Création automatique de dictionnaires bilingues d'entités nommées grâce à Wikipédia

Jean-Philippe Goldman & Yves Scherrer
Département de Linguistique
Université de Genève
{jean-philippe.goldman ; yves.scherrer}@unige.ch

Résumé

Nous présentons ici une étude visant à exploiter la richesse lexicale de l'encyclopédie en ligne Wikipédia. Cette ressource contient des informations structurées et traduites dans de nombreuses langues, ce qui en fait un outil idéal pour la création de dictionnaires bilingues. Nous nous focalisons ici sur un certain type d'entrées lexicales, appelées entités nommées, souvent composées de plusieurs mots. Celles-ci désignent des notions comme des personnalités, des organisations, des lieux ou encore des dates. Un des buts de ce travail est l'aide à la traduction automatique, ici pour la paire de langues français-anglais dans les deux sens, afin d'éviter des écueils grossiers comme traduire « Bush » par « buisson » quand il s'agit d'un président.

1. Introduction

Les lexiques bilingues sont indispensables tant à la traduction humaine qu'à la traduction automatique, et tout autant pour l'extraction d'informations multilingues ou encore pour l'apprentissage des langues. Ils sont très coûteux à enrichir manuellement, et plus spécialement pour ce qui concerne les lexiques spécialisés (dans un domaine particulier). Dans le cadre de la traduction automatique statistique, on utilise largement des corpus de textes multilingues parallèles afin d'en extraire des correspondances lexicales bilingues, mais les inexactitudes et les libertés prises par les traducteurs dans ces corpus peuvent nuire à la qualité des entrées lexicales inférées. De plus, le genre du corpus a une grande influence sur le vocabulaire pouvant être extrait. C'est pour ces raisons que nous nous sommes tournés vers l'encyclopédie libre Wikipédia afin d'en extraire des correspondances lexicales bilingues pour des noms de personnes et d'organisations (ce qu'on appelle habituellement des entités nommées), dans le but d'augmenter un lexique de traduction automatique existant et couvrant de manière satisfaisante le langage courant (Wehrli et al. 2009). Par rapport à des corpus de textes bruts, Wikipédia contient de nombreux titres et hyperliens. Cette

structuration facilite notamment l'extraction et la classification de données.

Dans la section suivante, nous définissons le concept d'entité nommée et passons en revue les différentes méthodes qui existent pour les extraire. Nous y explicitons également les motivations de notre travail. Dans la section 3, nous présentons l'encyclopédie Wikipédia et mettons en avant son utilité pour une telle tâche. Dans les sections 4 et 5, nous présentons la méthode d'extraction d'entités nommées multilingues, et nous confrontons le dictionnaire créé à un texte en anglais et en français pour estimer la représentativité des correspondances bilingues extraites. Nous filtrons le dictionnaire obtenu pour n'en retenir que les entrées apparaissant dans le texte. Enfin, nous expliquons en section 6, les méthodes mises au point pour enrichir chaque entité des traits grammaticaux nécessaires à son insertion dans un lexique de traduction automatique.

2. Entités nommées

On appelle *entité nommée* un mot ou un groupe de mots désignant une personne, une organisation ou entreprise, un lieu, une date ou encore une quantité. De nombreux travaux actuels se focalisent sur la reconnaissance, l'extraction et la classification des entités nommées dans un texte. Ainsi, dans un texte donné, il s'agit de décider quels mots font partie d'une entité nommée, et quel type d'entité nommée ce(s) mot(s) désigne(nt). En général, l'extraction d'entités nommées est faite soit à base de reconnaissance de patrons, soit de manière stochastique à l'aide d'un corpus d'entraînement manuellement annoté (voir par exemple, la compétition MUC-7 sur l'extraction des entités nommées (Chincor 1998)). Il existe également des techniques pour exploiter les ressources structurées comme Wikipédia pour établir une liste exhaustive d'entités nommés (Kazama & Torisawa 2007, Nothman et al. 2008, Nothman et al. 2009, Balasuriya et al. 2009). Ces travaux ont été menés dans une optique monolingue, en se concentrant sur une seule langue, le plus généralement l'anglais (notamment à cause des corpus d'entraînement disponibles pour cette langue).

