# Appariement d'entités nommées coréférentes : combinaisons de mesures de similarité par apprentissage supervisé

Erwan Moreau[1] François Yvon[2] Olivier Cappé[1]
(1) Institut Télécom ParisTech & LTCI CNRS
(2) Univ. Paris Sud & LIMSI CNRS
emoreau@enst.fr, yvon@limsi.fr, cappe@enst.fr

**Résumé.** L'appariement d'entités nommées consiste à regrouper les différentes formes sous lesquelles apparaît une entité. Pour cela, des mesures de similarité textuelle sont généralement utilisées. Nous proposons de combiner plusieurs mesures afin d'améliorer les performances de la tâche d'appariement. À l'aide d'expériences menées sur deux corpus, nous montrons la pertinence de l'apprentissage supervisé dans ce but, particulièrement avec l'algorithme C4.5.

**Abstract.** Matching named entities consists in grouping the different forms under which an entity may occur. Textual similarity measures are the usual tools for this task. We propose to combine several measures in order to improve the performance. We show the relevance of supervised learning in this objective through experiences with two corpora, especially in the case of the C4.5 algorithm.

**Mots-clés :** Entités nommées, Appariement, Mesures de similarité textuelle, Apprentissage supervisé.

**Keywords:** Named entities, Matching, Textual similarity measures, Supervised learning.

## 1 Introduction

La capacité à structurer de grandes quantités d'informations provenant de sources différentes est un enjeu technologique essentiel, et cette capacité réside notamment dans la possibilité de classer l'information par *entité*. Pour améliorer la recherche d'information, il est donc nécessaire de savoir reconnaître la même entité lorsqu'elle apparaît sous des formes différentes.

Nous traitons dans cet article du problème de l'identification de séquences de mots représentant une même entité nommée (EN), dans un cadre où l'on dispose des entités elles-même et du contexte dans lequel elles apparaissent. La difficulté porte sur les nombreuses variations textuelles possibles d'une EN : ces variations peuvent être volontaires et/ou naturelles (écriture différente selon la langue, abréviations ou extensions, surnoms, etc.) ou involontaires (erreurs typographiques ou orthographique, erreurs d'OCR, etc.). On s'intéresse ici particulièrement aux variations dues aux translittérations (traductions entre systèmes d'écriture différents). L'hypothèse de départ réside donc dans l'idée qu'un même référent conduit généralement à des formes "similaires", ainsi qu'à des ressemblances de leurs contextes d'occurrence.

L'appariement d'entités nommées coréférentes est notamment étudié dans la problématique du

liage d'enregistrements (*record linkage*), qui consiste à repérer deux enregistrements distincts représentant le même élément dans une base de données (pour la déduplication) ou dans deux bases différentes (fusion de bases) (Winkler, 1999; Bilenko *et al.*, 2003). Certains travaux, tels que (Cohen *et al.*, 2003; Christen, 2006), étudient plus spécifiquement l'appariement de noms de personnes. Dans (Freeman *et al.*, 2006), les problèmes de translittération spécifiques à la question de l'appariement sont traités dans le cas particulier anglais/arabe, tandis que dans (Pouliquen *et al.*, 2006) les auteurs présentent un système d'appariement multilingue.

Les mesures de similarité (ou de distance) textuelle sont les principaux outils utilisés pour ce type de tâche. On peut grossièrement classer les différents types de mesures existants en trois classes :

– *les méthodes basées sur les séquences de caractères*, qui définissent la similarité par la présence de caractères identiques à des positions similaires (e.g. Levenshtein, Jaro).
– *les méthodes de type "sac de mots"*, qui sont basées sur le nombre de mots en commun entre les deux chaînes, indépendamment de leur position. Notons que ces types de mesures sont également applicables aux n-grammes de caractères au lieu des mots.
– *les méthodes hybrides*, qui combinent les caractéristiques des deux précédents types de mesures.

Notre objectif est de combiner différentes mesures de similarité de façon à améliorer la qualité de l'appariemment d'entités. Dans cette optique, nous proposons de tirer profit aussi des similitudes éventuelles entre les contextes des entités, cette technique étant déjà utilisée pour le problème connexe de la désambiguisation d'homonymes, notamment dans (Pedersen & Kulkarni, 2007). La combinaison de plusieurs mesures est utilisée dans (Pouliquen *et al.*, 2006), où celle-ci consiste en une simple moyenne sur trois mesures. L'apprentissage supervisé permet une prise en compte plus fine des caractéristiques des mesures et des données, comme le montrent (Bilenko & Mooney, 2003) à l'aide de l'algorithme SVM dans le cadre des bases de données. À l'inverse, nous proposons ici d'appliquer différentes méthodes d'apprentissage dans le cas de données issus de textes non structurés : dans ce contexte, on ne dispose pas de l'information apportée par les différents champs d'un enregistrement, mais seulement du contexte dans lequel sont trouvées les EN.

