

Normalisation des entités nommées : pour une approche mixte et orientée utilisateurs

Vanessa Andréani

TecKnowMetrix – 4, rue Léon Béridot – ZAC Champfeuillet – 38500
Voiron – France

LIDILEM – Université Stendhal Grenoble 3 – Domaine universitaire – 1180,
avenue centrale – 38400 Saint Martin d'Hères – France
va@tkm.fr

Résumé La normalisation intervient dans de nombreux champs du traitement de l'information. Elle permet d'optimiser les performances des applications, telles que la recherche ou l'extraction d'information, et de rendre plus fiable la constitution de ressources langagières. La normalisation consiste à ramener toutes les variantes d'un même terme ou d'une entité nommée à une forme standard, et permet de limiter l'impact de la variation linguistique. Notre travail porte sur la normalisation des entités nommées, pour laquelle nous avons mis en place un système complexe mêlant plusieurs approches. Nous en présentons ici une des composantes : une méthode endogène de délimitation et de validation de l'entité nommée normée, adaptée à des données multilingues. De plus, nous plaçons l'utilisateur au centre du processus de normalisation, dans l'objectif d'obtenir des données parfaitement fiables et adaptées à ses besoins.

Abstract Normalization is involved in many fields of information processing. It improves performances for several applications, such as information retrieval or information extraction, and makes linguistic resources constitution more reliable. Normalization consists in standardizing each variant of a term or named entity into a unique form, and this way restricts the impact of term variation. Our work applies to named entity normalization, for which we implemented a complex system that mixes several approaches. We present here one of its components: an endogenous method to mark out and validate the normalized named entities. Moreover, we place the user in the center of our normalization process, in order to obtain fully reliable data that fit his needs.

Mots-clés : normalisation, entités nommées, traitement de l'information, analyse de corpus, méthodes endogènes, système complexe.

Keywords: normalization, named entities, information processing, corpus analysis, endogenous methods, complex system.

Introduction

Tout comme les unités lexicales « classiques », les entités nommées (EN) sont soumises à une grande complexité et à une variation linguistique importante. Leur normalisation est donc une

étape nécessaire pour un grand nombre de champs d'applications du Traitement Automatique des Langues (TAL) pour obtenir des données fiables et de qualité. C'est notamment le cas de l'extraction d'information, mais aussi de toutes les tâches qui traitent massivement les EN. Les méthodes de normalisation existantes relèvent la plupart du temps d'un seul type d'approche, ce qui les rend efficaces sur un aspect précis de la normalisation, mais parfois inopérantes sur d'autres points de difficulté. De plus, très peu de systèmes actuels placent l'utilisateur au centre du processus, alors que ce sont souvent ses connaissances et son point de vue qui permettent de réaliser une normalisation pertinente. Enfin, peu de systèmes de ce type sont portables d'une langue à l'autre sans une modification, parfois coûteuse, des ressources utilisées pour le traitement des données.

Dans une première partie de cet article, nous proposons de définir les notions liées au problème de la normalisation, et exposons les techniques existantes. Dans un deuxième temps, nous présentons le système de normalisation complexe que nous avons mis en place, puis démontrons l'intérêt d'une méthode de normalisation basée sur différents critères linguistiques et intégrant l'utilisateur. Enfin, nous détaillons notre approche endogène pour la normalisation, particulièrement adaptée à des données multilingues, et discutons nos premières évaluations.

1 La normalisation des entités nommées

Dans le cadre de l'exploitation de données textuelles pour l'analyse de l'information, la disparité des formes renvoyant à une même entité nommée (EN) ou à un même terme pose rapidement problème. Par conséquent, il est souvent nécessaire de procéder à une étape de normalisation de ces données, avant de débiter l'analyse à proprement parler.

Nous définissons la normalisation comme un processus permettant de ramener plusieurs formes de surface différentes renvoyant à un même référent à une forme standard, et dont la finalité est de reconnaître, dans des données structurées ou non, toutes les réalisations linguistiques pour une même entité.