Dans cette contribution, notre objectif est un peu différent car il s'agit d'extraire des paires bilingues d'entités nommées, c'est-à-dire des entités anglaises avec leur traduction en français. Ces paires bilingues sont ensuite utilisées dans le cadre de la traduction automatique. En effet, elles peuvent résoudre de nombreux problèmes rencontrés actuellement:

- En cas d'homographie entre une entité nommée (inconnue) et un nom commun (connu), un traducteur automatique préférera le

nom commun. Par exemple, *President Bush* est traduit incorrectement par *Président buisson*, l'initiale majuscule pouvant toutefois lever l'ambigüité dans certains cas.

- Un traducteur automatique pourrait tenter de décomposer un mot inconnu au lieu de le laisser tel quel. Cela arrive fréquemment en allemand; par exemple, le nom de ville *Mannheim* serait traduit littéralement par *foyer d'homme*.
- Une entité nommée à mots multiples ne devrait pas être séparée au milieu. La date *20 March 2005* pourrait être traduite par *20 marchent en 2005*, mais la structure figée de la date rend cette traduction très improbable.

À notre connaissance, il existe peu d'études avec une orientation multilingue. Yu & Tsuji (2009) utilisent le contenu des pages Wikipédia pour en extraire des entités nommées bilingues. Ils utilisent des mesures de similarité sémantique (contextuelle) et syntaxique pour apparier les entités des deux langues. Erdmann et al. (2008) utilisent une approche similaire à la nôtre pour inférer des correspondances bilingues en utilisant uniquement les titres et les liens des articles Wikipédia en vue de créer un lexique bilingue complet. Ils remarquent notamment que des mots du langage courant n'apparaissent généralement pas sous forme d'article complet. De plus, ils utilisent seulement des mots simples, pas des expressions à mots multiples.

Notre approche combine ces deux travaux: notre but est d'inférer seulement des entités nommées bilingues, et non pas un lexique complet. Dans ce cas, il est très judicieux de se focaliser sur les entités à mots multiples. Cependant, nous nous limitons aux titres et liens des articles Wikipédia; nous utilisons le contenu uniquement pour déterminer certains traits syntaxiques (voir section 6).

3. Wikipédia

Wikipédia est une encyclopédie collaborative en ligne, lancée en 2001. Elle est universelle, c'est-à-dire que son domaine de connaissance n'est pas limité et peut couvrir ainsi la quasi-totalité d'une langue. Elle est collaborative puisqu'elle est constamment alimentée par des contributeurs volontaires. Elle est libre de droit et donc exploitable à des fins de recherches. Les archives complètes sont mises à jour en permanence et librement téléchargeables.

Wikipédia est disponible dans 281 langues différentes. La version anglophone, la plus riche, compte plus de 3.5 millions d'articles, la version francophone plus d'un million d'articles.[1]

[1] Les chiffres reproduits ici reflètent l'état de Wikipédia en juillet 2011.

La caractéristique principale de Wikipédia est la présence de liens entre les articles. En effet, si un article mentionne un concept qui fait aussi l'objet d'un article, un hyperlien permet généralement de passer vers ce second article. Par ailleurs, si un article possède une traduction, celle-ci sera accessible par un lien en colonne gauche. Ce sont précisément ces liens interlangues qui nous intéressent ici.

Notons aussi que la qualité et la forme des articles peuvent être très variables. De plus, de par sa nature ouverte et libre, Wikipédia pourrait paraître vulnérable à des changements de contenu fortuits, au non-respect de la neutralité d'opinion ou encore à des conflits d'intérêts. Malgré ces inconvénients – généralement corrigés rapidement par la communauté des contributeurs – Wikipédia reste un des plus beaux succès de la ressource libre.[2]

Les articles de Wikipédia sont semi-structurés: ils comportent du texte courant, mais aussi des titres, des liens vers d'autres articles et d'autres langues, des boîtes d'information, des informations de catégorisation ontologique, des listes et des références. Les discussions et modifications des contributeurs sont également consultables.

