouver des experts afin d'effectuer une telle tâche. En effet, cela nécessiterait un travail fastidieux. Les experts se verraient proposer une quantité non négligeable de relations syntaxiques à expertiser. Il est par conséquent défini dans cet article un protocole expérimental automatique.

Le principe de ce protocole est d'utiliser deux corpus. Un premier, de test (appelé par la suite *corpus T*) duquel ont été extraites les relations syntaxiques qui vont ensuite être ordonnées par les différentes approches présentées section 3. Ce corpus contient 8 948 articles (16,5 Mo) en français et il est extrait du site Web d'informations de Yahoo (*http ://fr.news.yahoo.com/*). Il a été obtenu 60 460 relations syntaxiques induites. Un second corpus, également en français, est alors utilisé comme référence afin de valider les approches (appelé corpus V). Ce second corpus a la particularité d'être beaucoup plus important que le corpus T, contenant plus de 60 000 articles (125 Mo) issus du corpus du quotidien *Le Monde*. Il est de plus du même domaine, actualité avec un style journalistique.
Il est alors proposé de juger une relation syntaxique induite, créé à partir du corpus T, comme pertinente si celle ci est

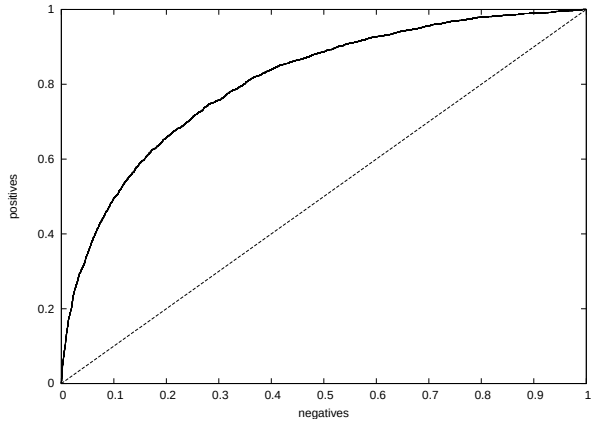

FIG. 3 – *Exemple de courbe ROC*

retrouvée dans le corpus V (comme relation syntaxique dite *classique* par opposition à *induite*). Concrètement, si une relation induite est retrouvée dans le *corpus V*, on la qualifiera de **positive**. Dans le cas contraire, elle sera jugée non pertinente et sera donc qualifiée de **négative**. L'intérêt de cette validation réside dans le fait qu'elle permette de mesurer de manière automatique la qualité des approches proposées, et ceci pour un très grand nombre de relations syntaxiques. Notons que les relations jugées négatives peuvent être de faux négatifs. En effet, une relation qui n'a pas été retrouvée dans le *corpus V* n'est pas pour autant non pertinente. Une fois la notion de positif et négatif définie, nous pouvons utiliser une approche couramment utilisée dans la littérature afin de mesurer la qualité d'une fonction de rang : les courbes ROC.

La méthode des courbes ROC (Receiver Operating Characteristic), détaillée par Ferri et al. (2002), fut utilisée à l'origine dans le domaine du traitement du signal. Cette méthode est fréquemment employée en médecine afin d'évaluer automatiquement la validité d'un diagnostique de tests. On trouve en abscisse des axes représentant une courbe ROC le taux de faux positifs (dans notre cas, le taux de relations syntaxiques induites non pertinentes, soit les relations non retrouvées dans le *corpus V*) et l'on trouve en ordonnée le taux de vrais positifs (dans notre cas les relations pertinentes, soit celles existantes dans le *corpus V*). La surface sous la courbe ROC ainsi créée est appelée AUC (*Area Under the Curve*). Un des avantages de l'utilisation des courbes ROC réside dans leur résistance à la non parité de la répartition du nombre d'exemples positifs et négatifs.
Une courbe ROC représentée par une diagonale correspond à un système où les relations syntaxiques ont une distribution aléatoire, la progression du taux de vrais positifs est accompagnée par la dégradation du taux de faux positifs. La courbe 3 est un exemple de courbe ROC, avec en diagonale, une distribution aléatoire. Considérons le cas d'une validation de relations syntaxiques induites. Si toutes les relations sont positives (ou pertinentes), l'AUC vaudrait 1, ce qui signifie avoir toutes les relations pertinentes en début de liste, donc ordonnées de manière optimale.