Les travaux présentés ici portent exclusivement sur la normalisation des EN. A la suite de (Poibeau, 2001), nous considérons que les EN sont « l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné »¹. Parmi tous ces types, nous ne nous intéresserons qu'aux noms de personnes et d'organisations, seules entités qui apparaissent dans les données que nous souhaitons traiter. Les EN révèlent une complexité semblable à celle des unités lexicales « classiques » (Ehrmann, Jacquet, 2006), et sont donc soumises à la variation linguistique au même titre que les termes « communs ». Par conséquent, pour les domaines du TAL traitant massivement les EN, la normalisation de ces entités doit réduire l'impact de cette variation au minimum. Ainsi, les entités nommées *President John Kennedy*, *John F. Kennedy* et *John Fitzgerald Kennedy* renvoient toutes au 35^e président des Etats-Unis, John Fitzgerald Kennedy, malgré les différences graphiques qu'elles présentent, et doivent être traitées en conséquence. Nous nous intéressons donc à une normalisation au niveau graphique, destinée à homogénéiser les formes de surface pour une même entité.

Les champs concernés par cette étape de normalisation sont relativement nombreux. Pour la recherche d'information (RI), la possibilité d'identifier toutes les formes possibles d'une entité *via* une étape de normalisation est un moyen d'améliorer les performances en

¹ L'auteur inclut également dans cette catégorie « les dates, les unités monétaires, les pourcentages, etc. »

augmentant significativement le taux de rappel. Une telle étape présente le même avantage pour l'extraction d'information (EI) et les systèmes question-réponse. Pour la constitution de ressources termino-ontologiques (RTO), c'est-à-dire des « modèle[s] de connaissances comportant un réseau conceptuel et des termes associés » (Aussenac-Gilles, 2007), une étape préalable de normalisation permet de standardiser les EN qui formeront les instanciations des nœuds du réseau. Cette normalisation impacte la qualité des résultats fournis par les applications utilisant la RTO ainsi constituée. Enfin, la normalisation s'avère nécessaire pour le traitement de données structurées : elle permet de régler les cas d'entrées dupliquées dans les bases de données, c'est-à-dire des cas où plusieurs entrées réfèrent à la même EN mais présentent des différences graphiques qui empêchent de les traiter comme identiques (Elmagarmid *et al.*, 2007).

Nous pouvons classer les approches de normalisation d'EN en trois catégories : les méthodes qui s'appuient sur des ressources externes, les approches fondées sur l'utilisation de patrons, et les techniques adaptées au cas particulier des données structurées en bases de données.

Les approches prenant appui sur des ressources externes

Pour la normalisation d'EN dans leur acception la plus restrictive, à savoir les noms propres de lieux, d'organisations ou de personnes, une approche courante consiste à utiliser un dictionnaire ou toute ressource externe suffisamment exhaustive, de manière à ramener chaque occurrence d'une EN à l'entrée « standard » qui lui correspond dans la ressource. Cela permet de résoudre les difficultés liées à la variation linguistique, notamment en termes de synonymie. (Khalid *et al.*, 2008) règle ce problème en utilisant Wikipédia en tant que ressource externe : il se sert des liens de redirection du site pour la gestion de la synonymie. En comparant les résultats obtenus par un système de RI avec et sans normalisation en amont, il démontre qu'il est plus efficace lorsque les EN ont été préalablement normalisées.

Les méthodes fondées sur des patrons d'extraction

Dans le cadre de systèmes d'EI dans des domaines spécialisés, l'utilisation de patrons pour la normalisation permet de régler efficacement les problèmes de variation linguistique. Selon les systèmes, l'acception du terme d'*entité nommée* peut être assez large et inclure d'autres éléments que les EN courantes. Les auteurs de (Alphonse *et al.*, 2004), dans le cadre du projet Caderige, travaillent à l'extraction d'information, et plus particulièrement d'interactions géniques, à partir de textes du domaine biomédical. Dans un premier temps, ils procèdent à une normalisation des noms de gènes et de protéines, en l'occurrence considérées comme des EN. Des amorces de synonymie telles que « formerly » ou « also called », et des patrons tels que « gene amorce gene », permettent de gérer la variation linguistique. Ils démontrent que cette étape de normalisation « facilite l'acquisition et l'apprentissage de règles d'extraction en offrant une représentation plus abstraite des phrases ». Avec le même objectif, mais dans le domaine financier cette fois, (Poibeau, 2003) prône lui aussi une phase de normalisation des EN préalable à l'extraction, et la qualifie d'« étape [...] essentielle ».