[2] Notre travail se base presque exclusivement sur les titres des pages Wikipédia, bien moins sujets à discussion que le contenu des pages.

Figure 1 : *Exemple d'article francophone avec infobox (à droite), hyper-liens intra-langue en texte (vers fromage, lait de vache,...), et inter-langue à gauche (vers catalan,…) et mention d'une redirection (sous le titre)*

Wikipédia est utilisée par de nombreux linguistes informaticiens pour des applications aussi variées que la construction d'ontologies et taxonomies (Strube & Ponzetto 2006, Ponzetto & Strube 2007, Medelyan et al. 2009), la désambiguïsation sémantique (Mihalcea 2007), les systèmes de question-réponse (Jijkoun & de Rijke 2006), la lexicologie et la traduction.

En plus des articles, qui sont des pages web consultables directement, il existe deux autres types de pages importants: les pages de redirection et les pages de désambiguïsation (ou homonymie).

Le mécanisme de **redirection** est mis en œuvre lorsqu'une entité peut être désignée de plusieurs manières. Il y a alors un renvoi automatique vers la page contenant les informations attendues. Par exemple, la requête *Fromage à raclette* renvoie vers *Raclette*. Les pages de redirection sont utilisées pour[3]:

[3] Exemple pris de http://fr.wikipedia.org/wiki/Aide:Redirection

- des abréviations: *SNCF* redirige vers *Société nationale des chemins de fer français*;
- des synonymes: *e-mail*, *courriel*, *mél* et *messagerie électronique* redirigent tous vers *courrier électronique*;
- des noms alternatifs: *Karol Wojtyła* redirige vers *Jean-Paul II*;
- capturer des erreurs probables: *Camberra* redirige vers *Canberra*; ceci empêche notamment la création répétée de doublons sous le nom fautif et rend la fonction de recherche plus facile.

Pour illustrer les deux derniers points, citons l'article *George W. Bush* qui peut être atteint depuis de nombreuses pages comme:

- *G.W. Bush* (sans espace)
- *G. W. Bush* (avec espace)
- *George Bush Jr.*
- *George Walker Bush*
- *Georges Bush* (avec un *s* au prénom)

Lorsqu'une requête ou un terme est trop ambigu, une page de désambiguïsation (ou homonymie) donne une liste des articles pertinents. Les abréviations courtes sont souvent concernées (*PAF* peut faire référence à *police aux frontières, patrouille de France, paysage audiovisuel français*), mais aussi des entités en toutes lettres comme *Mercure* (planète, élément chimique, dieu romain ou encore personnalités dont c'est le patronyme). Pour éviter de nombreux cas d'homonymie, les contributeurs sont incités à respecter des conventions sur les titres (ponctuation, ordre des mots) et d'ajouter si nécessaire une indication entre parenthèse: *Mercure (planète)*, *Mercure (mythologie)*, etc... Ces conventions doivent évidemment être prises en compte pour notre tâche de constitution d'un dictionnaire bilingue à partir de Wikipédia.

4. Création du répertoire bilingue

Dans cette première étude nous nous sommes limités aux correspondances entre le français et l'anglais. Après avoir téléchargé les fichiers d'archives de Wikipédia, 4 250 002 titres d'articles en français étaient disponibles. Après filtrage des pages de redirection, d'homonymie et des pages contenant des listes[4], il restait 3 990 623 titres. De plus, nous avons retenu uniquement les pages françaises qui étaient en lien avec une page anglaise. Ce filtre réduisait le nombre de titres à 759 459. Finalement, un filtrage similaire appliqué aux titres anglais (redirection, homonymie et listes) laisse 655 071 paires d'entités nommées.

[4] Par exemple, il existe une page nommée *Liste des Premiers ministres de Belgique*. Son titre ne dénote pas une entité nommée, et il convient donc de le retirer.

