

Terminologie & Ontologie : Théories et applications



Actes de la conférence

TOTh 2010

Annecy – 3 & 4 juin 2010

avec le soutien de :

- Ministère de la Culture et de la Communication, Délégation Générale à la Langue Française et aux Langues de France
- Association Européenne de Terminologie
- Société française de terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France



Institut Porphyre
Savoir et Connaissance

<http://www.porphyre.org>

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Franco Bertaccini	Professeur, Université de Bologne
Gerhard Budin	Professeur, Université de Vienne
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille 3
Rute Costa	Professeur, Universidade Nova de Lisboa
Luc Damas	MCF, Université de Savoie
Sylvie Desprès	Professeur, Université Paris 13
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section terminologie
Jean-Yves Gresser	Ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Equipe Condillac
Widad Mustafa	Professeur, Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge (The Irish-Language Body)
Henrik Nilsson	Terminologocentrum TNC, Suède
Jean Quirion	Professeur, Université d'Ottawa
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS, Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas
Samia Chouder, Joëlle Pellet

Avant propos



Cette année la conférence a été précédée d'une journée de formation consacrée à la terminologie et l'ontologie, à leurs liens et leurs apports mutuels. L'intérêt qu'a suscité cette journée nous amènera certainement à réitérer l'opération les années suivantes.

Le succès de la conférence d'ouverture de notre collègue Frédéric Nef, portant sur l'ontologie prise dans sa dimension philosophique, a montré, s'il en était encore besoin, la richesse d'une approche pluridisciplinaire.

Animées par différents présidents, les sessions ont alterné présentations théoriques et démonstrations de systèmes, offrant ainsi l'opportunité à plusieurs industriels de nous parler de leurs projets. L'éventail des sujets abordés, à travers les quatorze présentations retenues (incluant la conférence d'ouverture) réparties sur deux jours, illustre la richesse mais aussi la vitalité de notre communauté : aide à la traduction, thésaurus multilingue, phraséologie, entité nommée, recherche d'information, etc. L'« actualité » n'a pas été oubliée à travers une ontologie des risques financiers.

Enfin, les Conférences TOTh sont devenues internationales à partir de cette année avec le français et l'anglais comme langues officielles. Le comité de programme s'est ouvert à de nouveaux membres portant à dix le nombre de pays représentés et à plus de 40% le nombre de personnalités étrangères. Gageons que cette ouverture sera prometteuse.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

<i>L'Ontologie au miroir de la Terminologie</i>	9
Frédéric Nef	

ARTICLES

<i>Le travail sur la représentation (visuelle) des connaissances en terminologie : un retour d'expérience</i>	31
Dardo de Vecchi	
<i>Une « ontoterminologie » pour les interprètes de conférence</i>	53
Elisa Veronesi, Franco Bertaccini	
<i>Semiotic Triangle Revisited for the Purposes of Ontology-based Terminology Management</i>	83
Igor Kudashev, Irina Kudasheva	
<i>L'ontoterminologie pour la recherche d'information sémantique</i>	101
Luc Damas, Christophe Tricot	
<i>Modélisation des dénominations ontologiques</i>	117
Benjamin Diemert, Marie-Hélène Abel, Claude Moulin	
<i>Filtrage des Entités Nommées par des méthodes de Fouille de Textes</i>	141
Mathieu Roche	
<i>Ontologies des risques financiers – Continuité d'activité, gestion de crise, protection des infrastructures critiques financières</i>	155
Jean-Yves Gresser	
<i>Vers une ontologie pour le domaine de l'analyse de sécurité des systèmes de transport automatisés</i>	177
Lassaâd Mejri, Habib Hadj-mabrouk, Patrice Caulier	