Le cas particulier des données structurées

Le problème majeur qui se pose dans les données structurées est l'existence de doublons dans lesquels la forme des EN varie, rendant impossibles les recoupements. Ces doublons sont la conséquence d'erreurs de typographie, de l'utilisation d'abréviations, mais également de l'utilisation de plusieurs sources d'information, pour lesquelles les conventions graphiques peuvent varier (Jijkoun *et al.*, 2008). Les méthodes relèvent alors souvent de calculs de similarité entre champs, de manière à rapprocher deux ou plusieurs noms qui font référence au

même objet du monde. (Elmagarmid *et al.*, 2007) recense les méthodes utilisées pour la détection d'entrées dupliquées dans les BD, et donc pour leur normalisation.

Les méthodes utilisant des ressources exogènes sont efficaces sur du texte tout venant, *e.g.* des corpus de presse. Cependant, dès lors que les domaines abordés sont très spécialisés, ou que les EN sont des noms de petites entreprises par exemple, des ressources généralistes sont insuffisantes. L'alternative qui vise à constituer des ressources spécialisées est extrêmement coûteuse, sans garantie d'exhaustivité. Une méthode de normalisation par patrons peut quant à elle représenter une bonne alternative aux dictionnaires et autres ressources pour le traitement de textes spécialisés. Enfin, les calculs de similarité ont l'avantage d'être peu coûteux en termes de mise au point, et offrent des résultats satisfaisants pour certains aspects de la normalisation, en particulier pour la correction d'erreurs typographiques.

2 Un système de normalisation complexe

Notre travail intervient dans un contexte industriel, sur des données multilingues, et vise à normaliser les noms des organisations qui déposent des brevets ou publient des articles scientifiques. Les informations concernant ces publications sont stockées dans des bases de données contenant plusieurs millions d'entrées. L'objectif de cette normalisation est de fournir dans les tables contenant les noms d'organisations des données standardisées, et ce quelle que soit la langue employée, de manière à permettre des analyses de corpus fiables, qu'il s'agisse d'analyses statistiques, comme des comptages, ou d'analyses textuelles.

Nous avons conçu un système de normalisation des EN qui mêle plusieurs des méthodes que nous venons d'aborder, et qui implique la participation de l'utilisateur. Grâce à cette approche multiple, nous pouvons couvrir un grand nombre de cas de figure problématiques liés à la variation linguistique. Le système est composé de trois principaux modules :

- L'extraction et la réécriture des noms d'organisations grâce à des patrons ;
- La correction d'erreurs typographiques par des mesures de similarité entre les noms d'organisations de la base ;
- L'extraction du nom d'organisation le plus cohérent et le plus général via notre approche endogène, lorsque les deux premières étapes n'ont pas été suffisamment efficaces.

Le fait d'associer ces types d'approche permet de traiter les trois grandes catégories de difficultés rencontrées dans notre corpus :

- les erreurs typographiques et orthographiques courantes, comme pour *Mitsubisi* au lieu de *Mitsubishi*² ;
- la grande diversité des sources dont les données sont issues, ce qui entraîne une grande disparité dans la graphie et dans la description d'une même entité. C'est par exemple le cas pour *Mitsubishi Corp* et *Mitsubishi* ;

² Mitsubishi est une marque enregistrée de Mitsubishi Corporation

- pour les EN désignant des institutions publiques, il n'est pas rare de trouver à l'intérieur d'un même nom plusieurs sous-éléments d'une même EN. Un cas comme :

Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net

est assez représentatif. Cela pose problème pour dégager l'entité de niveau supérieur tout en mettant de côté les sous-organisations, qui viendraient bruyter les données.