4.2 Résultats et Discussions

	AUC
maximum	0,813
somme	0,814
IM max	0,760
IM somme	0,763

	AUC
IM3 max	0,810
IM3 somme	0,812
Dice max	0,800
Dice Somme	0,802

FIG. 4 – *Résultats obtenus pour le corpus de Validation*

La figure 4 présente les AUC obtenues pour les différentes approches définies section 3 en utilisant le corpus de validation dans sa totalité. Il en ressort que l'utilisation de la *somme* plutôt que le *maximum* pour la fonction $nb(v, o)$ donne de meilleurs résultats, mais assez peu significatifs. Par ailleurs la *somme* obtient de meilleurs scores que toutes les autres approches confondues, elle semble donc être l'approche la plus adaptée pour ordonner de manière automatique les relations syntaxiques induites. Toutefois, ces résultats devront être confirmés avec d'autres expérimentations, permettant alors de conclure sur la meilleure approche à utiliser.

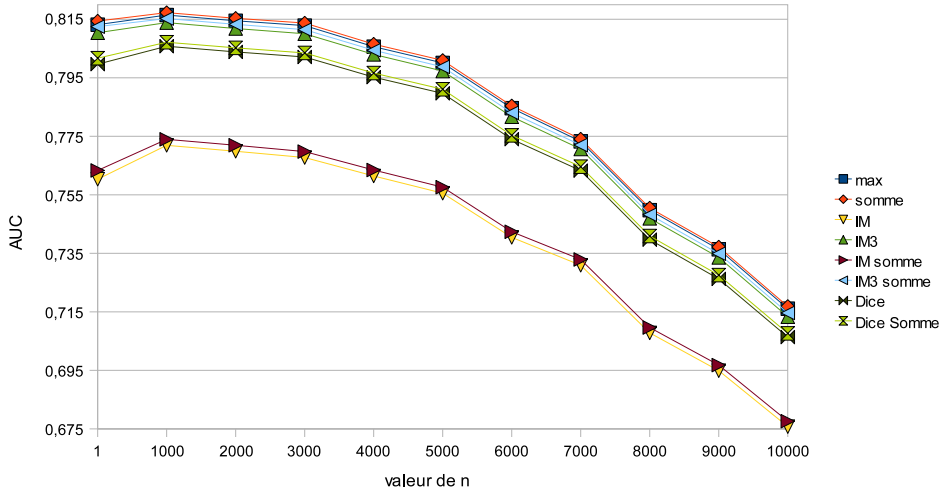

FIG. 5 – *AUC obtenues pour différentes valeurs de* n

Valeur de n	***Nb positifs***
Entier (n=1)	9080
1/1000	23,5
1/2000	11,8
1/3000	7,9
1/4000	5,9
1/5000	4,7

Valeur de n	***Nb positifs***
1/6000	3,9
1/7000	3,3
1/8000	2,9
1/9000	2,6
1/10000	2,3

FIG. 6 – *Nombre de relations couvertes en fonction de la proportion de corpus considérée*

Il est maintenant proposé d'étudier quelle doit être la taille minimum du corpus de test, afin d'évaluer correctement les approches présentées. Ainsi, le corpus de test original va être divisé en n parties. Chaque section de corpus va ensuite servir à valider les approches utilisant la validation Web, et ce en effectuant une validation croisée. Ainsi, pour un $n = 1000$, 1000 expérimentations vont être effectuées afin de calculer une AUC moyenne correspondant à un corpus d'évaluation d'une taille d'environ $60000/1000$ soit 60 articles en moyenne. Notons qu'un $n = 1$ revient à considérer la totalité du corpus de validation.

La figure 5 présente les différentes AUC obtenues pour des valeurs de n variants de 1 à 10 000, soit un corpus original divisé par 10 000. Ces résultats nous montrent que la taille du corpus peut être réduite de 5000, tout en conservant des

résultats équivalents à ceux obtenus avec le corpus de validation dans son ensemble. Pour un n variant de 1 à 5 000, nous observons en effet une variation des AUC de l'ordre de plus ou moins $0,01$ autour de l'AUC obtenue pour le corpus entier ($n = 1$), ce qui reste du même ordre. Pour une valeur de n supérieure à 5 000, les résultats se dégradent, quelque soit la mesure statistique utilisée, pour atteindre des AUC inférieures de $0,1$ points par rapport au corpus entier. Par exemple avec la *somme*, l'AUC passe de $0,81$ pour le corpus entier à $0,72$ pour un $n = 10000$. Cette baisse des AUC peut s'expliquer par une limite théorique du à un trop faible nombre de relations couvertes (nombre de relations syntaxiques induites retrouvées dans le corpus) pour des valeurs de n trop grandes. En effet, un nombre trop faible de relations couvertes reflète un manque de finesse dans les AUC résultantes.