Notre approche privilégie la robustesse de l'appariement, au sens où nous proposons de réaliser cette tâche sur tout type d'EN, indépendamment de ressources externes (e.g. dictionnaires de noms, heuristiques spécifiques à un domaine, etc.), et autant que possible sans distinction de langue. À ce titre nous travaillons sur des données potentiellement bruitées. En effet, dans le contexte d'applications réelles, cette tâche dépend de nombreuses phases en amont : la qualité du corpus d'origine, celle de l'extraction de sites web ainsi que celle de la phase de reconnaissance des EN. Notre but n'est donc pas d'obtenir les meilleurs résultats possibles sur des données spécifiques (objectif requérant un travail d'expertise précis et coûteux), mais plutôt des performances satisfaisantes facilement reproductibles sur différents types de données.

Nous présenterons dans un premier temps les données dont nous disposons et les principales mesures de similarité utilisées dans le système que nous avons implémenté. Nous exposerons ensuite notre approche face aux spécificités du problème, et enfin nous présenterons les expériences réalisées et les résultats obtenus.

# 2 Données et outils

## 2.1 Corpus

Le premier corpus (qui sera noté MNI par la suite), en anglais, est constitué d'un recueil d'articles de presse, de dépêches et de rapports officiels de provenance variée consacrés à la menace nucléaire en Iran. Celui-ci provient du site `www.nti.org`[1], dont nous utilisons la partie traitant de l'Iran (pour la période 1991-2006) parce qu'elle contient de nombreux cas de translittérations de noms arabes. Ce corpus compte 236 000 mots, et la reconnaissance des EN a été réalisée à l'aide de l'outil GATE[2]. Nous conservons seulement les noms de personnes, d'organisations et de lieux dans l'ensemble des EN ainsi constitué. Nous obtenons ainsi 35 000 EN (en nombre d'occurrences), contenant toutefois quelques erreurs de reconnaissance : principalement des cas de balisage erroné (entité tronquée ou mots superflus inclus dans l'entité) et de noms communs commençant par une majuscule. Nous travaillons sur les EN de fréquence supérieure ou égale à 2, ce qui restreint à 1588 le nombre d'entités distinctes (représentant 33 147 occurrences).

Le second corpus (noté MIF par la suite), long de 856 000 mots, est le résultat de l'aspiration du contenu de 20 sites web de médias d'information francophones. Ces médias ont été sélectionnés selon les critères suivants : volume suffisant, facilité d'accès et surtout diversité géographique, de façon à maximiser les chances d'y trouver des translittérations (c'est pourquoi nous avons notamment intégré une part non négligeable de médias d'Afrique du Nord). L'extraction a été réalisée durant 4 jours en juillet 2007 par Pertimm[3]. Le corpus ainsi obtenu a ensuite été traité par Arisem[4] pour la phase d'extraction des entités : 34 000 occurrences d'EN ont été reconnues comme noms de personnes, organisations ou lieux, parmi lesquelles on trouve encore quelques erreurs (principalement des noms communs). De même que pour le corpus MNI, on restreint l'ensemble des entités traitées à celles apparaissant au moins deux fois : on dénombre alors 3278 entités, correspondant à 23725 occurrences.

Rappelons que notre tâche est centrée sur l'appariement et non la reconnaissance des EN, ce qui signifie qu'on admet l'hypothèse que l'extraction des EN (en amont) est globalement "correcte". Nous n'avons pas cherché à corriger les erreurs de cette phase, puisque cela contredirait notre objectif de robustesse vis-à-vis des données disponibles.

## 2.2 Mesures de similarités

Nous présentons ci-dessous quelques-unes des principales mesures fréquemment utilisées pour l'appariement d'EN (Christen, 2006; Cohen *et al.*, 2003; Bilenko *et al.*, 2003).

*La distance d'édition de Levenshtein (et variantes).* Cette mesure de distance $d$ représente le nombre minimal d'insertions, suppressions ou substitutions nécessaires pour transformer une chaîne $x$ en une chaîne $y$. Exemple : $d(kitten, sitting) = 3$ ($k \mapsto s$, $e \mapsto i$, $\varepsilon \mapsto g$). La similarité $s(x, y)$ normalisée sur $[0, 1]$, est définie par $s = 1 - d/max(|x|, |y|)$.

