Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ?

Adam Lion-Bouton, Loïc Grobol, Jean-Yves Antoine, Sylvie Billot, Anais Anais Lefeuvre-Halftermeyer

▶ **To cite this version:**

Adam Lion-Bouton, Loïc Grobol, Jean-Yves Antoine, Sylvie Billot, Anais Anais Lefeuvre-Halftermeyer. Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ?. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), Jun 2020, Nancy, France. pp.10-18. hal-02750222v4

HAL Id: hal-02750222
https://hal.science/hal-02750222v4

Submitted on 23 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ?

Adam Lion-Bouton[1], Loïc Grobol[2, 3], Jean-Yves Antoine[1], Sylvie Billot[4], Anaïs Lefeuvre-Halftermeyer[4]
(1) LIFAT, ICVL, Université de Tours, 41000 Blois,France
(2) LLF, 8, Rue Albert Einstein 75013 Paris, France
(3) Lattice, 1 Rue Maurice Arnoux, 92120 Montrouge, France
(4) LIFO, ICVL, Université d'Orléans, 45000 Orléans, France
lion.adam.otman@gmail.com, loic.grobol@ens.psl.eu,
Jean-Yves.Antoine@univ-tours.fr,
{Sylvie.Billot, Anais.Halftermeyer}@univ-orleans.fr

RÉSUMÉ

Cet article présente un travail qui consiste à étudier si les scores les plus utilisés pour l'évaluation de la résolution des coréférences constituent des métriques de similarité normalisées. En adoptant une démarche purement expérimentale, nous avons vérifié si les scores MUC, B^3, CEAF, BLANC, LEA et le meta-score CoNLL respectent les bonnes propriétés qui définissent une telle métrique. Notre étude montre que seul le score $CEAF_m$ est potentiellement une métrique de similarité normalisée.

ABSTRACT

Do the standard scores of evaluation of coreference resolution constitute metrics ?

This paper presents an experimental research that investigates whether the most commonly used scores for evaluating the resolution of co-references constitute normalized similarity metrics. Considering systematic test suites, we verified whether the MUC, B3, CEAF, BLANC, LEA and CoNLL scores comply with the formal properties that define such a metric. Our study shows that only the CEAFm score is potentially a normalized similarity metric.

MOTS-CLÉS : coréférence, évaluation, métrique de similarité, MUC, B^3, CEAF, BLANC, LEA, CoNLL.

KEYWORDS: coreference, evaluation, similarity metric, MUC, B^3, CEAF, BLANC, LEA, CoNLL.

1 Introduction

Disposant de ressources linguistiques d'envergure servant aussi bien à l'apprentissage de modèles qu'à leur test, le TAL a recours à des mesures quantitatives pour évaluer ses avancées et étudier la pertinence d'approches alternatives. Généralement, l'évaluation consiste à comparer les réponses du système à une référence idéale (appelée GOLD standard), à l'aide d'un score bien choisi. Ce rôle central de l'évaluation quantitative interroge notre discipline d'un point de vue méthodologique et déontologique. En effet, si une réflexion éthique en recherche doit concerner les productions de cette recherche, elle doit également concerner la manière dont ces recherches sont conduites. On sait en effet que par delà le modèle poppérien de réfutabilité, la recherche est une activité sociale (Latour &

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 10–18.
HAL : hal-02750222.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

Woolgar, 1986) dont il est bon d'interroger les pratiques. Lors de la mise en place d'une campagne d'évaluation, de nombreuses questions méritent ainsi d'être posées :

— quelles sont la qualité et la représentativité des données de test utilisées ?
— comment s'assurer de la significativité statistique des différences de performances observées ?
— comment interpréter les résultats de l'évaluation en termes de qualité perçue par l'utilisateur ?
— quels biais peuvent altérer l'interprétation des résultats obtenus ?
— à quelle question répond réellement un score d'évaluation donné, et quel score est le plus appropriée dans une situation donnée ?

Le choix d'un score pertinent est assez évident dans les cas simples. Si l'on considère une tâche de catégorisation telle que l'attribution d'une valence émotionnelle (positif, négatif, aucune) à un tour de parole, le recours à des scores standards pour la classification (rappel, précision, F-mesure) est assez naturel. Notons toutefois que retenir la F-mesure (non pondérée) comme juge de paix pose déjà question : certaines applications peuvent privilégier un besoin en rappel ou au contraire en précision, plutôt que de viser l'optimisation conjointe des deux, telle qu'évaluée par la F-mesure.