Le tableau 6 présente le nombre de relations couvertes en fonction de la taille du corpus considéré (la taille du corpus correspond à celle du corpus de validation divisée par n). Les AUC se dégradent pour une valeur de n supérieure à 5 000. Il est constaté qu'avec moins de 4 relations syntaxiques couvertes par un corpus, les résultats donnés par le protocole utilisé sont biaisés. Néanmoins, cela signifie également qu'avec seulement 5 relations syntaxiques retrouvées dans un corpus, les AUC obtenues avec le protocole sont équivalents à ceux obtenus avec la totalité du corpus. La figure 7 confirme que

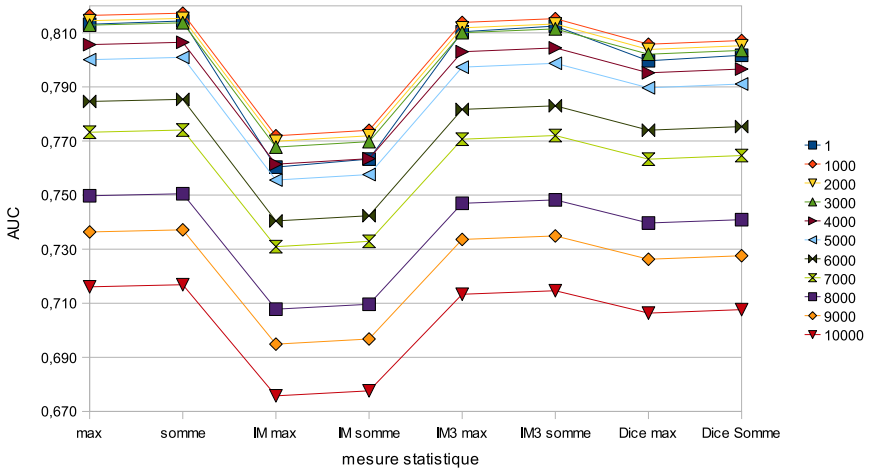

FIG. 7 – *AUC obtenues en fonction des mesures statistiques*

les AUC obtenues, pour un n compris entre 1 et 5 000 sont du même ordre. Cette figure présente les AUC en fonction de la mesure statistique utilisée. Elle permet de constater que l'allure des courbes, suivant la valeur de n, reste la même. Cela signifie que quelque soit la valeur de n (entre 1 et 10 000 ici) on retrouve le même le classement de qualité des mesures statistiques. Par exemple, la mesure *somme* reste la meilleure, quelque soit n.

Le protocole d'évaluation proposé permet donc d'utiliser un petit corpus comme corpus de validation. Il reste cependant important de mesurer la robustesse du protocole vis à vis des relations syntaxiques jugées non pertinentes. Cela correspond à des relations syntaxiques qui n'ont pas été retrouvées dans le corpus de validation. Certaines de ces relations peuvent être des faux positifs, un corpus d'un même domaine ne contenant pas nécessairement toutes les relations induites provenant d'un autre corpus. Une évaluation humaine triviale a donc été effectuée afin de mesurer la quantité de faux négatifs, permettant également de confronter une évaluation automatique à une évaluation humaine. Le protocole suivant est décrit ci-dessous.
Cent relations syntaxiques induites parmi les 60 460 issus du corpus de test préalablement triées par la validation Web avec la *somme*, sont extraites de manière homogène. Ces 100 relations ont alors été évaluées manuellement par un expert, suivant le critère : *La relation syntaxique est elle sémantiquement cohérente ?* avec seulement deux réponses possibles, oui ou non. En d'autre terme, le verbe v peut-il être rencontré avec l'objet o. Avec cette réponse binaire, et l'homogénéité de l'échantillon de relations syntaxiques testées, nous pouvons calculer une AUC moyenne correspondant à une extrapolation des résultats que l'on obtiendrait avec la totalité des relations syntaxiques induites.
Le tableau 8 présente les AUC obtenues avec cette évaluation manuelle ainsi qu'avec l'évaluation automatique. Ils sont également comparés à l'AUC obtenue avec l'ensemble des relations syntaxiques. Les AUC obtenues pour l'approche automatique avec les 100 relations et avec l'ensemble des relations syntaxiques sont très proches. Cela permet de confirmer l'homogénéité de l'échantillon sélectionné. La comparaison avec l'approche manuelle reste assez proche également avec une variation de $0,06$ points. Les faux positifs sont donc existants avec le protocole d'évaluation automatique mais restent

tolérables. Les résultats obtenus avec la validation manuelle confirme également la qualité de la validation Web avec la *somme*.