[1] *Nuclear Threat Initiative*, qui recense par pays toutes les données disponibles liées au risque nucléaire.
[2] http ://gate.ac.uk
[3] `www.pertimm.com`
[4] `www.arisem.com`

*La métrique de Jaro*. Cette mesure est basée sur le nombre et l'ordre des caractères communs entre deux chaînes. Étant données deux chaînes $x = a_1 \dots a_n$ et $y = b_1 \dots b_m$, soit $H = \min(n, m)/2$ : un caractère $a_i$ de $x$ est *en commun* avec $y$ s'il existe $b_j$ dans $y$ tel que $a_i = b_j$ et $i - H \leq j \leq i + H$. Soit $x' = a'_1 \dots a'_{n'}$ (respectivement $y' = b'_1 \dots b'_{m'}$) la séquence de caractères de $x$ (resp. $y$) en commun avec $y$ (resp. $x$), dans l'ordre où les caractères apparaissent dans $x$ (resp. $y$). Toute position $i$ telle que $a'_i \neq b'_i$ est appelée une *transposition*. Soit $T$ le nombre de transpositions entre $x'$ et $y'$ divisé par 2, la mesure de similarité de Jaro est définie[5] par : $Jaro(x, y) = \frac{1}{3} \times \left( \frac{|x'|}{|x|} + \frac{|y'|}{|y|} + \frac{|y'| - T}{|y'|} \right)$.

*Les mesures de type "sac de mots" ou "de n-grammes de caractères"*. Pour ces mesures, chaque entité est traitée comme un ensemble d'éléments (les mots ou les n-grammes). Soient $X = \{x_i\}_{1 \leq i \leq n}$ et $Y = \{y_i\}_{1 \leq i \leq m}$ les ensembles représentant les EN $x, y$ à comparer. Les mesures les plus simples ne prennent en compte que le nombre d'éléments en commun[6], par exemple :

$$Jaccard(x, y) = \frac{|X \cap Y|}{|X \cup Y|}; \; Overlap(x, y) = \frac{|X \cap Y|}{min(|X|, |Y|)}; \qquad Cos(x, y) = \frac{|X \cap Y|}{\sqrt{|X|}\sqrt{|Y|}}$$

Certaines mesures plus élaborées s'appuient sur une représentation vectorielle des ensembles $X$ et $Y$, qui peut tenir compte de paramètres extérieurs aux ensembles eux-même. Soient $A = (a_1, \dots, a_{|\Sigma|})$ et $B = (b_1, \dots, b_{|\Sigma|})$ ces vecteurs[7], la similarité définie par le cosinus de l'angle formé par $A$ et $B$ est fréquemment utilisée : $cos(A, B) = \frac{A^T B}{\|A\| \times \|B\|}$. La représentation des éléments par leurs poids TF-IDF (*Term Frequency-Inverse Document Frequency*) est l'une des plus classiques. Il s'agit dans notre cas de mesurer l'importance d'un élément $w$ pour une EN $x$ parmi un ensemble $E$ d'entités[8] :

$$tf_{w,x} = \frac{n_{w,x}}{\sum_{w' \in \Sigma} n_{w',x}}, \quad idf_w = \log \frac{|E|}{|\{x \in E | w \in x\}|}, \quad tfidf_{w,x} = tf_{w,x} \times idf_w.$$

*Les combinaisons de mesures*. Leur principe est la combinaison des propriétés des différents types de mesures présentés ci-dessus. Il s'agit généralement d'appliquer une "sous-mesure" $sim'$ aux mots des deux EN à comparer, puis d'en déduire un éventuel alignement optimal des EN. Il s'agit donc d'appliquer une méthode de type "sac de mots", mais sans subir la rigidité d'un test d'identité entre mots : par exemple, les entités *"Director ElBaradei"* et *"Director-General ElBareidi"* présentent des similarités importantes que les mesures "sac de mots" classiques ne prennent pas en compte. La sous-mesure doit bien sûr être choisie judicieusement.

– La mesure de Monge-Elkan calcule simplement la moyenne des meilleurs paires de mots trouvés : $sim(x, y) = \frac{1}{n} \sum_{i=1}^{n} \max_{j=1}^{m}(sim'(x_i, y_j))$
– La mesure Soft-TFIDF proposée dans (Cohen *et al.*, 2003) est une forme assouplie du cosinus sur les vecteurs de poids TF-IDF : grossièrement, deux mots différents peuvent être considérés comme identiques selon que leur score de sous-mesure dépasse ou non un seuil.

Enfin, on peut mesurer la *similarité des contextes* des EN. On nomme *contexte* d'une occurrence d'une EN l'ensemble des $n$ mots qui la suivent et qui la précèdent, et le contexte (global) d'une entité distincte est formé par l'union des contextes de toutes ses occurrences. De façon

[5] Il est utile de noter que cette mesure n'est pas symétrique. On trouve dans la littérature et dans les implémentations existantes diverses variantes pour contourner ce problème.

[6] Avec $|E|$ le cardinal de l'ensemble $E$.

[7] $\Sigma$ est l'ensemble des éléments considérés (e.g. tous les mots apparaissant dans au moins une EN).