Le choix d'un score adéquat est au contraire bien moins évident dans le cas des systèmes end-to-end de TAL, et ce pour deux raisons principales :

— d'une part, il n'est pas toujours évident de définir une référence qui propose une vérité terrain incontestable, de même qu'il n'est pas facile d'évaluer qualitativement si un écart à la référence est plus grave qu'un autre. Si nous prenons ainsi l'exemple de la traduction automatique, comment caractériser une bonne traduction ?
— d'autre part, les tâches complexes ajoutent souvent au problème de la catégorisation celui de la segmentation en unité, ou de l'alignement entre unités. Là encore, il est alors délicat de savoir quelle segmentation est plus contestable qu'une autre

Face à ces difficultés conceptuelles, la communauté scientifique a souvent proposé plusieurs fonctions d'évaluation alternatives pour une tâche donnée. Des études ont été conduites pour étudier les biais et limites de chaque proposition, sans arriver le plus souvent à démêler cet écheveau. Cette situation était prévisible, car souvent, chaque score répond à un objectif d'évaluation particulier. Ce constat se retrouve précisément dans la problématique étudiée dans cet article : la résolution des coréférences. De multiples fonctions d'évaluation ont été proposées, qui répondent chacune à des manières peu conciliables d'appréhender la segmentation d'un texte en chaînes de références. Ceci conduit à des résultats sensiblement différents suivant le score utilisée (Moosavi & Strube, 2016)

Le travail présenté dans cet article vise à clarifier cette situation en vérifiant quelles propriétés formelles vérifient les scores les plus utilisés. Plus précisément, nous proposons de vérifier quels scores peuvent être considérés comme des métriques de similarité normalisées. Ces contraintes formelles ne constituent pas le seul critère de choix d'un score d'évaluation. Elles ne rendent pas compte des biais statistiques et ne parlent que de manière indirecte des liens qui existent entre évaluation objective et jugement humain. Si l'on cherche à mesurer la qualité d'un système, il nous semble toutefois important d'étudier les propriétés formelles de l'outil de mesure utilisé.

Dans un premier temps, nous présentons la tâche de résolution des coréférences, puis les différents scores qui ont été proposés pour l'évaluer. La section suivante décrit les propriétés formelles qui définissent une métrique de similarité normalisée. Nous présentons ensuite notre protocole d'étude. La présentation des résultats est enfin l'occasion de donner un tableau synthétique de l'intérêt de ces scores d'évaluation très répandus.

2 L'évaluation pour la coréférence

On définit une **mention** comme un mot ou groupe de mots référant à un élément de l'univers du discours appelé **entité**, et une **chaîne de coréférences** comme l'ensemble des mentions référant à une même entité. On dit de deux mentions référant à une même entité qu'elles sont **coréférentes** et on nomme **singleton** toute entité qui ne compte qu'une unique mention. Par définition, une mention réfère nécessairement à une unique entité. L'ensemble des chaînes de coréférences d'un document forme donc une partition de l'ensemble des mentions de ce document, dont les chaînes de coréférences sont les parties. Considérons l'énoncé suivant :

> Elles$_1$ tournent, tournent les aiguilles$_1$, ne m'$_2$ dis plus qu'c'est passager
>
> Furax Barbarossa - Fin 2006

Dans cet exemple ainsi que dans les suivants, chaque mention sera indicée par le numéro de l'entité auquel elle réfère. Cet exemple est composé de trois mentions et des deux entités suivantes :

1 : Elles , les aiguilles
2 : m' (singleton)

Étant donné un document et l'ensemble des mentions le composant [1], la tâche de résolution des chaînes de coréférences consiste à produire une partition de mentions (que l'on appellera SYS) aussi proche de la vérité (appelé GOLD) que faire se peut. L'évaluation consiste ainsi à attribuer un score à la partition SYS en fonction de sa similarité avec la partition GOLD. Or, il s'avère qu'une partition peut être caractérisée de plusieurs manières, induisant chacune différentes méthodes de comparaison. Deux mentions coréférentes peuvent par exemple être vues respectivement comme appartenant à un même ensemble, référant à une même entité, ou encore comme étant **liées** par un *lien* (relation) de coréférence. Il existe alors plusieurs manières d'évaluer les différences (et donc les similarités) entre le GOLD et le SYS suivant la vision de la coréférence adoptée.