DEMONSTRATIONS

<i>Une « ontoterminologie » pour les interprètes de conférence – Un outil développé au sein de l’environnement académique</i>	203
Elisa Veronesi, Franco Bertaccini	
<i>ITM, une infrastructure sémantique pour la maintenance du thésaurus multilingue Eurovoc</i>	207
Thomas Francart, Charles Teissède	
<i>Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext</i>	211
Falaise Achille, Tutin Agnès	
<i>Ontoterminologie : méthode et mises en œuvre</i>	217
Marie Calberg-Challot, Christophe Tricot	
<i>Libellex, plateforme de travail multilingue et référentiel terminologique d’entreprise</i>	225
François Brown de Colstoun, Estelle Delpech	
<i>Pages blanches</i>	230

Filtrage des Entités Nommées par des méthodes de Fouille de Textes

Mathieu Roche

Résumé : Cet article présente une approche de Traitement Automatique du Langage (TAL) afin de filtrer les Entités Nommées à partir d'une liste de candidats à la collocation. La méthode proposée s'appuie uniquement sur des mesures statistiques associées aux ressources du Web. L'évaluation à partir de candidats à la collocation de type Nom-Nom issus d'un corpus en français (corpus de CVs) permet de valider l'approche et de discuter des limites de cette dernière.

Mots-clés : Traitement Automatique du Langage (TAL), Fouille de Textes, Collocations, Entités Nommées

1. Introduction

Dans cet article, nous nous intéressons à l'étude des groupes de mots souvent appelés des collocations. Ces groupes peuvent être extraits par des méthodes de TAL (Traitement Automatique du Langage). Plus formellement, [Clas, 1994] donne deux propriétés définissant une collocation. Premièrement, une collocation est définie comme un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe. Par exemple, « lumière vive » est considéré comme une collocation car le sens global de ce groupe de mots peut être déduit des deux mots « lumière » et « vive ». En nous appuyant sur cette définition, l'expression « tirer son chapeau » n'est pas une collocation car son sens ne peut pas être déduit de chacun des mots. De telles formes sont appelées des **combinaisons figées**. Une deuxième propriété est ajoutée par [Clas, 1994] pour définir une collocation. Le sens des mots qui composent la collocation doit être limité. Par exemple « acheter un chapeau » n'est pas une collocation car le sens de « acheter » et de « chapeau » n'est pas limité. En effet, de multiples objets, voire des personnes, peuvent être achetés. De tels groupes de mots sont appelés des **combinaisons libres**. Notons cependant qu'il reste très difficile de différencier par des méthodes automatiques issues du TAL les locutions figées, libres et les collocations.

La définition générale des collocations étant donnée, elle peut être enrichie avec deux caractéristiques supplémentaires : les aspects sémantiques et syntaxiques [Heid, 1998; Laurens, 1999]. Le premier point s'appuie sur des caractéristiques sémantiques communes de certaines collocations. Par exemple, « lait tourné » et « beurre rance » ont des sens très proches liés à un phénomène de dégradation. Les aspects sémantiques définissant formellement les collocations sont pris en considération dans de nombreux travaux [Melcuk *et al.*, 1984-1999; Heid, 1998; Laurens, 1999]. Ainsi, [Melcuk *et al.*, 1984-1999] ont introduit les fonctions lexicales qui s'appuient sur des caractéristiques sémantiques pour définir les relations entre les unités des collocations. La deuxième caractéristique est liée à la structure syntaxique des collocations. A titre d'exemple, « lumière vive » et « marque distinctive » ont une même structure syntaxique de type Nom-Adjectif. Une classification de la structure syntaxique des collocations que nous donnons ci-dessous est proposée dans de nombreux travaux [Clas 94; Laurens 1999] : Nom-Verbe (par exemple, « interpréter un film »), Nom-Adjectif (par exemple, « cinéma muet »), Nom-Nom/Nom-Préposition-Nom (par exemple, « plateau de cinéma »), Verbe-Adverbe (par exemple, « boire goulûment »), Adverbe-Adjectif (par exemple, « gravement malade »).

Même si les méthodes de TAL ne permettent pas toujours d'identifier les collocations proprement dites qui s'appuient sur les définitions linguistiques énoncées, dans cet article, nous allons nous intéresser à l'extraction des *candidats à la collocation* qui respectent les patrons syntaxiques suivants : Nom-Nom, Nom-Préposition-Nom, Adjectif-Nom et Nom-Adjectif. La terminologie nominale de ce type est par exemple étudiée dans [Daille, 1996; Bourigault et Jacquemin, 1999]. Cependant, l'originalité de l'approche décrite dans cet article réside dans l'**identification automatique des *Entités Nommées*** à partir des candidats extraits.