[8] Avec $n_{w,x}$ le nombre d'occurrences de $w$ dans l'EN $x$, et $\Sigma$ le vocabulaire.

classique (Pedersen & Kulkarni, 2007), nous calculons l'ensemble des vecteurs représentant le contexte de chaque entité, chaque vecteur contenant les poids TF-IDF des mots de ce contexte. La similarité entre les contextes de deux EN est alors le cosinus de leurs vecteurs respectifs.

# 3 Approche proposée

Nous avons implémenté un prototype d'évaluation de mesures de similarités entre EN. À partir d'une liste d'entités et de leur contexte, celui-ci calcule le score de similarité obtenu par chaque couple d'EN pour un ensemble prédéfini de mesures. 48 mesures sont disponibles, dont une vingtaine proviennent de deux librairies publiques : SimMetrics[9] de S. Chapman et SecondString[10] de W. Cohen, P. Ravikumar et S. Fienberg.

## 3.1 Difficultés posées par l'étiquetage

Il est bien entendu nécessaire de disposer de données étiquetées, d'une part pour pouvoir tester et comparer les performances des différentes mesures, et d'autre part pour pratiquer l'apprentissage supervisé. Cependant, la tâche d'appariement présente certaines spécificités qui rendent la phase d'étiquetage de données difficile. En effet, nous cherchons à classer des couples d'EN comme positifs (coréférence) ou négatif (non coréférence). Or pour $n$ entités distinctes l'ensemble des couples potentiel comprend $n \times (n-1)/2$ éléments, il serait donc excessivement coûteux en temps d'envisager l'étiquetage manuel de cet ensemble (pour les valeurs de $n$ étudiées). Dans de telles circonstances, une technique usuelle consiste à n'étiqueter qu'un sous-ensemble de couples tirés aléatoirement. Mais cette alternative n'est pas envisageable ici, à cause de la disproportion entre couples positifs et négatifs : dans nos données, on ne trouve respectivement que 0,06% (pour MNI) et 0,02% (pour MIF) de couples positifs.

C'est pourquoi notre approche vise à extraire de façon semi-automatique un ensemble contenant tous les couples positifs. Seul cet ensemble sera examiné au cours de l'étiquetage manuel. Cette approche repose sur l'hypothèse selon laquelle les couples positifs seront jugés similaires par au moins une des métriques ; à l'inverse, les couples qui ne sont bien classés par aucune métrique sont considérés négatifs. Cette méthodologie n'est pas sans biais, mais une analyse approfondie d'un ensemble d'entités nous a permis de constater que ce biais était en réalité faible, du fait de la multiplicité et de la diversité des mesures utilisées.

Les critères d'étiquetage manuel ainsi que les méthodes automatiques de recherche de couples candidats ont été affinés pour le traitement du corpus MIF, grâce à l'expérience acquise avec le corpus MNI. C'est pourquoi nous ne détaillons ci-dessous que la méthode employée sur MIF, sachant que le changement principal réside dans une définition beaucoup plus stricte de la coréférence.

Pour la recherche de couples candidats, notre système propose d'abord les couples obtenant les $k$ meilleurs scores selon chaque mesure. Deux autres techniques sont également mises en œuvre : la première consiste à appliquer automatiquement les relations de transitivité (si les EN $A$ et $B$ sont coréférentes et que $B$ et $C$ le sont aussi, alors $A$ et $C$ sont coréférentes). La seconde vise à repérer d'éventuels couples difficiles à trouver de façon globable (par exemple,

[9] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html
[10] http://secondstring.sourceforge.net

les EN courtes sont défavorisées par la majorité des mesures) : dans ce but, l'ensemble des EN est parcouru, en proposant pour chaque EN les $n$ entités les plus proches selon $m$ "bonnes" mesures. Les couples sont classés en trois catégories : les *positifs* (coréférence stricte, au moins dans le corpus) ; les *négatifs* (non-coréférence stricte) ; les *couples incertains* (ne pouvant être acceptés comme positifs mais présentant toutefois un lien étroit[11]). Enfin certaines EN sont éliminées (principalement les erreurs de reconnaissance ou les EN mal formées, mais aussi quelques cas ambigus).

Par rapport au corpus MNI, plus de temps a été consacré à la recherche de couples coréférents parmi les entités. En particulier, bon nombre d'acronymes ont été appariés manuellement avec leur forme étendue, ainsi que quelques cas tels que *"Quai d'Orsay"* et *"Ministère des affaires étrangères*[12]. Le parcours d'appariement local, au cours duquel chaque EN est prise comme référence, a permis de repérer une douzaine de couples positifs supplémentaires parmi environ 30 000. Pour toutes ces raisons, nous pensons que la probabilité pour un couple positif de n'avoir pas été étiqueté est très basse.

| Corpus | EN | EN éliminées | positifs | négatifs | incertains | total |
|---|---|---|---|---|---|---|
| Corpus MNI | 1588 | 0 | 805 | 1 877 | 3 836 | 1 260 078 |
| Corpus MIF | 3278 | 745 | 741 | 32 348 | 419 | 3 206 778 |

Dans les colonnes 2 et 3 de ce tableau sont représentés des nombres d'*entités*, tandis que les suivantes contiennent des nombres de *couples d'entités*. Ainsi dans MIF 3278-745 EN représentent 3 206 778 couples, parmi lesquels 33 508 ont été étiquetés. Pour MNI, aucune EN reconnue n'avait été éliminée (car la méthode employée pour l'étiquetage était différente), c'est pourquoi la proportion d'*incertains* est si importante (elle reste toutefois négligeable par rapport à l'ensemble des couples).