Parmi ces 655 071 entités nommées restantes, on compte des expressions qui vont jusqu'à 20 mots en français comme

(1) Décennie internationale de la promotion d'une culture de la non-violence et de la paix au profit des enfants du monde

ou 18 mots en anglais:

(2) United Nations Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families

Pour ce genre de syntagmes nominaux complexes, la traduction est transparente et ne devrait pas produire d'erreurs. Mais leur caractère d'expression figée et leur grande longueur incite à les lexicaliser. Il existe d'autres titres longs comme

(3) Ecole supérieure de physique et de chimie industrielles de la ville de Paris

pour lesquels l'article Wikipédia en anglais garde un titre en français[5]. En anglais, il existe des cas similaires, comme:

(4) Sex and the City[6]

La traduction automatique a grand intérêt à connaître ces particularités car il serait malheureux de tenter une traduction. Il existe encore des cas pour lesquels la traduction littérale n'a aucune chance:

(5) Mercedes Cup ~ Tournoi de Stuttgart

(6) Two Cars in Every Garage and Three Eyes on Every Fish ~ Sous le signe du poisson[7]

Nous pouvons donc noter pour ces cas, qu'une correspondance bilingue mémorisée est tout à fait pertinente, soit parce que les expressions sont longues et figées, soit parce qu'elles n'ont pas lieu d'être traduites ou encore parce qu'elles sont non-transparentes.

Signalons à l'inverse des cas critiques de non-correspondance entre les titres anglais et français de Wikipédia, l'un contenant plus d'information que l'autre.

(7) Politburo of the Central Committee of the Communist Party of the Soviet Union ~ Politburo[8]

(8) Conference of Community and European Affairs Committees of Parliaments of the European Union ~ COSAC

(9) $1 - 2 + 3 - 4 + \cdots$ ~ Série alternée des entiers

---

[5] Une visite de la version anglaise du site officiel www.espci.fr confirme qu'il n'existe pas de traduction officielle du nom de l'école en anglais.

[6] Traduit pourtant au Québec par *Sexe à New York.*

[7] Quatrième épisode de la saison 2 de la série télévisée d'animation *Les Simpson*. *Poisson nucléaire* est le titre en québécois.

[8] A noter que Wikipédia contient également la correspondance *politburo ~ bureau politique*.

Ceci incite à ne pas accepter aveuglément l'ensemble des correspondances des entités nommées récoltées.

Deux observations supplémentaires peuvent être faites sur cette liste de 655 071 correspondances bilingues. D'une part, on remarque que 57% des termes sont identiques dans les deux langues. Il s'agit surtout de noms de personnes et de noms de lieux. À l'inverse, les abréviations, les acronymes (*UNO ~ ONU*), les noms de pays (*Mexique ~ Mexico*) et de grandes villes (*Londres ~ London*), les institutions, les titres d'œuvres, et les positions officielles sont généralement traduites. D'autre part, 30% des entités en anglais sont à mot unique (contre 34% pour la liste en français).

Anglais	identique	non-identique	Total
mot unique	24%	6%	30%
mots multiples	33%	37%	70%
total	57%	43%	100%

Français	identique	non-identique	Total
mot unique	24%	10%	34%
mots multiples	33%	33%	66%
total	57%	43%	100%

Tableau 1 : *Pourcentages d'entités nommées à mot(s) unique/multiples et de traductions semblables pour l'anglais et le français*

D'après ce tableau, on pourrait supposer une légère propension à traduire des titres anglais à mots multiples (37% de multimots pour 6% de monomots) vers un titre à un seul mot en français (resp. 32% et 10%). En réalité, cette différence de 4% correspond à 6% des entrées qui ont effectivement plusieurs mots en anglais et un seul en français, mais qui est compensée par 2% pour la situation inverse. Dans un sens comme dans l'autre, une grande majorité est due à un choix de rédaction comme évoqué plus haut pour COSAC (expansion d'un acronyme) ou Politburo (surspécification dans une des langues). Mais la raison peut aussi être inhérente aux langues comme pour:

(10) 1960s ~ Les années 1960

(11) mooncake ~ gatêau de lune

(12) Republic of Ireland ~ République Irlandaise

(13) Irish language ~ irlandais

(14) Irving Texas ~ Irving

Il existe également, mais en très petite quantité, des expressions à mots multiples qui ne sont pas considérées comme entités nommées mais plutôt comme collocations (Nerima et al. 2006), mais qui sont tout aussi utiles pour la traduction:

(15) graph paper ~ papier millimétré

5. Pertinence par rapport à un corpus textuel

Pour évaluer la pertinence de cette liste d'entités nommées, il convient de la confronter à un corpus textuel externe pour chacune des deux langues. Nous avons choisi dans un premier temps un extrait d'un corpus parallèle de nouvelles journalistiques. Comme ces textes ont les mêmes contenus en anglais et en français, l'effet du filtrage des entités nommées sera équivalent dans les deux langues. Ce corpus de nouvelles journalistiques (news) compte environ 2,5 millions de mots pour l'anglais et 2,9 millions de mots pour le français.[9]

Il a fallu repérer automatiquement dans chacune des phrases du texte la présence éventuelle d'une ou plusieurs entités nommées parmi notre liste bilingue de 655 071 entrées. Si plusieurs entités se partageaient un ou plusieurs mots du texte, alors la plus longue était privilégiée.

Au final 5926 expressions uniques ont été relevées dans le corpus anglais et 4757 en français. Dans ce dernier, parmi les plus longues repérées, il y a par exemple:

(16) Office de secours et de travaux des Nations unies pour les réfugiés de Palestine dans le Proche-Orient (17 mots - 1 occurrence)

(17) Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes (13 mots - 1 occ.)

(18) Fonds mondial de lutte contre le SIDA, la tuberculose et le paludisme (12 mots - 3 occ.)

Pour ce qui concerne le nombre d'occurrences, voici les plus fréquentes:

(19) Amérique latine (2 mots - 505 occurrences)

(20) Corée du Sud (2 mots - 247 occ.)

(21) George W. Bush (3 mots - 222 occ.)

(22) Guerre froide (2 mots - 187 occ.)

On constate également qu'un grand nombre d'expressions repérées sont à mot unique. Elles représentent 27% des types d'expressions en anglais et 66% en français. Elles peuvent être des noms communs, des noms propres, des chiffres, des dates, des abréviations ou encore des caractères uniques. Il en ressort qu'une bonne partie ne pourrait être ajoutée à un lexique bilingue d'entités nommées sans une vérification minutieuse. En revanche, les expressions à mots multiples semblent beaucoup plus pertinentes et ont fait l'objet de comptages spécifiques dans le tableau ci-dessous.

[9] Ce corpus, appelé *News Commentary*, a été mis à disposition de la communauté scientifique dans le cadre des *Workshops on Machine Translation*, par exemple sous http://www.statmt.org/wmt11.

Corpus	Mots dans le corpus	Expres-sions	Expressions différentes repérées	Mots dans les expressions repérées	Nb moyen de mots par expression	% mots / corpus
News Anglais	2 521 334	Toutes	5926	519897	1.06	20.6%
		Poly	4306	59373	2.21	**2.4%**
News Français	2 897 193	Toutes	4757	388785	1.03	13.4%
		Poly	1634	41270	4.77	**1.4%**

Tableau 2 : *Nombre d'expressions uniques repérées, nombres de mots concernés, longueur moyenne des expressions et couverture par rapport aux corpus en anglais et en français, en considérant toutes les expressions puis uniquement les expressions à mots multiples (« Poly »).*

La principale conclusion est que 2.4% des mots du corpus anglophone concernent une entité nommée mentionnée dans un titre d'article de Wikipédia (1.4% pour le français). Ces proportions peuvent paraître insignifiantes, mais elles représentent tout de même un mot sur 25 (1 mot sur 70 pour le français).

Un deuxième corpus de plus grande envergure a également été parcouru. Il s'agit des transcriptions en anglais du Parlement Européen (Europarl, Koehn 2005). Nous avons donc également confronté nos entités nommées à ce corpus de 45 millions de mots en anglais. Il est remarquable de constater que, malgré sa grande taille par rapport au corpus news, moins du double d'expressions différentes ont été repérées (11945 contre 4306). La couverture est similaire (1.8% soit 1 mot sur 55) à celle de la sélection faite sur le premier corpus (2.4%).