Nous illustrons dans le tableau qui suit les traitements effectués par chaque module :

Modules \ Nom d'organisation brut	<i>Tulane-Xavier Center for Bioenvironmental Research, Department of Pharmacology, Tulan University Medical Center, LA 70112,USA.</i>
1. Extraction et réécriture par patrons	<i>Univ Tulan Medical Center</i>
2. Correction d'erreurs typographiques par mesures de similarité	<i>Univ Tulane Medical Center</i>
3. Extraction du nom d'organisation par méthode endogène	<i>Univ Tulane</i>

Tableau 1: traitements effectués par les modules du système de normalisation

L'utilisateur intervient à la fin des phases 2 et 3, et doit valider les suggestions de correction du système³. Notons que l'utilisateur a le choix de procéder ou non à cette vérification, en fonction des contraintes liées à son activité. De plus, chacun de ses choix est gardé en mémoire, ce qui permet au système d'apprendre les corrections adéquates pour un nom déjà rencontré auparavant. Ainsi, le nombre de corrections à valider par l'utilisateur diminue au fil des normalisations. Cette supervision par l'utilisateur expert du domaine traité permet de garantir la qualité et la fiabilité des données qui seront stockées dans les bases.

Placer cette étape endogène à la fin du traitement permet d'éliminer un maximum de bruit par des méthodes de *pattern matching* peu coûteuses en temps de calcul, et de concentrer ensuite la procédure endogène, plus lourde en termes de coût calculatoire, sur les cas irrésolus.

3 Une approche endogène pour la normalisation

3.1 Principe et objectifs

Un système endogène trouve les informations dont il a besoin dans les données qu'il doit traiter. Ces informations permettent de résoudre des cas problématiques pour lesquels les méthodes à base de règles et de patrons atteignent leurs limites.

³ Cette validation se fait *via* une interface dédiée, dans laquelle l'utilisateur doit cocher la suggestion la plus pertinente. Il a aussi la possibilité de rentrer manuellement un nom d'organisation, lorsque les suggestions ne sont pas adaptées.

En France, les principaux travaux dans ce domaine ont été menés avec pour objectif de réaliser des analyses syntaxiques et / ou morphologiques là où les systèmes préexistants n'étaient pas assez efficaces en raison d'une variation linguistique trop importante (Vergne, 2004 ; Bourigault, 1993 ; Bourigault et Frérot, 2006). (Frérot *et al.*, 2003) démontre que des approches de ce type, c'est-à-dire des « procédures non supervisées d'apprentissage sur corpus qui [...] permettent d'exploiter le corpus d'analyse pour acquérir les informations nécessaires » aux traitements, sont particulièrement efficaces lorsque le domaine des textes à traiter n'est pas connu à l'avance, et que des ressources constituées *a priori* sont par conséquent peu adaptées. Ces méthodes se fondent essentiellement sur le principe de productivité, c'est-à-dire sur le nombre de contextes différents avec lesquels peut apparaître le mot ou le terme étudié, et s'appuient de façon privilégiée sur les redondances du corpus.

Une autre approche consiste à utiliser la longueur et la fréquence des mots pour réaliser une analyse syntaxique partielle sans ressources (Vergne, 2004). Le fait d'utiliser de tels critères d'analyse permet de traiter des textes multilingues sans coût supplémentaire, puisqu'aucun lexique n'est nécessaire.

Notre objectif diffère de ceux des travaux évoqués ci-avant, puisque nous cherchons pour notre part à normaliser des noms d'organisations en vue d'un traitement de l'information stratégique. Cependant, une approche endogène fondée sur la récurrence de séquences nous paraît tout à fait appropriée au vu de nos données, et peut être un moyen de déterminer la structure d'un nom d'organisation. En effet, en nous fondant sur les fréquences de segments de différentes longueurs, et non plus sur des structures syntaxiques ou des indices lexicaux, nous évitons plusieurs écueils. Tout d'abord, cela évite d'avoir à constituer des lexiques adaptés, ce qui serait très coûteux puisque nous travaillons sur des données multilingues. D'autre part, du fait que nous traitons exclusivement des noms propres, aucun dictionnaire ne serait suffisamment exhaustif. Enfin, une telle approche permet d'avoir un traitement parfaitement adapté au corpus traité, quel que soit le domaine : un système endogène est forcément adapté au corpus qu'il traite, puisque par essence, les données qu'il utilise pour traiter le corpus sont issues du corpus lui-même.