## 3.2 Apprentissage supervisé : motivations et outils

Nous développerons dans la partie 4.1 les résultats obtenus par les mesures de similarité testées. On peut néanmoins déjà déduire de leurs définitions (cf. partie 2.2) qu'elles ont chacune des propriétés spécifiques qui les rendent potentiellement complémentaires. Nous pouvons observer ces différences sur quelques cas positifs non triviaux issus du corpus MIF dans le tableau 3.2.

| Couple | Levenshtein | TF-IDF mots | TF-IDF trigrammes | TF-IDF contextes |
|---|---|---|---|---|
| "Fatah al-Islam" / "Fateh el-Islam" | 281 | > 3000 | 686 | > 3000 |
| "Mosquée Rouge" / "Mosquée rouge d'Islamabad" | > 3000 | 233 | 449 | > 3000 |
| "Omar al-Baghdadi" / "Omar de Bagdad" | 887 | 1802 | 2318 | 10 |
| "Recep Tayyip Erdogan" / "Erdogan" | > 3000 | 746 | 2406 | 2510 |

Les valeurs indiquées représentent la position du couple dans la liste triée par score décroissant de chaque mesure[13] : le couple "Fatah al-Islam" / "Fateh el-Islam" est ainsi classé 281e par la mesure de Levenshtein.

Ces exemples illustrent le fait qu'aucune de ces mesures n'est capable de prendre en compte tous les types d'indices permettant de statuer sur l'éventuelle coréférence d'un couple d'entités. C'est pourquoi nous proposons d'utiliser l'apprentissage supervisé : nous espérons ainsi déterminer une manière optimale de combiner les scores obtenus à l'aide de différentes mesures, dans le but d'améliorer les performances de la tâche d'appariement d'EN.

[11] Exemples : *"ONU"* et *"Conseil de sécurité de l'ONU"*, ou *"Russie"* et *"Gouvernement russe"*.

[12] Notons que ce type de couple est hors de portée des mesures de similarité textuelle.

[13] Rappelons que ce corpus contient 741 positifs, ce qui signifie qu'un seuil (choisi de façon à garantir une précision minimale) sur une seule mesure ne permettrait de conserver au mieux que les $500 \pm 100$ meilleurs scores.

Dans les données fournies à l'algorithme d'apprentissage, chaque couple d'entités est représenté par un ensemble de paramètres choisis pour leur contribution potentielle à la détection d'une coréférence. Parmi ces paramètres figurent bien entendu les scores de similarité obtenus avec différentes mesures, mais aussi certaines caractéristiques du couple d'EN telles que leurs longueurs (en nombre de caractères et de mots) et leurs fréquences minimales et maximales. Nous utilisons le logiciel Weka (Witten & Frank, 2005) pour réaliser l'apprentissage[14], et testons deux méthodes de classification : La *régression logistique,* qui apprend un séparateur linéaire, et L'*algorithme C4.5* (Quinlan, 1993), qui apprend un arbre de décision.