Les Entités Nommées (EN)

Les EN sont classiquement définies comme les noms de Personnes, Lieux et Organisations. Initialement, une telle définition est issue des campagnes d'évaluation américaines MUC – *Message Understanding Conferences* qui furent organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les EN dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Aujourd'hui, de telles campagnes d'évaluation couvrent des tâches très variées sur la base de textes de différents domaines (textes spécialisés en biologie, dépêches d'actualités, blogs, etc). Nous pouvons, entre autres, citer les challenges TREC – *Text REtrieval*

Conférence (international) et DEFT – *DEfi Fouille de Textes* (francophone) qui sont aujourd'hui très actifs dans la communauté « fouille de textes ».

Comme le précisent [Daille *et al.*, 2000], les classes de base d'EN définies dans le cadre de MUC doivent être enrichies. Par exemple, outre les classes relatives aux Personnes, Lieux et Organisations, [Paik *et al.*, 1994] définissent de nouvelles classes telles que *Document* (logiciels, matériels, machines) et *Scientifique* (maladie, médicaments, etc).

Pour identifier les EN, de nombreux systèmes s'appuient sur la présence de majuscules [Daille *et al.*, 2000]. Cependant ceci peut se révéler peu efficace dans le cas d'EN non capitalisées et pour le traitement de textes non normalisés (mails, blogs, textes ou fragments de textes inégalement en majuscule ou minuscule, etc). A titre d'exemple, certaines données du défi DEFT'06 étaient constituées de discours politiques entièrement capitalisés [Azé *et al.*, 2006] (corpus disponible à l'adresse suivante : <http://deft.limsi.fr/>). Ainsi, nous avons choisi dans nos travaux de ne pas exploiter ce type d'informations pour identifier les EN. Notons cependant que de telles caractéristiques pourraient être intéressantes à associer à l'approche essentiellement statistique présentée dans cet article.

Plus formellement, pour caractériser les EN, les critères d'unicité référentielle (c'est-à-dire, un nom propre renvoie à une entité référentielle unique) et une stabilité dénominate (c'est-à-dire, peu de variations possibles) sont notamment précisées par [Fort *et al.*, 2009]. Nous allons nous appuyer sur ce dernier critère pour identifier les EN parmi les candidats à la collocation dont l'extraction est décrite dans la section suivante.

2. Méthode de fouille de textes pour l'extraction de candidats à la collocation

Nous présentons ci-dessous un processus automatique pour extraire les candidats à la collocation. Dans un premier temps, il est nécessaire de rassembler les textes à traiter. Ces textes devront être homogènes dans la spécialité étudiée (textes sur la biologie, les régimes politiques, etc). Nous appelons l'ensemble de ces textes des *corpus*. Après l'acquisition d'un corpus, l'étape suivante du traitement consiste à normaliser les textes. Cette tâche consiste par exemple à supprimer les caractères qui peuvent provoquer des erreurs dans les traitements automatiques (tirés d'énumérations, balises HTML, etc). Avec les textes normalisés nous pouvons appliquer un étiqueteur grammatical qui appose une étiquette grammaticale à chacun des mots du

corpus. Par exemple, dans la phrase *L'ouvrier règle la machine outil*. Les mots *ouvrier*, *machine* et *outil* auront une étiquette Nom, le mot *règle* aura une étiquette Verbe. Ces étiquettes sont apposées en utilisant des systèmes d'étiquetage morpho-syntaxique automatique [Brill, 1994]. L'étude effectuée dans cet article concerne l'extraction des candidats à la collocation à partir des textes étiquetés. Par exemple, avec le fragment étiqueté *L'/Article ouvrier/Nom règle/Verbe la/Article machine/Nom outil/Nom*, nous pouvons extraire le candidat à la collocation « *machine outil* » qui est de type Nom-Nom. Les candidats ainsi extraits sont utilisés pour des tâches précises : extraction d'informations, traduction automatique, classification de documents, etc. Notons que le filtrage grammatical ainsi appliqué permet de désambiguïser le mot « *règle* » qui peut avoir plusieurs rôles grammaticaux (nom, verbe).