Considérons par exemple les deux partitions suivantes — respectivement GOLD et SYS :

> J'$_1$ suis bien chez moi$_1$ mais j'$_1$ comprends l'gitan$_2$ quand il$_2$ m'$_1$ rappelle qu' il$_2$ est libre
> J'$_1$ suis bien chez moi$_1$ mais j'$_3$ comprends l'gitan$_2$ quand il$_2$ m'$_3$ rappelle qu' il$_2$ est libre
>
> Furax Barbarrossa - Fin 2012

Une première approche représente une chaîne de coréférence comme la suite des liens entre les mentions qui la composent, en respectant leur *succession* dans le texte. Ainsi, sur la figure 1.a, on décrit l'entité 1 du GOLD par les trois liens J'$_1$ — moi$_1$, moi$_1$ — j'$_1$, j'$_1$ — m'$_1$, tandis que les entités 1 et 3 détectées dans SYS se décrivent par les deux liens J'$_1$ — moi$_1$, j'$_3$ — m'$_3$. Dans cette approche, ajouter au SYS le lien moi — j' fusionnerait les entités 1 et 3 du SYS, résultant en un SYS égal au GOLD. En somme, GOLD ne diffère de SYS que par une seule opération.

On pourrait au contraire considérer qu'une chaine de coréférences est décrite par l'ensemble de *tous* les liens possibles entre les mentions qui la composent (figure 1.b), indépendamment de leur ordre d'occurrence. Dans cette approche, l'entité 1 du GOLD est décrite par les six liens J'$_1$ — moi'$_1$, J'$_1$ — j'$_1$, J'$_1$ — m'$_1$, moi$_1$ — j'$_1$, moi$_1$ — m'$_1$, j'$_1$ — m'$_1$ et les entités 1 et 3 du SYS

1. On ne s'intéresse ici pas à la tâche consistant à retrouver les mentions. De même, notre étude ne concerne que la résolution des coréférences, et non celle des anaphores pour lesquelles l'identification d'une référence est nécessaire pour l'interprétation d'une mention, sans qu'il y ait nécessairement identité de référence