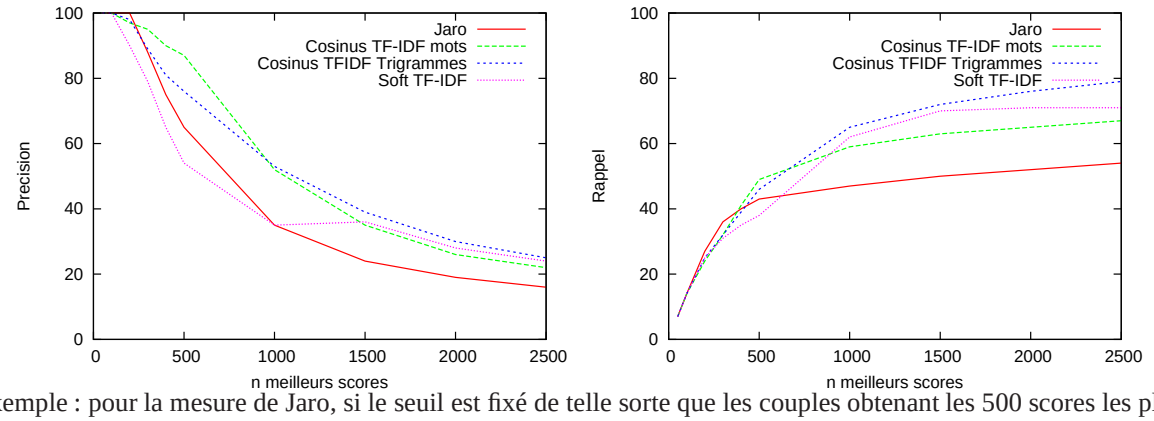
# 4 Expérimentations et résultats

Conformément à l'approche que nous avons suivie pour l'étiquetage des données, les résultats[15] détaillés ci-dessous sont évalués sous les hypothèses suivantes : tout couple non étiqueté est assimilé à un couple négatif ; tout couple marqué comme "incertain" est simplement ignoré.

## 4.1 Observations générales

Tout d'abord, nous constatons que les mesures se comportent de façon similaire sur les deux corpus. Des différences de performances importantes sont observées, mais celles-ci sont principalement dues aux critères d'étiquetage différents (voir partie 3.1).

FIG. 1 – Précision et rappel pour 4 mesures de similarité (corpus MIF)




Exemple : pour la mesure de Jaro, si le seuil est fixé de telle sorte que les couples obtenant les 500 scores les plus élevés soient classés positifs, la précision est de 65% et le rappel de 46% (rappel : MIF contient 741 positifs).

La typologie des ressemblances reconnues par type de mesure laisse apparaître quelques grandes lignes : sur les mots simples qui présentent de légères différences textuelles, souvent des noms de lieux ou de personnes, les mesures de type Levenshtein/Jaro sont performantes. Mais celles-ci deviennent inadaptées dès que plusieurs mots sont présents, ce qui est essentiellement le cas des noms d'organisation mais aussi souvent des noms de personnes (e.g. avec/sans prénom/titre) : les mesures "sac de mots" sont alors nettement meilleures (voir table 3.2).

[14]Cet outil public propose un ensemble varié d'algorithmes d'apprentissage prêts à l'emploi pour la fouille de donnée.

[15]L'apprentissage est réalisé par validation croisée en 10 sous-ensembles. Dans tous les cas étudiés, le taux d'erreur global est très faible (inférieur à 0,1%) puisque le taux de couples négatifs est très élevé.

En général, les mesures qui obtiennent les meilleures performances sont de type “sac de mots” ou “sac de n-grammes”, tandis que les mesures basées sur les séquences de caractères, moins souples, ne permettent d’identifier sans erreur que les couples positifs très proches. Sans surprise, la prise en compte de l’IDF améliore assez nettement les résultats des mesures de type sac de mots/n-grammes. Individuellement, la mesure de similarité des contextes n’a pas d’intérêt[16].

## 4.2 Apport de l’apprentissage

### 4.2.1 Mesures individuelles

Dans le tableau 1 sont indiquées les performances obtenues par quelques-une des meilleures mesures individuelles. Celles-ci sont calculées selon les deux méthodes d’apprentissage, de façon à pouvoir servir de référence par rapport aux combinaisons de mesures décrites ci-après. Dans ce même tableau nous évaluons l’apport des paramètres de longueur/fréquence des EN. On peut constater que les résultats sont très proches avec les deux méthodes dans le cas des mesures seules, tandis que le C4.5 tire beaucoup mieux profit des paramètres de longueur/fréquence : la F-mesure va jusqu’à augmenter de 26 (MNI) ou 15 (MIF) points pour la mesure de Jaro.

TAB. 1 – Mesures individuelles avec/sans longueurs et fréquences (pourcentages)

| Paramètres | Corpus MNI | | | | | | Corpus MIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Régr. log. | | | C4.5 | | | Régr. log. | | | C4.5 | | |
| | P | R | F | P | R | F | P | R | F | P | R | F |
| Jaro seule | 66,0 | 25,0 | 36,2 | 74,3 | 17,6 | 28,5 | 89,4 | 40,2 | 55,4 | 91,6 | 38,7 | 54,4 |
| Jaro + l/f | 67,4 | 34,2 | 45,3 | 81,8 | 42,5 | 55,9 | 84,1 | 43,0 | 56,9 | 88,2 | 57,0 | 69,2 |
| TF-IDF mots seule | 85,6 | 58,5 | 69,5 | 83,3 | 60,0 | 69,7 | 81,9 | 51,0 | 62,9 | 91,5 | 46,7 | 61,8 |
| TF-IDF mots + l/f | 86,3 | 58,8 | 69,9 | 88,4 | 63,6 | 74,0 | 82,0 | 50,6 | 62,6 | 86,9 | 48,4 | 62,2 |
| TF-IDF trigrammes seule | 79,5 | 64,7 | 71,4 | 76,0 | 67,2 | 71,3 | 73,1 | 49,4 | 58,9 | 70,7 | 52,1 | 60,0 |
| TF-IDF trigrammes + l/f | 84,7 | 71,3 | 77,4 | **87,1** | **68,9** | **77,0** | 77,4 | 55,0 | 64,3 | **84,5** | **62,1** | **71,6** |

P/R/F = Précision/Rappel/F-mesure. “+ l/f” signifie “avec les paramètres de longueurs et fréquences”.