La section suivante décrit une **méthode statistique** de sélection des EN à partir des candidats à la collocation obtenus après l'application du processus de Fouille de Textes.

3. Filtrage des Entités Nommées

Principe général

Il est fréquent que les candidats à la collocation de type Nom-Nom aient des formes variées comme nous le montrerons dans la section « Expérimentations » de cet article. Par exemple, la collocation « fichier clients » peut se décliner sous les formes Nom-Préposition-Nom : « fichier de clients », « fichiers pour clients », etc.

A contrario, les EN sont peu sujettes aux variations [Fort *et al.*, 2009] telles que les « variations prépositionnelles ». Nous allons nous appuyer sur cette constatation pour identifier avec des méthodes de TAL les EN nominales à partir d'une liste de candidats à la collocation de type Nom-Nom.

Pour cela, pour chaque candidat Nom-Nom, l'approche que nous décrivons dans cet article va consister à :

- (1) Construire artificiellement une collocation prépositionnelle de type Nom-Préposition-Nom à partir du candidat Nom-Nom.
- (2) Mesurer la « pertinence » de la collocation prépositionnelle construite en mesurant la dépendance entre chaque mot par des méthodes statistiques.

- (3) Sélectionner les collocations prépositionnelles ayant des scores faibles (c.-à-d. collocation construites peu pertinentes). En effet, si les possibilités de variations du candidat Nom-Nom sont faibles, nous pouvons supposer que ce candidat à la collocation peut potentiellement être une EN.

Nous allons maintenant décrire de manière précise chacune de ces étapes en nous appuyant sur les exemples « fichier clients » et « logiciel ciel ». Rappelons que le but de nos travaux est de déterminer automatiquement que le second candidat est en fait une EN.

Description du processus

Etape 1 – Construction

Nous allons dans une première étape construire des candidats prépositionnels en nous appuyant, dans ces travaux, sur la préposition « de » qui demeure la plus courante. En appliquant ce principe avec nos deux exemples, nous obtenons les résultats suivants :

fichier clients $_{NN}$ \rightarrow fichier de clients $_{N-Prep-N}$

logiciel ciel $_{NN}$ \rightarrow logiciel de ciel $_{N-Prep-N}$

Notons que lorsque le second Nom du terme de base commence par une voyelle, la préposition qui sera appliquée sera « d' » :

mission intérim $_{NN}$ \rightarrow mission d'intérim $_{N-Prep-N}$

Etape 2 – Mesure

Le but de la deuxième étape est de mesurer la dépendance entre chaque mot composant les collocations prépositionnelles construites. Pour cela nous allons nous appuyer sur une des mesures couramment utilisée en Fouille de Textes qui est le coefficient de Dice [Smadja *et al.*, 1996]. Le choix de cette mesure est motivé par son bon comportement que nous avons montré dans nos précédents travaux [Roche et Kodratoff, 2009]. Une telle mesure est définie par la formule suivante :

$$Dice(X, Y) = \frac{2 \times nb(X, Y)}{nb(X) + nb(Y)}$$

[Petrovic *et al.*, 2006] présentent une extension de la formule d'origine de Dice à trois éléments :