3.2 Application à nos données

Les données que nous normalisons permettent de mener des analyses textuelles et statistiques et de tirer parti d'informations stratégiques. Ainsi, à partir des séquences suivantes :

1. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, Whitney-Hendrickson Building, Lexington 40536, USA.*
2. *University of Kentucky, Lexington 40536-0098, USA. runge@pop.uky.edu*
3. *University of Kentucky Research Institute, USA*

une méthode endogène devra permettre d'extraire la séquence récurrente *University of Kentucky*, qui représente le nom normalisé que doivent prendre ces trois noms d'organisations bruts. De cette manière, il sera possible d'identifier toutes les publications émanant de cette organisation, et ce quel que soit le département ou la division.

Nous avons réduit l'utilisation des ressources exogènes au minimum : nous avons notamment établi une liste d'une centaine d'amorces, c'est-à-dire de mots qui permettent de détecter la présence d'une entité nommée, ainsi que son niveau hiérarchique. Par exemple, une université, détectée par l'amorce *Univ* suivie ou non de plusieurs caractères (*University*,

Université, ...) est considérée comme hiérarchiquement supérieure à un centre, repéré par l'amorce *Center*, *Centre*, *Centro*, etc. Nous avons en effet observé que lorsque ces deux types d'entités nommées se trouvaient dans le même nom d'organisation brut, le centre était rattaché à l'université, et représentait donc une « sous-partie » de l'université en question.

Pour mettre en pratique une approche endogène adaptée à nos besoins, nous nous fondons sur deux calculs : la fréquence des sous-séquences et la surface des sous-séquences.

La fréquence des sous-séquences

Nous définissons une séquence comme un segment d'un nom d'organisation placé entre deux virgules. Une sous-séquence est donc une suite de n items issus de cette séquence, dans une fenêtre dont la taille s'incrémente d'un item à chaque passe.

Grâce à notre liste d'amorces, nous pouvons déterminer que l'amorce de plus haut niveau en 1 est *University* : nous ne conserverons donc que cette séquence. En revanche, elle pose problème puisqu'elle comporte une deuxième amorce, *Center*, de niveau inférieur. Toute la difficulté est donc de délimiter le nom d'organisation le plus cohérent au sein de cette séquence, c'est-à-dire d'extraire la sous-séquence la plus pertinente. C'est l'étape endogène qui permettra de déterminer que *University of Kentucky* doit être conservé comme nom d'organisation, et non, par exemple, *University of Kentucky Medical*.

Nous partons de l'hypothèse selon laquelle la sous-séquence la plus fréquente est le nom d'organisation le plus cohérent. Pour parvenir à extraire ce segment, nous comptabilisons le nombre d'occurrences de chaque sous-séquence de *University of Kentucky Medical Center* comportant l'amorce de niveau supérieur, en l'occurrence *University of*. Le cas échéant, les premiers mots sélectionnés sont ceux situés à gauche de l'amorce, du voisin immédiat jusqu'au plus lointain. Cela se justifie par le fait que la plupart des données problématiques à ce stade sont des données en anglais, et que la qualification des noms dans cette langue se fait souvent de droite à gauche. Puis les mots placés à droite sont à leur tour ajoutés un par un. Les sous-séquences testées pour notre exemple sont donc :

(1.a) *University of Kentucky*

(1.b) *University of Kentucky Medical*

Le système ne va pas plus loin, puisqu'il rencontre après *Medical* la seconde amorce, *Center*, qui n'appartient pas au nom d'organisation le plus cohérent.

Pour chacune de ces sous-séquences, le système va donc calculer leur fréquence dans l'intégralité de notre corpus de plusieurs millions de noms d'organisations. Dans ce corpus, nous avons relevé 71 occurrences de la suite (1.a) et 9 occurrences de la suite (1.b). De fait, en sélectionnant la sous-séquence la plus fréquente, nous sommes bien en mesure de détecter le nom d'organisation le plus cohérent, soit : (1.a) *University of Kentucky*.

Les mêmes calculs sont effectués sur les deux autres séquences, soit *University of Kentucky* et *University of Kentucky Research Institute*. *University of Kentucky* étant toujours la séquence la plus fréquente, nous considérons qu'il s'agit du nom d'organisation le plus cohérent.