Exemple : sur le corpus MNI, l’apprentissage par C4.5 sur les paramètres constitués du score de TF-IDF sur les mots et des longueurs et fréquences min. et max. de chaque couple donne un rappel de 63,6%.

### 4.2.2 Combinaisons de mesures

Nous avons testé plusieurs sélections de mesures comme paramètres de l’apprentissage. Les résultats de ces expérimentations pour deux sélections de mesures et quelques variantes sont fournis dans le tableau 2. On constate globalement une nette amélioration des performances par rapport au cas des mesures individuelles : en comparant les meilleurs cas des deux situations, le rappel passe ainsi de 69% à 83% sur le corpus MNI et de 62% à 75% sur le corpus MIF. C’est encore une fois l’algorithme C4.5 qui combine les différents paramètres de façon optimale.

En revanche, la contribution de la mesure de similarité des contextes, particulièrement étudiée ici, est quasiment nulle. Cependant, en considérant un ensemble restreint de mesures de façon à analyser plus en détail cette mesure (tableau 2), on constate un apport faible mais significatif de celle-ci : l’algorithme C4.5 en permet un usage positif, puisque la F-mesure gagne 2,7 points

[16] Elle n’atteint jamais les 20% de précision. Pourtant, on observe que ce score est bien représentatif d’une proximité sémantique, mais celle-ci s’avère trop peu précise pour marquer une éventuelle coréférence (par exemple, on trouve souvent parmi les bons scores des couples formés d’une organisation et du nom de son représentant).

TAB. 2 – Performances de différentes combinaisons de mesures (pourcentages)

| Paramètres | Corpus MNI | | | | | | Corpus MIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Régr. log. | | | C4.5 | | | Régr. log. | | | C4.5 | | |
| | P | R | F | P | R | F | P | R | F | P | R | F |
| A seules | 85,9 | 75,0 | 80,1 | **85,1** | **83,2** | **84,2** | 78,7 | 58,8 | 67,3 | 89,7 | 67,6 | 77,1 |
| A + l/f | 87,2 | 76,4 | 81,5 | 85,2 | 80,9 | 83,0 | 79,0 | 59,2 | 67,6 | 87,3 | 71,0 | 78,3 |
| A' seules | 86,2 | 76,6 | 81,1 | 86,3 | 82,0 | 84,1 | 82,9 | 63,9 | 72,2 | 87,1 | 73,9 | 80,0 |
| A' + l/f | 87,5 | 77,4 | 82,1 | 84,5 | 80,5 | 82,4 | 82,2 | 63,8 | 71,8 | 85,1 | 75,2 | 79,8 |
| B seules | 82,9 | 69,1 | 75,3 | 83,1 | 76,3 | 79,5 | 74,4 | 51,6 | 60,9 | 87,2 | 71,0 | 78,2 |
| B + l/f | 84,9 | 73,3 | 78,7 | 82,5 | 78,3 | 80,3 | 80,5 | 59,4 | 68,4 | 87,0 | 70,7 | 78,0 |
| B' seules | 84,4 | 71,3 | 77,3 | 82,1 | 77,5 | 79,7 | 80,7 | 59,6 | 68,5 | **87,5** | **75,3** | **81,0** |
| B' + l/f | 86,6 | 74,2 | 79,9 | 83,8 | 79,3 | 81,5 | 81,9 | 64,0 | 71,9 | 85,9 | 75,4 | 80,3 |

L'ensemble de mesures A est constitué des mesures Cosinus (nombre de mots), Jaro, Smith-Waterman-Gotoh, TFIDF mots, TFIDF trigrammes et Soft-TFIDF. L'ensemble B est constitué des mesures Levenshtein, Jaro, TFIDF mots, TFIDF bigrammes, TFIDF trigrammes et une combinaison par couplage de mots basée sur Jaro. A' (resp. B') est l'ensemble A (resp. B) auquel est ajouté le TFIDF sur les contextes.

dans le cas où ce paramètre est ajouté à deux autres bonnes mesures. Un gain similaire est observé entre l'une des mesures prise individuellement (tableau 1) et la même avec le contexte.