$$Dice(X, Y, Z) = \frac{3 \times nb(X, Y, Z)}{nb(X) + nb(Y) + nb(Z)}$$

Le coeur de cette mesure consiste à calculer le nombre d'occurrences de chaque mot « a » ($nb(a)$) ou collocation « a b c » ($nb(a, b, c)$). En règle générale, le nombre d'occurrences, c'est-à-dire la fréquence d'apparition des mots/collocations est calculée relativement à un corpus [Daille, 1996]. Dans notre cas, la mesure de Dice va être appliquée dans un contexte de fouille du web (web mining). Ainsi, la fréquence d'apparition nb correspondra au nombre de pages web contenant les mots ou les collocations. Ce nombre est retourné par des requêtes issues des moteurs de recherche (Google, Yahoo, Exalead, etc). Par exemple, $nb(\text{fichier})$ correspond au nombre de pages retourné avec le seul mot clé *fichier* et $nb(\text{fichier, de, client})$ correspond au nombre de pages retourné avec la requête "*fichier de clients*" (utilisation des guillemets pour rechercher une expression exacte). Les valeurs obtenues avec la mesure de Dice appliquée avec les deux requêtes « *fichier de clients* » et « *logiciel de ciel* » sont données ci-dessous.

$$Dice(\text{fichier, de, clients}) = \frac{3 \times 999.000}{37.200.000 + 6.350.000.000 + 208.000.000} = 0,000454$$

$$Dice(\text{logiciel, de, ciel}) = \frac{3 \times 89.800}{35.000.000 + 6.350.000.000 + 35.400.000} = 0,0000419$$

Ce résultat montre que le score le plus faible dans des proportions importantes (facteur dix) est donné par « *logiciel de ciel* ». Ainsi, notre mesure peut prédire que le candidat « *logiciel ciel* » de type Nom-Nom a statistiquement plus de chance d'être une EN comparativement à « *fichier clients* ». Ceci est tout à fait pertinent car cette EN fait référence à un logiciel de gestion et de comptabilité, ce qui correspond au type d'EN appelé *Document* [Daille *et al.*, 2000, Paik *et al.*, 1994].

Les mesures Web donnent une indication de popularité des mots/collocations tout à fait intéressante lorsque des données issues d'un domaine plus ou moins général sont traitées. Par ailleurs, l'avantage de ces connaissances « externes » au corpus (c.-à-d. Web) tient au fait que nous sommes moins sensibles à la taille des données traitées (c.-à-d. corpus). En effet, cette taille et donc la fréquence d'apparition des mots/collocations doit être assez significative lorsque des méthodes statistiques sont appliquées. Avec nos approches de type « Fouille du Web », nous n'avons pas de telles contraintes liées à la fréquence d'apparition des éléments dans les corpus eux-mêmes.

Etape 3 – Sélection

Les candidats à la collocation de type Nom-Nom qui obtiennent de faibles scores représentent des éléments peu enclins à la variation. Dans notre approche, de tels candidats seront considérés comme des EN. La section suivante évaluera la proportion de candidats sélectionnés qui représentent réellement des EN. Dans notre approche, nous allons introduire un paramètre S qui représente un seuil de sélection. Par exemple, avec un seuil $S=10$, les dix candidats ayant les scores les plus faibles seront sélectionnés comme EN potentielles. Les résultats selon différentes valeurs de S seront discutés dans la section « Expérimentations » de cet article.

Quid de notre approche en anglais ?

La terminologie nominale de type Nom-Nom possède des formes variantes différentes en anglais. Ainsi, les variantes fréquentes d'un candidat à la collocation de type Nom-Nom (par exemple, « knowledge discovery ») sont constituées d'une préposition associée à une permutation entre les noms (par exemple, « discovery of knowledge »). L'ensemble de ces règles pour caractériser les collocations variantes sont détaillées dans [Jacquemin, 1997].

Après avoir décrit, notre approche d'identification des EN à partir de candidats à la collocation, la section suivante présente les résultats expérimentaux obtenus sur des données réelles.

4. Expérimentations

Les corpus

Le premier corpus traité est composé de 1144 Curriculum Vitae (noté **CV**) fournis par la société *VediorBis* (120.000 mots). Une des particularités de ce corpus tient au fait qu'il est composé de phrases très courtes avec de nombreuses énumérations. Les travaux à partir de ce corpus consistaient à déterminer les concepts les plus significatifs pour le domaine [Roche et Kodratoff, 2006]. D'autres travaux sur ce même corpus avaient pour but de classer les CVs, c'est-à-dire classer les Curriculum Vitae en deux catégories : CVs de cadres et de non cadres [Clech et Zighed, 2003].