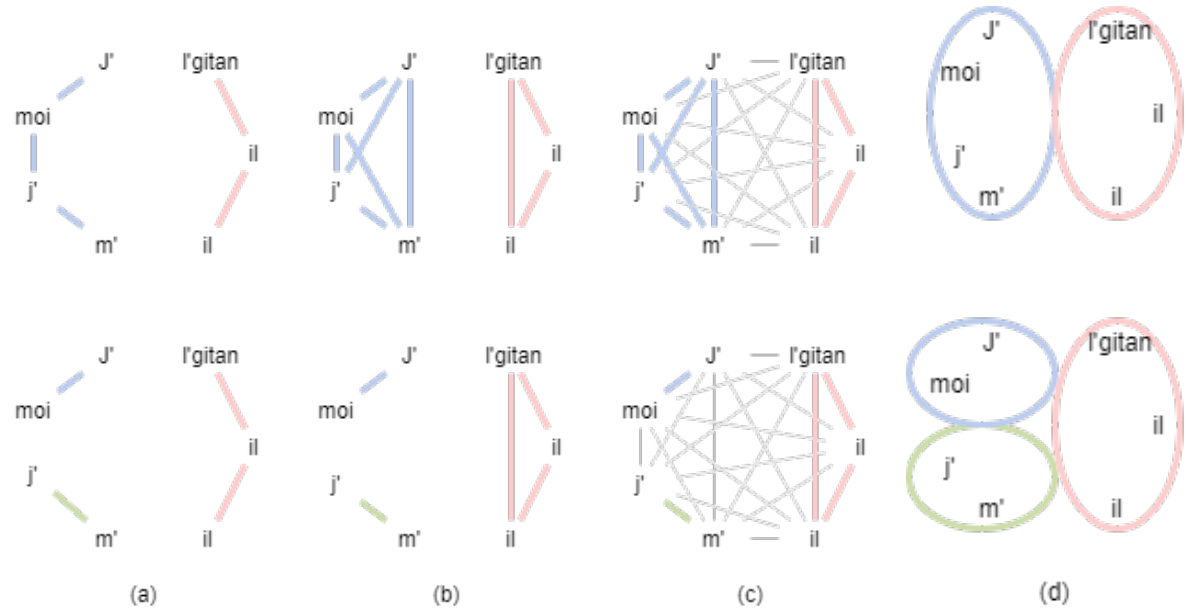

FIGURE 1 – Représentations des partitions de nos exemple GOLD (en haut) et SYS (en bas) suivant différentes approches de la coréférence : (a) chaînes de maillons (b) ensemble de liens de coréférence (c) ensemble de liens de coréférence, en gras et de non coréférence, en traits fin (d) vision ensembliste

par les deux liens J'$_1$ — moi$_1$, j'$_3$ — m'$_3$. Dans ce cas, il faudrait ajouter au SYS les quatre liens suivants pour retrouver le GOLD : J' — j', J' — m', moi — j', moi — m' pour que SYS et GOLD soient égaux.

Ce raisonnement peut être poursuivi dans une approche qui se focalise sur une vision ensembliste (figure 1 ;d). Dans ce cas, il faut déplacer les mentions j' et m' dans l'ensemble représentant l'entité 1 pour identifier GOLD et SYS. Soit donc deux opérations. Arrêtons là la démonstration pour simplement constater que plusieurs scores ont été proposés, qui répondent à ces différentes approches :

— **MUC**, proposé par Vilain *et al.* (1995) comme fonction d'évaluation de la tâche MUC6 (MUC Consortium, 1995). MUC considère le nombre minimum de liens nécessaires pour décrire les deux partitions et calcule une similarité en fonction de la proportion de ces liens correctement résolus (i.e. qui sont communs au GOLD et au SYS).
— **B^3** proposé par Bagga & Baldwin (1998) s'intéresse, pour toutes les paires d'entités de GOLD et de SYS, au nombre de mentions correctement résolues.
— **CEAF**, proposé par Luo (2005), choisit le meilleur appariement entre les entités GOLD et SYS puis s'intéresse à la proportion de mentions correctement résolues, mais uniquement pour les entités ainsi appariées. La définition de meilleur appariement n'a pas une solution unique, Luo (2005) propose ainsi deux alternatives, l'une centrée sur les mentions, nommée **CEAF$_\mathbf{m}$**, l'autre centrée sur les entités et nommée **CEAF$_\mathbf{e}$**
— **BLANC** (Recasens & Hovy, 2011) considère les mentions non-coréférentes comme liées par un lien de non-coréférence (figure 1.c) et évalue la proportion de liens (de coréférences *et* de non-coréférence) correctement résolus en s'abstrayant de la notion d'entité.
— **LEA** (Moosavi & Strube, 2016) s'intéresse à la fraction de liens de coréférences bien résolus.

Aucun de ces scores n'est parvenu à faire l'unanimité. Face à cela, Denis & Baldridge (2009) proposent

MELA, une moyenne de MUC, B^3, et $CEAF_e$, employé par Pradhan *et al.* (2012) comme fonction d'évaluation de la tâche **CoNLL**-2012 (et pour cette raison souvent appelée «score CoNLL»).

Ces fonctions de scores sont définies par une évaluation de la proximité entre SYS et une unique partition GOLD. Dans une perspective d'interprétabilité, il est intéressant de les étudier plus généralement comme des estimations de la similarité entre deux partitions, sans nécessairement que l'une d'entre elles ait un rôle privilégié. Ainsi, il serait par exemple souhaitable que deux sorties systèmes proches donnent également des scores proches quand elles sont comparées au GOLD et, par extension, que deux systèmes donnant des sorties similaires obtiennent également des scores proches.

Dans une perspective de conception de systèmes, pouvoir estimer directement la proximité de sorties systèmes sans passer par une référence serait également souhaitable, par exemple pour étudier la sensibilité d'un système au variations dans un choix de paramètres. Dans la suite de cet article, nous étudions précisément les propriétés formelles de ces scores, non pas en termes d'adéquation avec une référence mais en terme de similarité entre deux partitions. Pour ce faire, nous nous intéressons en particulier aux propriétés définissant une métrique de similarité normalisée.