FIG. 2 – Influence de la mesure sur les contextes (corpus MIF) (pourcentages)

| Paramètres | Régr. log. | | | C4.5 | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| TFIDF Trigrammes + Contexte + l/f | 80,4 | 61,6 | 69,7 | 81,9 | 70,8 | 76,0 |
| TFIDF Mots + Contexte + l/f | 85,7 | 52,0 | 64,7 | 83,2 | 55,2 | 66,4 |
| TFIDF Trigrammes + TFIDF Mots + l/f | 77,4 | 57,5 | 66,0 | 86,0 | 64,2 | **73,5** |
| TFIDF Trigrammes + TFIDF Mots + Contexte + l/f | 80,5 | 62,0 | 70,0 | 83,1 | 70,3 | **76,2** |

Le choix des mesures est une question complexe : tout d'abord, on constate assez naturellement que plus on fournit de paramètres à l'algorithme d'apprentissage, meilleur est le modèle qu'il produit. Ainsi, nous avons testé le C4.5 sur le corpus MNI avec 18 mesures différentes (et les paramètres de longueur/fréquence) : ceci permet d'obtenir jusqu'à 84,9%/85,0% de précision/rappel. Toutefois cet apport est limité : le gain obtenu en combinant les scores de seulement deux mesures (pertinentes) est important, mais il a tendance à s'affaiblir avec l'augmentation du nombre de mesures. Ceci est dû en partie au fait qu'une combinaison de mesures n'a d'intérêt que si celles-ci sont complémentaires, or on retrouve rapidement des mesures de même type. De plus, dans un cadre d'utilisation réelle, les ressources matérielles ne permettent pas de recourir au calcul de dizaines de mesures pour des volumes de données importants. C'est pourquoi il semble judicieux d'adopter un compromis raisonnable en sélectionnant de 3 à 5 mesures complémentaires. À ce titre, soulignons que malgré ses performances assez faibles la mesure de similarité des contextes est un bon candidat sur le plan de la complémentarité.

# 5 Conclusion et perspectives

Pour conclure, dans cet article nous avons montré l'intérêt de combiner les scores de différentes mesures de similarité pour l'appariement d'EN extraites de textes non structurés. Les expériences menées sur deux corpus montrent que l'apprentissage supervisé permet de telles combinaisons, et que celles-ci améliorent de façon significative les performances de la tâche d'appariement. Pour cet apprentissage, la comparaison de la régression logistique et de l'algo-

rithme C4.5 est nettement en faveur de ce dernier. Dans ce cadre, nous avons également étudié l'apport d'une mesure de similarité des contextes, qui semble faible mais non négligeable.

L'inconvénient le plus important de cette méthode est certainement la nécessité de données étiquetées, très difficiles et/ou coûteuses à obtenir à cause des spécificités de cette problématique (nombre de couples potentiels très élevé et disproportion entre positifs et négatifs). C'est pourquoi il nous semble pertinent d'étudier les possibilités d'apprentissage non-supervisé ou semi-supervisé (par exemple en sélectionnant judicieusement les couples à étiqueter).

## Remerciements

Ces travaux ont été financés par le projet Cap Digital - Infom@gic. Nous remercions L. Rigouste (Pertimm), N. Dessaigne et A. Migeotte (Arisem) pour nous avoir fourni le corpus MIF annoté.

## Références

BILENKO M. & MOONEY R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. In P. DOMINGOS, C. FALOUTSOS, T. SENATOR, H. KARGUPTA & L. GETOOR, Eds., *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03)*, p. 39–48, New York : ACM Press.

BILENKO M., MOONEY R. J., COHEN W. W., RAVIKUMAR P. & FIENBERG S. E. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, **18**(5), 16–23.

CHRISTEN P. (2006). *A Comparison of Personal Name Matching : Techniques and Practical Issues*. Rapport interne TR-CS-06-02, Department of Computer Science, The Australian National University, Canberra 0200 ACT, Australia.

COHEN W. W., RAVIKUMAR P. & FIENBERG S. E. (2003). A comparison of string distance metrics for name-matching tasks. In S. KAMBHAMPATI & C. A. KNOBLOCK, Eds., *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico*, p. 73–78.

FREEMAN A., CONDON S. L. & ACKERMAN C. (2006). Cross linguistic name matching in English and Arabic. In R. C. MOORE, J. A. BILMES, J. CHU-CARROLL & M. SANDERSON, Eds., *HLT-NAACL* : The Association for Computational Linguistics.

PEDERSEN T. & KULKARNI A. (2007). Unsupervised discrimination of person names in Web contexts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

POULIQUEN B., STEINBERGER R., IGNAT C., TEMNIKOVA I., WIDIGER A., ZAGHOUANI W. & ZIZKA J. (2006). Multilingual person name recognition and transliteration. *CORELA - Cognition, Representation, Langage.*

QUINLAN J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

WINKLER W. E. (1999). *The state of record linkage and current research problems.* Rapport interne RR99/04, US Bureau of the Census.

WITTEN I. H. & FRANK E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.