Le second corpus de spécialité étudié (noté **RH**) est composé d'un ensemble de textes également écrits en français qui sont issus du domaine des Ressources Humaines (société *PerformanSe* : <http://www.performanse.fr/>). Les textes correspondent à des commentaires de tests de psychologie de 378 individus (600.000 mots). Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique.

Extraction des candidats à la collocation

Le principe d'élagage des candidats à la collocation consiste à considérer seulement les candidats présents un nombre de fois minimum dans le corpus. L'élagage permet, dans la majeure partie des cas, d'exclure les candidats trop rares qui sont souvent peu représentatifs du domaine [Roche et Kodratoff, 2006]. Ainsi, classiquement un élagage à 3 est effectué [Jacquemin, 1997; Thanopoulos *et al.*, 2002]. Le tableau ci-dessous présente le nombre de candidats à la collocation obtenu avant et après élagage à 3.

Corpus	<i>Avant élagage</i>		<i>Après élagage</i>	
	RH	CV	RH	CV
Nom-Nom	98	1781	11	162
Nom-Préposition-Nom	4703	3634	1268	307
Adjectif-Nom	1260	1291	478	103
Nom-Adjectif	5768	3455	1628	448

Nombre de candidats à la collocation obtenus avant et après élagage.

Suivant les domaines de spécialité écrits dans une même langue, les résultats peuvent différer de manière importante. Par exemple, sur le corpus de CVs après élagage, le nombre de candidats à la collocation de type Nom-Nom (162) est beaucoup plus important que celui du corpus des Ressources Humaines (11) également écrit en français. Le corpus des Ressources Humaines a pourtant une taille cinq fois plus importante que le corpus de CVs. Ceci est dû au fait que les CVs sont écrits de manière condensée en employant un vocabulaire très spécifique : « *emploi solidarité* », « *action communication* », « *fichier client* », « *service achat* », etc. De tels candidats pourraient être assimilés à des collocations de type Nom-Préposition-Nom : « *emploi de solidarité* », « *action de communication* », « *fichier des clients* », « *service des achats* », etc. Dans la section suivante, nous allons nous appuyer sur les candidats à la collocation Nom-Nom du corpus de CVs. Nous allons filtrer les EN à partir de ces candidats en utilisant la méthode statistique présentée dans cet article.

Filtrage des Entités Nommées

Le but de cette section est d'estimer si les candidats à la collocation sélectionnés par notre approche représentent des EN réellement pertinentes. Dans ce cadre, nous nous sommes appuyés sur 70 candidats à la collocation de type Nom-Nom les plus fréquents qui sont estimés pertinents (en tant que terme ou EN). Ces candidats ont été évalués manuellement (18 EN ont été identifiées sur les 70 candidats).

Ces candidats ont alors été classés par la mesure de Dice décrite dans la section « Filtrage des Entités Nommées ». Nous avons donc évalué la qualité des candidats sélectionnés en utilisant différentes valeurs du seuil S (sélection des S candidats ayant les valeurs de Dice les plus faibles). L'objectif est alors d'évaluer si les candidats sélectionnés par notre système correspondent à des EN pertinentes.

Notons que les mesures appliquées (mesures de Dice) avec les candidats ont nécessité l'exécution automatique de 210 requêtes avec le moteur de recherche Exalead (<http://www.exalead.com/>) qui utilise des ressources en français assez riches. Nous avons alors effectué 70 requêtes pour les numérateurs et 140 requêtes pour les deux noms propres aux dénominateurs (les requêtes pour les prépositions ont été appliquées une seule fois pour l'ensemble des calculs).