La surface des sous-séquences

Les calculs de fréquence permettent de couvrir un grand nombre de cas semblables à ceux que nous venons d'évoquer. Cependant, il reste des problèmes non résolus. Ainsi les exemples :

4. *New York University Medical School*

5. *New York University Dental Center*

Ces séquences présentent un problème supplémentaire, puisque le nom de l'université, soit *New York*, est un mot composé, et que le nom d'université *York* existe par ailleurs. Le calcul de fréquences de sous-séquences donne les résultats suivants :

- *York University* 182 occurrences
- *New York University* 174 occurrences
- *New York University Medical* 24 occurrences
- *New York University Dental* 1 occurrence

En suivant notre hypothèse de départ, le nom d'organisation le plus cohérent serait donc *York University*, avec 182 occurrences. Or, nous savons que dans les exemples que nous étudions, le nom devrait être *New York University*. Pour résoudre cette difficulté, nous avons mis en place un calcul de surface des sous-séquences, qui met en rapport leur fréquence et leur longueur. Le calcul de surface est le suivant :

$$\text{Surface} = \text{nombre de tokens} * \text{nombre d'occurrences}$$

Nous présentons les résultats de ces calculs sur la figure 1 :

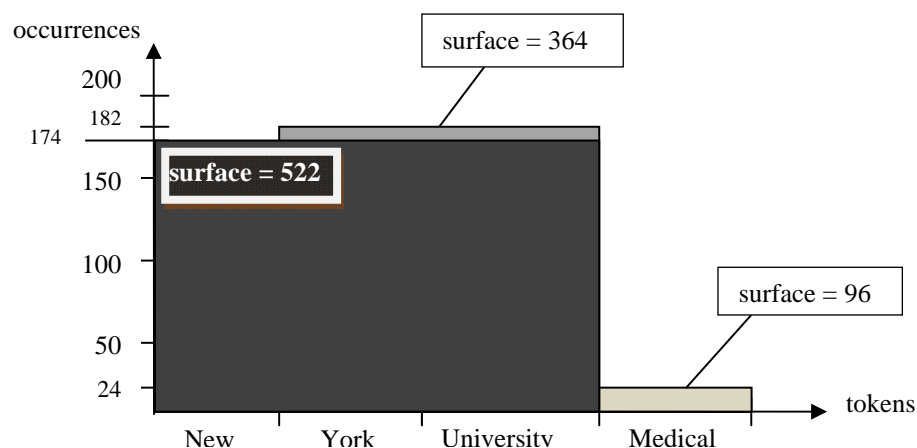


Figure 1: Calcul de surfaces pour les sous-séquences de *New York University Medical Center*

Ici, le meilleur score de surface est celui de la séquence *New York University*, soit le nom que nous avons identifié manuellement comme le plus cohérent. Les calculs de surface nous permettent donc, dans un certain nombre de cas, de résoudre le problème des mots composés.

3.3 Evaluation

Nous exposons ici les résultats d'une évaluation manuelle sur un échantillon de données. Par la suite, nous mènerons une évaluation plus poussée sur un plus grand nombre de données, pour chaque bloc de traitement ainsi que pour l'ensemble de notre système.

Pour l'heure, sur un échantillon composé de 15 noms d'organisations bruts, dont la majorité contiennent des noms d'organisations composés et donc les cas les plus difficiles, le calcul de fréquences permet de déterminer le nom d'organisation le plus cohérent dans 8 cas. Sur les mêmes données, le calcul de surfaces extrait 10 noms d'organisations correctement. Les cas

mal identifiés quant à eux, permettent d'identifier les difficultés qui expliquent que le calcul de surface ne soit pas efficace. Soit le nom brut suivant :

6. *Nagoya City University Medical School*

Notons que nous considérons *Nagoya City* comme un nom composé, puisqu'il s'agit du nom complet de l'université. Les trois sous-séquences testées sont les suivantes :

Sous-séquence	Fréquence	Surface
<i>City University</i>	151	302
<i>Nagoya City University</i>	48	144
<i>Nagoya City University Medical</i>	15	60

Tableau 2: Fréquences et surfaces pour les sous-séquences de *Nagoya City University Medical School*

Le calcul de fréquence comme le calcul de surface sont inefficaces sur ce nom. En effet, s'ils permettent d'éliminer *Medical* sans équivoque, ils ne permettent pas de conserver *Nagoya*, puisque la fréquence et la surface de *Nagoya City University* sont inférieures à celles de *City University*. La présence de *City*, un mot trop « générique » et donc trop fréquent dans notre corpus, empêche une normalisation correcte de ce nom d'organisation.