Nb de relations	100		60460
Type d'évaluation	manuelle	automatique	automatique
AUC	0,88	0,82	0,81

FIG. 8 – *Comparaison de l'évaluation manuelle et automatique*

5 Conclusion

Cet article a présenté un protocole d'évaluation automatique, permettant de se passer d'une évaluation couteuse et fastidieuse que serait une évaluation manuelle. Ce protocole propose de mesurer la qualité d'approches visant à mesurer la cohérence et à ordonner des relations syntaxiques dites induite. Les relations syntaxiques induites sont des relations qui ne sont pas originalement présentes dans un corpus, nécessitants une évaluation de leur cohérence. Le protocole d'évaluation automatique présenté se fonde sur l'utilisation d'un second corpus d'une thématique proche du corpus étudié. Une relation syntaxique induite est alors jugée positive si celle-ci est retrouvée dans le second corpus. Il est alors discuté la taille minimale ainsi que le nombre minimal de relations syntaxiques qui doivent être retrouvées dans le second corpus.
Les expérimentations menées ont permis de montrer que l'on pouvait réduire de manière considérable la taille du corpus utilisé afin de valider les relations syntaxiques (jusqu'à 5 000 fois). La validation humaine proposée a permis de constater le biais occasionné par les faux négatifs, mais ce biais reste faible. Une évaluation manuelle, plus complète, avec notamment plusieurs experts, devra néanmoins être effectuée pour confirmer ces résultats. Des corpus d'autres domaines devrons être également être traités, afin de tester le protocole et la validation de relations syntaxiques induites. Enfin, il sera proposé d'utiliser d'autres relations syntaxiques que les Verbe-Objet, dont notamment les relations Sujet-Verbe.

Références

Béchet, N., M. Roche, et J. Chauché (2008). How the ExpLSA approach impacts the document classification tasks. In *Proceedings of the International Conference on Digital Information Management, ICDIM'08*, University of East London, London, United Kingdom,.

Béchet, N., M. Roche, et J. Chauché (2009). Comment valider automatiquement des relations syntaxiques induites. In *EGC'09, à paraître*.

Bourigault, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN, Nancy*, pp. 75–84.

Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours. In *Proceedings of Coling, Standford University, California*, pp. 11–15.

Church, K. W. et P. Hanks (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, Volume 16, pp. 22–29.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*(1), 37.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.

Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *P. Resnik and J. Klavans (eds). The Balancing Act : Combining Symbolic and Statistical Approaches to Language, MIT Press*, pp. 49–66.

Dale, R., H. L. Somers, et H. Moisl (Eds.) (2000). *Handbook of Natural Language Processing*. New York, NY, USA : Marcel Dekker, Inc.

Faure, D. (2000). *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph. D. thesis, Université Paris-Sud.

Faure, D. et C. Nédellec (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In P. Velardi (Ed.), *LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications*, Granada Espagne, pp. 5–12.

Ferri, C., P. Flach, et J. Hernandez-Orallo (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of ICML'02*, pp. 139–146.

Mitkov, R. (2003). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*. Oxford University Press.

Pospescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en tal. *TAL (Traitement Automatique des Langues) 47*(2).

Salton, G. et M. J. Mcgill (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.

Smadja, F., K. R. McKeown, et V. Hatzivassiloglou (1996). Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics 22*(1), 1–38.

Thanopoulos, A., N. Fakotakis, et G. Kokkianakis (2002). Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of LREC'02*, Volume 2, pp. 620–625.

Summary

This paper presents an evaluation protocol in order to validate the quality of approaches. These approaches allow to rank automatically syntactic relations Verb-Object called induced, by querying a Web search engine. Then, some statistic measures are applied: The mutual information, the cubic mutual information, the Dice's coefficient and the frequency. Two corpora are needed to apply the evaluation protocol. The first one is a test corpus, and the second one is a validation corpus. The idea is to recover induced syntactic relations, which are not originally present in the first corpus, in the second corpus. Finally, the minimum size of the validation corpus is discussed.