Mesures d'évaluation du filtrage des Entités Nommées

De multiples critères d'évaluation sont disponibles dans le domaine de la fouille de textes. Nous citerons et appliquerons dans nos travaux les deux critères d'évaluation couramment usités que sont la *précision* et le *rappel*. La précision mesure le nombre d'EN pertinentes sélectionnées par rapport à l'ensemble des EN candidates retournées par un système. Le rappel mesure quant à lui le nombre d'EN pertinentes sélectionnées par rapport au nombre total d'EN pertinentes. Une précision de 100% signifie que toutes les EN extraites par le système sont correctes et un rappel de 100% signifie que toutes les EN correctes sont extraites.

Pour résumer, les mesures de précision et de rappel sont calculées de la manière suivante :

$$\text{Précision} = \frac{\text{nb EN candidates pertinentes}}{\text{nb EN candidates}}$$

$$\text{Rappel} = \frac{\text{nb EN candidates pertinentes}}{\text{nb EN pertinentes}}$$

Par ailleurs, une mesure très utilisée qui combine la précision et le rappel est la pondération nommée F-mesure :

$$F\text{-mesure} = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

Résultats du filtrage des Entités Nommées

Nous allons mesurer les résultats selon différents seuils (S) sur la base de ces trois critères. Les résultats sont présentés dans le tableau ci-dessous.

<i>Seuil de Sélection (S)</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>
10	0.60	0.33	0.43
20	0.45	0.50	0.47
30	0.37	0.61	0.46
40	0.35	0.78	0.48
50	0.36	1	0.53
60	0.30	1	0.46
70	0.26	1	0.41

Précision, Rappel et F-mesure selon différentes valeurs de S – Corpus de CVs.

Les résultats montrent que la meilleure valeur de F-mesure est obtenue lorsque nous considérons les 50 premiers candidats ($S=50$). Ceci s'explique par le rappel maximum (de valeur 1) car toutes les EN se situent parmi les 50 premiers candidats.

Les résultats du tableau montrent également que les premiers candidats sélectionnés sont assez souvent des EN avec notamment une précision de 60% pour les dix premiers candidats retournés ($S=10$). Ces derniers sont : *lotus note*, *ciel paie*, *agent recenseur*, *chauffeur livreur*, *go sport*, *rayon fruit*, *accueil client*, *france télécom*, *paris nord*, *front page*. Six candidats sont effectivement des EN (société, lieu, logiciel). Remarquons que deux candidats non pertinents ont été sélectionnés par notre approche (*agent recenseur*, *chauffeur livreur*) car, dans certains cas, notre méthode fondée sur la construction de termes variants de type prépositionnel n'est pas adaptée. En effet, les collocations construites (*agent de recenseur*, *chauffeur de livreur*) sont erronées. La mesure de Dice a donc retourné un score très faible pour ces collocations qui présentent peu de dépendance entre les trois mots les formant. Notre approche a alors naturellement déterminé ces collocations comme des EN potentiellement intéressantes. Dans ce cas, il aurait été plus pertinent de nous appuyer sur des règles de variation utilisant des conjonctions de coordination (*agent et recenseur*, *chauffeur et livreur*). Dans de prochains travaux, nous ajouterons de telles règles afin d'améliorer les résultats de précision.

Notons enfin qu'un classement aléatoire retourne une précision de 25% avec $S=10$. Ceci confirme donc que notre méthode, avec laquelle nous obtenons une précision de 60% dans les mêmes conditions, est tout à fait pertinente.

5. Conclusion et Perspectives

Cet article présente une méthode de fouille de textes permettant (1) d'extraire des candidats à la collocation, (2) de déterminer des Entités Nommées (EN) à partir de cette liste de candidats. La méthode de filtrage des EN s'appuie uniquement sur une approche statistique. Celle-ci utilise la mesure de Dice en exploitant les résultats de requêtes issues d'un moteur de recherche. Les EN étant *a priori* peu « stables », nous construisons des candidats variants et vérifions leur popularité via les moteurs de recherche. Si les candidats variants construits sont peu pertinents (c.-à-d. valeur faible de la mesure statistique), ils sont potentiellement considérés comme des EN.