3 Métrique de similarité normalisée

La notion de métrique de distance fait l'objet d'une définition reconnue, à l'inverse de celle de similarité qui est le plus souvent ramenée au complément d'une distance (Deza & Deza, 2009). Pour répondre à cette carence formelle, Chen *et al.* (2009) dérive de la définition d'une métrique de distance celle d'une métrique de similarité : on appelle **métrique de similarité** toute fonction $s : X \times X \longrightarrow \mathbb{R}$ respectant les cinq propriétés suivantes pour tout $(a, b, c) \in X^3$:

$$s(a,b) = s(b,a) \quad \text{(symétrie)} \tag{1}$$
$$s(a,a) \geq 0 \tag{2}$$
$$s(a,b) \leq s(a,a) \tag{3}$$
$$s(a,b) + s(b,c) \leq s(b,b) + s(a,c) \quad \text{(inégalité triangulaire)} \tag{4}$$
$$s(a,a) = s(b,b) = s(a,b) \text{ si et seulement si } a = b \quad \text{(identité des indiscernables)} \tag{5}$$

On dira de plus que s est **normalisée** si

$$s(a,b) \leq 1 \tag{6}$$

Enfin, si s est une métrique de similarité normalisée telle que

$$s(a,a) = 1 \quad \text{(identité à 1)} \tag{7}$$
$$s(a,b) \geq 0 \quad \text{(positivité)} \tag{8}$$

alors $1 - s$ est une métrique de *distance* normalisée. Dans ce cas, les propriétés 2 à 5 peuvent être réécrites sous une forme plus naturelle :

$$s(a,a) = 1 \quad \text{(équivalent à la propriété 7)} \tag{9}$$
$$s(a,b) \leq 1 \quad \text{(équivalent à la propriété 6)} \tag{10}$$
$$s(a,b) + s(b,c) \leq 1 + s(a,c) \tag{11}$$
$$s(a,b) = 1 \Leftrightarrow a = b \tag{12}$$

Vérifier si un score est une métrique de similarité normalisée nous parait important. En effet, une métrique de similarité normalisée dispose de bonnes propriétés qui rendent intelligibles les écarts de performance qu'elle peut mesurer. Lorsque l'on compare les performances de différents systèmes, on opère des jugements cognitifs de similarité. À la suite de Shepard (1962), de multiples travaux en psychologie ont suggéré que ces jugements relèvent d'une évaluation de distance dans un espace métrique, ceci même si certains biais peuvent conduire à des violations de la contrainte de métricité sur des cas limites (Laub *et al.*, 2007). Ces violations n'empêchent pas de même Tversky & Gati (1982) de considérer que la propriété d'inégalité triangulaire favorise l'intelligibilité des comparaisons.

D'un point de vue pratique, le statut de métrique de similarité est préservé par plusieurs opérations élémentaires telles que la somme et le produit avec une autre métrique de similarité, ainsi que par l'application d'une fonction convexe positive strictement croissante (Chen *et al.*, 2009). Cette dernière opération permet un rééquilibrage du score pour obtenir une distribution plus uniforme des valeurs sur l'intervalle $[0, 1]$ et donc d'obtenir un score plus aisément interprétable, tandis que somme et produit sont utiles pour construire des scores d'évaluation combinés tel que CoNLL.

4 Protocole

Pour cette étude, nous avons défini un protocole de vérification des propriétés précédentes pour chacun des scores présentés plus haut. On adopte une démarche qui a pour objectif non pas de valider ces propriétés (ce qui ne peut se faire que formellement), mais de tenter de les falsifier à l'aide de jeux de tests bien choisis . Cette méthode a l'avantage de s'appliquer à toute fonction de score, que son fonctionnement interne soit connu ou pas. Elle s'applique par ailleurs à des fonctions de score qui ne seraient pas définies formellement, mais répondraient par exemple à des estimations statistiques. Notons que dans les cas de falsification, notre démarche a force de preuve analytique : nous sommes en effet en mesure de donner un contre-exemple.