Pour dépasser cette difficulté, nous envisageons de coupler ces calculs à un calcul de fréquence des items de manière isolée. Nous avons pu observer en corpus que lorsqu'un mot voisin d'une amorce était le moins fréquent d'une séquence, il permettait de situer le point de rupture entre le nom d'organisation cohérent et le « bruit ». Généralement, ce mot peu fréquent est inclus dans le nom d'organisation, et la rupture se situe juste avant ou après lui. Par exemple, pour le nom 6, nous avons les fréquences suivantes :


tokens	Nagoya	City	University	Medical
fréquences	756	3153	40118	14930
courbe				

Figure 2: Courbe des fréquences des tokens pour la sous-séquence *Nagoya City University Medical*

De cette manière, nous pouvons déterminer que *Nagoya* fait bien partie du nom d'organisation le plus cohérent. Le calcul de fréquences ou de surfaces permettra d'éliminer *Medical*, et nous obtenons donc en sortie du traitement *Nagoya City University* comme nom le plus cohérent.

Conclusion

A travers cet article, nous avons abordé le problème complexe de la normalisation des entités nommées, situé en amont de différentes tâches de traitement de l'information. Les retombées impliquées par une normalisation précise des entités nommées sont très importantes lorsque les tâches de traitement de l'information doivent atteindre un degré de précision important, comme c'est le cas pour la plupart des acteurs du domaine du traitement automatique des langues et de l'ingénierie linguistique.

Une approche endogène permet de couvrir des cas non résolus par d'autres méthodes, particulièrement dans des domaines spécialisés pour lesquels il n'existe pas ou peu de ressources, et d'autant moins lorsque le travail porte sur les entités nommées.

Le fait d'allier cette méthode à d'autres techniques, comme l'utilisation de patrons ou des mesures de similarité, permet d'obtenir une couverture bien plus large que ce vers quoi nous pourrions tendre avec un système n'utilisant qu'un seul type de processus. De cette manière, tous les points de difficulté peuvent être traités. De plus, l'intervention de l'utilisateur dans la constitution des données garantit leur qualité et leur fiabilité, indispensables aux traitements ultérieurs qui prennent ces informations en entrée.

Références

ALPHONSE E., AUBIN S., BESSIERES P., BISSON G., HAMON T., LAGARRIGUE S., NAZARENKO A., NEDELLEC C., OULD ABDEL VETAH M., POIBEAU T., WEISSENBACHER D. (2004). Extraction d'information appliquée au domaine biomédical - apprentissage et traitement automatique de la langue. Actes de *CIFT*.

AUSSENAC-GILLES N. (2007). *Projet DaFOE4App - Dossier A.0 / Document A.0.1 - Etat de l'art et étude des besoins pour une plateforme de construction d'ontologies*. Rapport de contrat, IRIT/RT—2007-2—FR, IRIT.

BOURIGAULT D. (1993). An endogenous Corpus Based Method for Structural Noun Phrase Disambiguation. Actes de la *Conference of the European Chapter of ACL (EACL)*, 81-86.

BOURIGAULT D., FREROT C. (2006). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *TAL* 47, 141-154.

EHRMANN M., JACQUET G. (2006). Vers une double annotation des Entités Nommées. *TAL* 47, 63-88.

ELMAGARMID A.K., IPEIROTIS P.G., VERYKIOS V.S. (2007). Duplicate Record Detection : A Survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1-16.

JIJKOUN V., KHALID M.A., MARX M., DE RIJKE M. (2008). Named Entity Normalization in User Generated Content. Actes de *SIGIR 2008 – Workshop on Analytics for Noisy Unstructured Text Data*.

KHALID M.A., JIJKOUN V., DE RIJKE M. (2008). The Impact of Named Entity Normalization on Information Retrieval for Question Answering. *LNCS* 4956.

POIBEAU T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Paris : Hermès.

VERGNE J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. Actes des *JADT 2004* 2, 1158-1164.