Corpus	Mots dans le corpus	Express.	expressions différentes repérées	mots dans les expressions repérées	nb moyen de mots par express.	% mots / corpus
Europarl Anglais	45 682 992	Toutes	11945	7 963 307	1.06	17.4 %
		Poly	7965	850 403	2.32	1.8 %

Tableau 3 : *Comptages similaires au tableau précédent pour le corpus Europarl*

6. Enrichissement de traits

Nous préparons les paires bilingues d'entités nommées en vue de leur utilisation dans un système de traduction automatique basé sur des règles linguistiques. Dans ce cas, les entités doivent être annotées avec un certain nombre de traits morphosyntaxiques. En particulier, le système de traduction doit savoir si l'entité nommée en question s'utilise avec ou sans déterminant (*les Etats-Unis* vs. *Jacques Delors*). Si un déterminant est requis, il faut indiquer son nombre (singulier ou

pluriel) et – pour le français – son genre[10]. Nous avons développé quatre heuristiques pour obtenir ces traits.

La grammaire des dates: Un nombre important de pages de Wikipédia concernent des dates de type *20 novembre*, *Mai 2004*. Ces dates suivent un schéma relativement rigide qui peut être détecté à l'aide d'un nombre restreint d'expressions régulières.[11] Par exemple, en français, les dates commençant par un jour sont précédées du déterminant masculin *le*, tandis que les autres dates s'utilisent sans déterminant.

L'entité nommée elle-même: La page intitulée *Le Caire* ne laisse pas de doute par rapport aux traits morphosyntaxiques: *Caire* prend un déterminant obligatoire au masculin singulier. Ce trait grammatical est évidemment indispensable pour privilégier des traductions comme *au Caire* plutôt que *à Le Caire*.

Le corps de la page: Nous comptons le nombre d'occurrences de l'entité nommée précédée d'un article et le nombre d'occurrences sans article. Par exemple, l'entité *Salvador Allende* apparaît 23 fois dans la page en question, mais jamais avec un déterminant. Par contraste, l'entité nommée *Royaume-Uni* apparaît 45 fois sur 47 avec un déterminant. Pour que ces comptages soient fiables, il faut un certain nombre d'occurrences. Il est grammatical d'utiliser *Royaume-Uni* sans article dans certains contextes, tout comme on peut trouver des séquences comme *le New York Police Department*, laissant croire que *New York* s'utilise avec un déterminant masculin. Des tests préliminaires ont montré que les occurrences avec articles étaient plus fiables que les occurrences sans article. Ainsi, nous pondérons les premières trois fois plus.

L'analyseur syntaxique Fips: Dans les cas où le corps de la page nous donne seulement des informations imprécises (la présence des articles *l'* et *les* ne donne pas d'indication quant au genre de l'entité nommée) ou que les comptages sont trop faibles pour être fiables, nous utilisons l'analyseur syntaxique Fips (Wehrli 2007) pour trouver les traits. Dans une expression à mots multiples, l'idée est de trouver sa tête grammaticale et de vérifier si celle-ci est un nom commun. Par

[10] On ne se préoccupe pas ici de savoir si les noms de personne, utilisés sans déterminant, se réfèrent à des hommes ou des femmes. Des heuristiques simples pourraient être employées pour déterminer cela, comme par exemple la présence de *né* ou *née* dans le texte, ou la proportion des pronoms *il* et *elle*.

[11] Les expressions calendaires, étant une classe fermée et composée de mots communs (et de chiffres), ne sont pas indispensables à un analyseur syntaxique, et peuvent même nuire à la qualité de ses analyses. Selon les cas, il peut donc être plus judicieux de ne pas entrer ces expressions en tant qu'entités nommées figées. Dans ce travail, nous gardons les dates, tout en les annotant comme telles pour faciliter un filtrage ultérieur.

exemple, l'expression *Office fédéral de la statistique* a comme tête *Office*, mot qui est contenu dans le lexique de Fips avec les traits masculin singulier. Cette méthode est utilisée pour pratiquement toutes les expressions anglaises, puisque le déterminant *the* ne donne pas d'indication sur le nombre (singulier ou pluriel).

Les entités nommées anglaises et françaises ont été annotées séparément à l'aide des heuristiques détaillées ci-dessus. Ensuite, elles sont remises ensemble dans une même entrée lexicale bilingue, avec leurs annotations respectives. Les entités dont l'annotation est restée incomplète sont écartées.