Précisons que ces méthodes ne prétendent pas filtrer de manière exhaustive les EN mais permettent d'obtenir des résultats intéressants à présenter aux experts pour une phase de validation. Dans ce cas, nous devons, en général, privilégier une valeur élevée de précision. Dans le cas du traitement automatique de quantité importante de données (par exemple, pour des tâches d'Extraction d'Information ou de Recherche d'Information), il est souvent nécessaire de privilégier une valeur élevée de rappel même si les méthodes retournent du bruit.

Dans nos futurs travaux, nous envisageons d'enrichir les règles de recherche de variantes car cet article s'appuie sur des méthodes de base simples. Ce point reste donc crucial à améliorer afin de couvrir la grande majorité des variations linguistiquement pertinentes qui seront validées par les approches statistiques décrites dans ce document.

Enfin, nous combinerons ces approches uniquement statistiques pour prendre en compte certaines spécificités lexicales des mots, en particulier la présence de majuscules lorsque cela se révèle possible.

Bibliographie

Azé J., Heitz T., Mela A., Mezaour A.D., Peinl P., Roche M. *Présentation de DEFT'06 (DEfj Fouille de Textes)*. Dans les actes de l'atelier DEFT'06, SDN'06 (Semaine du Document Numérique), 2006

Bourigault D., Jacquemin C. *Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology*. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99), p.15-22, 1999.

Brill E. *Some Advances in Transformation-Based Part of Speech Tagging*. Proceedings of AAAI, Vol. 1, p. 722-727, 1994.

Clas A. *Collocations et langues de spécialité*. Meta, Vol 39, No 4, p. 576-580, 1994

Clech J, Zighed D.A. *Data Mining et Analyse des CV : Une Expérience et des Perspectives*. Actes de la conférences EGC, p.189-200, 2003

Daille B. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*. The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press. p.49-66, 1996

Daille B., Fourour N., Morin E. *Catégorisation des noms propres : une étude en corpus*. Cahiers de Grammaire, Volume 25, p.115-129, 2000.

Fort K., Ehrmann M., Nazarenko A. *Vers une méthodologie d'annotation des entités nommées en corpus*. Actes de TALN 2009 (Traitement Automatique des Langues Naturelles), 2009

Heid U. *Towards a corpus-based dictionary of German noun-verb collocations*. Proceedings of the Euralex International Congress, p. 301-312, 1998

Jacquemin C. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes, 1997.

Laurens M. *La description des collocations et leur traitement dans les dictionnaires*, In Romaneske, Vol 4, 1999

Melcuk I.A., Arbatchewsky-Jumarie N., Elnitsky L., Lessard A. *Dictionnaire explicatif et combinatoire du français contemporain*. Vol 1, 2, 3, 4, Presses de l'Université de Montréal, 1984, 1988, 1992, 1999

Paik W., Liddy E.D., Yu E., McKenna, M. Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. In B. Boguraev, & J. Pustejovsky (eds), *Corpus Processing for Lexical Acquisition*, MIT Press, chap. 4., 1994

Petrovic S., Snajder J., Dalbelo-Basic B., Kolar M. *Comparison of collocation extraction measures for document indexing*. Proceedings of Information Technology Interfaces (ITI), p.451-456, 2006.

Roche M., Kodratoff Y. *Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition*. Proceedings of onToContent'06 workshop (Ontology content and evaluation in Enterprise) - OTM'06, Springer-Verlag, LNCS, p.1107-1116, 2006

Roche M, Kodratoff Y. *Text and Web Mining Approaches in Order to Build Specialized Ontologies*. Journal of Digital Information (JoDI), Vol 10, No 4, 2009

Smadja F., McKeown K. R., Hatzivassiloglou V. *Translating collocations for bilingual lexicons : A statistical approach*. Computational Linguistics, vol. 22, No 1, p. 1-38, 1996

Thanopoulos A., Fakotakis N., Kokkianakis G. *Comparative Evaluation of Collocation Extraction Metrics*. Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02), p.620-625, 2002.

A propos des auteurs

Équipe TAL - LIRMM
UMR 5506, CNRS, Univ. Montpellier 2
34392 Montpellier Cedex 5 - France
mroche@lirmm.fr

<http://www.lirmm.fr/~mroche>