La démarche employée tire parti de deux traits spécifiques à l'évaluation de la coréférence. Premièrement, les scores ne sont calculés que pour évaluer des partitions d'un même ensemble de mentions. Deuxièmement, la similarité entre deux partitions ne dépend que des relations définies par les partitions, et aucunement d'une éventuelle sémantique propre aux mentions. Ainsi, une propriété vérifiée pour toutes les partitions d'un ensemble de n éléments particuliers le sera également pour toutes les partitions de *tous* les ensembles à n éléments. Réciproquement, un contre-exemple valable pour un ensemble à n éléments particulier le sera pour tous les ensembles à n éléments.

$$s(\{\{1,2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1,2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$

FIGURE 2 – Exemple de la procédure de falsification

Suivant ce principe, pour chaque propriété — par exemple : (3) $(s(a,b) < s(a,a)$ — on génère toutes les partitions de l'ensemble des n premiers nombres entiers positifs, et on remplace chacune des variables dans la définition de la propriété — a et b pour la propriété 3 — par toutes les combinaisons de partitions de cet ensemble. Ainsi, si l'on considère $n = 2$, on obtient deux partitions : $\{\{1,2\}\}$ et $\{\{1\},\{2\}\}$. On évalue alors les cas présentés en figure 2.

On procède exhaustivement jusqu'à $n = 5$ dans le cas de la propriété 4 (soit environs 140 000 combinaisons à évaluer) et jusqu'à $n = 6$ pour les autres (environ 40 000 combinaisons par propriété).

5 Résultats

Les résultats des expérimentations sont résumés dans la table 1.

	1	2	3	4	5	6	7	8
MUC				X			X	
B^3				X				
$CEAF_m$								
$CEAF_e$				X				
CoNLL				X			X	
BLANC	X			X				
LEA				X				

TABLE 1 – Propriétés non-respectées par les scores : (X) si la propriété n'est pas respectée

Symétrie — On remarque tout d'abord que BLANC ne respecte pas la propriété 1 et n'est donc pas symétrique. Par exemple, on observe que :

$$BLANC(\{\{1,2,3\}\},\{\{1,2\},\{3\}\}) = 0{,}5 \neq 0{,}25 = BLANC(\{\{1,2\},\{3\}\},\{\{1,2,3\}\})$$

Nos tests montrent que BLANC est le plus souvent symétrique. La propriété n'est enfreinte que lorsqu'une des partitions considérées est composée uniquement de singletons ou d'une entité unique. BLANC est en effet défini par une formule générale, mais aussi par des cas particuliers dans ces situations. En l'occurrence, ces cas particuliers s'avèrent être la cause de la non-symétrie du score.

On pourrait penser que ces cas particuliers sont rares et peuvent être négligés. Il convient cependant de se rappeler que plus l'ensemble de mentions considéré est petit, plus la probabilité de rencontrer l'un de ces cas particuliers est élevée. Ainsi, 6 des 15 combinaisons de deux partitions créées sur un ensemble de trois mentions débouchent sur l'un de ces cas particuliers non-symétriques.

Identité à 1 — En second lieu, on remarque que ni MUC, ni CoNLL ne respectent la propriété 7, d'identité à 1 lorsqu'une partition est comparée à elle-même. Comme précédemment, cette propriété semble être généralement respectée et uniquement enfreinte dans le cas où la partition comparée à elle-même est composée uniquement de singletons.

MUC a été définie dans le cadre d'une tâche où la notion de singleton était ignorée. MUC n'a donc pas été conçue pour en tenir compte et vaut 0 dans le cas où l'une des partitions n'est composée que de singletons. CoNLL étant défini comme une moyenne de scores intégrant MUC, lorsque MUC est égal à 0, CoNLL ne peut nécessairement pas être supérieur à $\frac{2}{3}$.

Un score ne respectant pas l'identité à 1 a pour défaut de rendre l'interprétation de score plus complexe. Dans le cas présent, afin d'être correctement interprétés les scores MUC et CoNLL devraient être accompagnés d'une information supplémentaire, indiquant si l'une des partitions est composée uniquement de singleton.

Inégalté triangulaire — Enfin, on remarque que ni MUC, B^3, $CEAF_e$, CoNLL, BLANC, ni LEA ne respectent la propriété 4, aucun de ces scores ne respecte donc l'inégalité triangulaire. Contrairement à précédemment, cette propriété est falsifiée même en dehors de cas limites. Les trois exemples suivants falsifient respectivement l'inégalité triangulaire pour les scores MUC, B^3 et LEA ; les scores $CEAF_e$ et BLANC ; et le score CoNLL.

$$a = \{\{1,2,3\},\{4,5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2\},\{3\},\{4\},\{5\}\}$$
$$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2,3,5\},\{4\}\}$$
$$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4,5\}\}, c = \{\{1\},\{2\},\{3\},\{4,5\}\}$$

Un score ne respectant pas cette propriété pose de forts problèmes d'interprétation. Il est alors possible pour deux partitions SYS très similaires d'obtenir des scores très différents. L'inégalité triangulaire est ainsi la propriété donnant son coeur à la notion de métrique.