Rappelons que 655 071 paires d'entités nommées ont été extraites de Wikipédia (section 4). Parmi ces paires, 11 945 entités à un ou plusieurs mots ont été retrouvées dans le corpus de référence Europarl, et nous en avons retenu les 7965 entités à plusieurs mots. (section 5). Ces entités filtrées sont annotées à l'aide de nos heuristiques; les tableaux ci-dessous montrent les statistiques exactes.

Français:

7965 (100%)	Entités dans le fichier source
188 (2%)	Entités annotées par la grammaire des dates
138 (2%)	Entités annotées en utilisant l'entité nommée elle-même
6028 (76%)	Entités complètement annotées en utilisant le contenu de la page
1064 (13%)	Entités annotées en utilisant Fips
7418 (93%)	Entités complètement annotées en utilisant les quatre heuristiques

Anglais:

7965 (100%)	Entités dans le fichier source
187 (2%)	Entités annotées par la grammaire des dates
0 (0%)	Entités annotées en utilisant l'entité nommée elle-même
4122 (52%)	Entités complètement annotées en utilisant le contenu de la page
3158 (40%)	Entités annotées en utilisant Fips
7467 (94%)	Entités complètement annotées en utilisant les quatre heuristiques

Tableau 4 : *Nombre et pourcentage d'entités nommées annotées pour chaque heuristique en français et en anglais.*

On voit qu'en français, un pourcentage élevé d'entités peuvent être annotées en genre et en nombre en regardant directement les occurrences de l'entité dans le contenu de la page Wikipédia. En anglais, seules les entités sans déterminant peuvent être annotées complètement par le contenu de l'article; le nombre (singulier ou pluriel) est toujours détecté uniquement par Fips.

Dans l'annotation par Wikipédia, on trouve beaucoup de cas où le nombre d'occurrences est insuffisant pour prendre une décision fiable d'annotation (640 cas en français, 361 cas en anglais). En français, les entités introduites par les déterminants *l'* ou *les* sont sous-spécifiées (913 cas).

De même, l'annotation avec Fips peut échouer, soit parce que Fips ne connaît pas la tête de l'entité (228 en français, 474 en anglais), soit parce que l'entité contient à la fois des mots masculins et féminins (267 en français), soit parce que l'analyse de l'entité fournie par Fips est mauvaise (si la tête d'une entité est analysée faussement comme un verbe, il est impossible d'en récupérer le genre) (12 en français, 24 en anglais).

Une fois les entités anglaises et françaises annotées, on les remet ensemble pour former des paires bilingues d'entités nommées. Le tableau suivant résume les chiffres correspondants:[12]

7965	Paires d'entités dans le fichier source
918	Paires d'entités où au moins une langue a une annotation incomplète
40	Paires d'entités où la version française réfère à une liste
2	Autres erreurs d'alignement
7005 (88%)	Paires d'entités complètement annotées

Tableau 5 : *Nombre de paires d'entités nommées anglais-français entièrement annotées*

7. Conclusion

Nous avons proposé une procédure d'extraction et de validation de paires bilingues d'entités nommées à partir de l'encyclopédie Wikipédia. L'objectif de ce travail était de fournir un dictionnaire d'entités nommées afin d'enrichir un dictionnaire de la langue courante existant. Ces informations sont à intégrer dans un système de traduction automatique à base de règles. Dans cette optique, nous avons choisi un type d'annotation syntaxique plutôt que sémantique. Si de nombreux travaux concernent l'annotation sémantique en catégories (personne, pays, date, organisation), nous avons procédé à une annotation syntaxique (utilisation ou non d'un déterminant, genre, nombre) directement utilisable pour la traduction automatique, en analyse et en génération. Cependant, on pourrait étendre nos heuristiques relativement facilement pour obtenir des annotations en catégories sémantiques.