6 Conclusion

Dans cet article, nous nous sommes interrogés sur les propriétés formelles que vérifient les principaux scores utilisés pour évaluer la résolution de la coréférence. Nous avons vu que ces scores diffèrent par la vision qu'ils proposent de la coréférence Afin de comparer ces scores, nous avons considéré les bonnes propriétés qu'ils doivent vérifier afin d'être considérés comme des métriques de similarité normalisées. Nous avons alors cherché à falsifier par des jeux de test l'hypothèse selon laquelle les scores étudiés constituent (ou non) de telles métriques. Cette étude a prouvé que MUC, B^3, $CEAF_e$, CoNLL, BLANC et LEA ne sont pas des métriques alors que $CEAF_m$ semble être une métrique de similarité normalisée. Du moins, il ne nous a pas été possible de falsifier cette hypothèse.

$CEAF_m$ apparaît ainsi comme une métriques intéressante d'un point de vue formel, car son possible statut de métrique de similarité normalisée répond aux modèles cognitifs décrivant les processus de catégorisation, de comparaison et d'estimation de ressemblances réalisés par les humains. Par delà cette étude purement formelle, il est toutefois important de s'intéresser également aux biais statistiques (influence de la taille des entités, influence de la prévalence des singletons, etc.) qui peuvent influencer les scores d'évaluation obtenus par ces métriques. Nous sommes précisément en train d'étudier systématiquement ces types de biais, à l'aide d'une approche expérimentale sur jeux de tests. Nous étudions également le degré de corrélation qui existe entre scores d'évaluation et jugements humains de la qualité des systèmes (Holen, 2013). Ce tableau complet (propriétés formelles, tolérance aux biais, proximité avec une évaluation subjective) devrait alors permettre au concepteur de système ou de campagne d'évaluation de choisir, en toute connaissance, le score d'évaluation le plus adapté à ses besoins, ou à mieux connaître les avantages et faiblesses de chacun.

Remerciements

Ce travail a bénéficié du soutien du RTR DIAMS de la région Centre-Val-de-Loire.

Références

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, p. 563–566.

CHEN S., MA B. & ZHANG K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, **410**(24-25), 2365–2376.

DENIS P. & BALDRIDGE J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, **42**.

DEZA M. M. & DEZA E. (2009). Encyclopedia of distances. In *Encyclopedia of distances*, p. 1–583. Springer.

HOLEN G. I. (2013). Critical reflections on evaluation practices in coreference resolution. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, p. 1–7.

LATOUR B. & WOOLGAR S. (1986). *Laboratory Life : The Construction of Scientific Facts*, volume 80. Princeton University Press.

LAUB J., MÜLLER K.-R., WICHMANN F. A. & MACKE J. H. (2007). Inducing metric violations in human similarity judgements. In *Advances in neural information processing systems*, p. 777–784.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 25–32 : Association for Computational Linguistics.

MOOSAVI N. S. & STRUBE M. (2016). Which coreference evaluation metric do you trust ? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 632–642.

MUC CONSORTIUM (1995). Appendix D : Coreference Task Definition (v2.3). In *Sixth Message Understanding Conference*, Columbia, Maryland : Morgan Kaufmann.

PRADHAN S., MOSCHITTI A., XUE N., URYUPINA O. & ZHANG Y. (2012). Conll-2012 shared task : Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, p. 1–40 : Association for Computational Linguistics.

RECASENS M. & HOVY E. (2011). Blanc : Implementing the rand index for coreference evaluation. *Natural Language Engineering*, **17**(4), 485–510.

SHEPARD R. N. (1962). The analysis of proximities : multidimensional scaling with an unknown distance function. i. *Psychometrika*, **27**(2), 125–140.

TVERSKY A. & GATI I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, **89**(2), 123.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, p. 45–52 : Association for Computational Linguistics.