Dans une prochaine étape, il s'agira de tester dans quelle mesure ce dictionnaire d'entités nommées améliore les capacités d'analyse syntaxique et de traduction automatique. Des tests préliminaires n'ont pas donné les résultats attendus, principalement à cause de difficultés techniques d'intégration. Il faudrait aussi voir si le filtrage par corpus

[12] Quarante entités nommées françaises se référaient à des listes et avaient échappé au filtre décrit dans la section 4 (voir aussi note 4). Nous indiquons ces cas séparément dans le tableau.

est trop agressif, dans le sens où il garderait uniquement les entités nommées fréquentes, qui sont déjà présentes dans le lexique.

Le travail présenté ici a été appliqué à la paire de langue anglais-français. Toutefois, il peut être adapté facilement à d'autres paires de langues avec une bonne couverture sur Wikipédia; la seule adaptation concerne les heuristiques d'annotation. En particulier, les paires de langues contenant l'allemand, l'italien ou l'espagnol seraient des candidats intéressants pour des extractions d'entités nommées.

Bibliographie

BALASURIYA D., RINGLAND N. NOTHMAN J., MURPHY T. & CURRAN J.R. (2009). « Named entity recognition in Wikipedia », in *Proceedings of the 2009 Workshop on The People's Web Meets NLP*, Morristown, NJ, 10–18.

CHINCHOR, N.A. (1998), Overview of MUC-7/MET-2, http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html

CUCERZAN, S. (2007), « Large-scale named entity disambiguation based on Wikipedia data », in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, Prague, 708–716.

ERDMANN M., NAKAYAMA K., HARA T. & NISHIO S. (2008), « Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia », in *IPSJ Journal of Information Processing*, http://www.jstage.jst.go.jp/article/imt/3/3/564/_pdf

JIJKOUN, V. & DE RIJKE M. (2006), « Overview of the WiQA Task at CLEF 2006 », in *Evaluation of Multilingual and Multi-modal Information Retrieval - Revised Selected Papers of the 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, 265-274.

KAZAMA J. & TORISAWA K. (2007), « Exploiting Wikipedia as external knowledge for named entity recognition », in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 698–707.

KOEHN P. (2005), « Europarl: A parallel corpus for Machine Translation », in *Proceedings of the MT Summit 2005*, Phuket, 79-86.

MEDELYAN O., MILNE D., LEGG C. & WITTEN I.H. (2009), « Mining meaning from Wikipedia », in *International Journal of Human-Computer Studies*, 67 (9), 716-754.

MIHALCEA, R. (2007), « Using Wikipedia for automatic word sense disambiguation », in *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 196–203.

NERIMA L., SERETAN V. & WEHRLI E. (2006), « Le problème des collocations en TAL », in *Nouveaux cahiers de linguistique française*, 27, 95–115.

NOTHMAN J., CURRAN J. R. & MURPHY T. (2008), « Transforming Wikipedia into Named Entity Training Data », in *Proceedings of the Australian Language Technology Workshop*, Hobart, 124–132.

NOTHMAN J., MURPHY T. & CURRAN J. (2009), « Analysing Wikipedia and gold-standard corpora for NER training », in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athènes, 612–620.

PONZETTO, S.P. & STRUBE M. (2007), « Deriving a large scale taxonomy from Wikipedia », in *Proceedings of the 22nd National Conference on Artificial Intelligence*, Vancouver, B.C., 1440–1447.

RICHMAN A.E. & SCHONE P. (2008), « Mining wiki resources for multilingual named entity recognition », in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus OH, 1–9.

STRUBE M. & PONZETTO S.P. (2006), « WikiRelate! Computing semantic relatedness using Wikipedia », in *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, 1419–1424.

WEHRLI E. (2007), « Fips, a Deep Linguistic Multilingual Parser », in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Parsing*, Prague, 120-127.

WEHRLI E., NERIMA L. & SCHERRER Y. (2009), « Deep linguistic multilingual translation and bilingual dictionaries », in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athènes, 90–94.

WENTLAND W., KNOPP J., SILBERER C. & HARTUNG M. (2008), « Building a Multilingual Lexical Resource for Named Entity Recognition and Translation and Transliteration », in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, 3230-3237.

YU K. & TSUJII J. (2009), « Bilingual dictionary extraction from Wikipedia », in *Proceedings of Machine Translation Summit XII*, Ottawa, 379-386.