BiTermEx : un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle

Emmanuel Planas[1,2]
(1) UNAM, LINA, 2 rue de la Houssinière, BP 92208, 44322 Nantes
(2) UNAM, UCO, ST, 3, place André Leroy, 49008 Angers
emmanuel.planas@univ-nantes.fr

RÉSUMÉ

Nous décrivons BiTermEx, un prototype d'expérimentation de l'extraction de terminologie bilingue de mots composés, à partir de documents comparables, via la méthode compositionnelle. Nous expliquons la variation morphologique et la combinaison des constituants lexicaux des termes composés. Cette permet une précision TOP1 de 92% et 97,5% en français anglais, et de 94% en français japonais pour l'alignement de termes composés (textes scientifiques et de vulgarisation scientifique).

ABSTRACT

BiTermEx , A prototype for the extraction of multiword terms from comparable documents through the compositional approach.

We describe BiTermEx, a prototype for extracting multiword terms from comparable corpora, using the compositional method. We focus on morphology-based variations of multiword constituents and their recombinaison. We experimented our approach on scientific and popular science corpora. We record TOP1 precisions of 92% and 97,5% on French to English alignments and 94% on French to Japanese.

Mots-clés : extraction terminologique, prototype, terminologie bilingue, documents comparables, méthode compositionnelle, mots composés, corpus.

Keywords : term extraction, prototype, bilingual terminology, comparable documents, compositional method, multiword terms, corpus.

1 Introduction

Les documents comparables sont caractérisés par le partage d'un ensemble significatif de termes traduits en commun, tels les articles de Wikipédia relatifs à un sujet. (Déjean & Gaussier, 2011) en donnent cette définition : « Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1. » Ils sont plus difficiles à utiliser pour un alignement de termes que les documents parallèles qui sont, eux, des traductions l'un de l'autre (ex : les manuels de téléphones portables en plusieurs langues). Ceci provient du fait que les premiers ne présentent pas de repères positionnels et distributionnels présents dans les derniers. Ils ont cependant un avantage important : leur nombre nettement plus élevé (Fung, 1998).

En outre, Internet en est un réservoir important, qui s'incrémente quotidiennement.

2 Principes d'extraction de terminologie de corpus comparables

Les méthodes d'alignement de termes d'une langue à l'autre sont habituellement classées en deux grandes catégories : la **méthode contextuelle** (Fung, 1998) et la **méthode compositionnelle** (Robitaille et al., 2006). Elles reposent sur quatre phases.

La **Phase 1** de collecte les corpus sources et cibles. Elle peut être automatique, semi-automatique ou manuelle, éventuellement guidée par une liste fermée de termes sources et de leur traduction (graines), comme par exemple dans (Robitaille et al., 2006). Des outils ont été précédemment développés pour cette tache d'extraction, comme BootCat (Baroni & Bernardini, 2004.), ou Babouk (De Groc, 2011).

La **Phase 2** a pour but d'extraire une liste de termes sources et cibles candidats, souvent après une phase de pré-traitement du texte extrait des corpus ; léger : un effacement de mots vides (Fung, 1995), ou plus profond : analyse syntaxique chez (Yu & Tsujii, 2009). L'identification de termes candidats peut se faire : de façon monolingue, par exemple à l'aide de patrons de catégories grammaticales comme dans (Takeuchi et al., 2009) ; qui utilisent Acabit (Daille, 2003) ; ou encore de façon bilingue, comme dans (Fung, 1998), via la recherche d'une corrélation entre les termes source et cible.

La différentiation entre les deux grandes méthodes (**contextuelle et compositionnelle**) s'effectue en **Phase 3**. Celle-ci a pour mission de construire une liste de candidats cibles associés à un candidat source dont on cherche la traduction.

Dans la **méthode contextuelle**, cette construction se fait par le rapprochement statistique de **contextes** construits autour des mots sources d'une part et des mots cibles d'autre part. Les « contextes » peuvent être des vecteurs lexicaux (Fung, 1998) ou syntaxiques (Yu & Tsujii, 2009). La méthode contextuelle s'applique mieux aux termes simples car ceux-ci ont en général une fréquence plus importante que les termes complexes, qui peut être exploitée par les méthodes statistiques.

Dans la **méthode compositionnelle**, les termes sources dont on cherche la traduction sont «transposés» en des candidats cibles par construction morphologique, lexicale, syntaxique, ou sémantique à partir de leurs composants lexicaux (Morin & Daille, 2009). L'ensemble des combinaisons des composants est généré et traduit pour obtenir des candidats cibles.

C'est à la **Phase 4** que sont sélectionnées la où les meilleures traductions. En contextuel, cela s'effectue souvent à l'aide d'une mesure de similarité entre les contextes source et cible ; dans la méthode compositionnelle, les candidats résultent d'une « construction théorique ». La sélection des meilleurs candidats peut alors se faire par simple « projection » : on ne retient les candidats qui apparaissent dans la liste des termes cibles.

Pour compléter ce tour d'horizon, on pourra consulter (Laroche & Langlais, 2010) qui passent en revue l'ensemble des facteurs de la méthode contextuelle, et (Morin & Daille, 2009) qui présentent une large vue de la méthode compositionnelle.

3 Description de BiTermEx

BiTermEx est un prototype permettant de tester la méthode d'alignement de terminologie compositionnelle (Morin & Daille, 2009). Il est écrit en Java. Il a été testé sur Linux Ubuntu, Windows XP et 7. Nous décrivons ici les choix théoriques.

3.1 Phase 1 : Identification de corpus comparable

La collecte des corpus se fait actuellement manuellement : l'automatisation sera traitée dans une version ultérieure de l'outil.

3.2 Phase 2 : Extraction de termes candidats sources et cibles

BiTermEx extrait des listes de mots composés candidats via l'application de patrons de catégories grammaticales et de lemmes. Ceci est réalisé après une catégorisation et identification des lemmes relatifs aux mots du corpus. Cette analyse est obtenue par l'intégration du TreeTagger d'Helmut Schmid (Schmid, 1994). En post traitement de l'analyse de Tree Tagger, nous effectuons une standardisation des étiquettes des catégories grammaticales, de telle façon à pouvoir exprimer les patrons d'extraction de terminologie dans un langage commun à l'ensemble des langues traitées (Ex : français : DET:ART → DET@art ; anglais: RB → DET).

Chaque phrase du texte est exprimée comme une chaîne de caractères résultant de la concaténation de l'analyse de chacun de ses mots (la catégorie grammaticale et le lemme), en voici un exemple ; les mots sont séparés par des tirets bas :

…._cat=VERB@be3sp:lem=be_cat=DET:lem=a_cat=NOUN@sing:lem=design_cat=NOUN@sing:lem=concept_cat=PREP@in:lem=for_cat=DET:lem=a_…

Les règles d'extraction de terminologie sont du type suivant :

[cat1=DET_cat2=NOUN_cat3=NOUN_lem4=for#lem2_lem3]

Dans cet exemple, la correspondance entre le texte analysé et la règle d'extraction se fait successivement sur le *DET/a*, *NOUN/design*, *NOUN/concept*, et *lem/for*. Le terme extrait est *design concept*.

Cette méthode permet non seulement d'identifier des patrons, mais aussi de les contextualiser (*design concept* est extrait entre *DET/a* et *lem/for*). Nous avons ici une contribution à la question de la « termicité » (« termhood » en anglais : la séquence extraite est-elle vraiment un « vrai » terme ?) des termes candidats extraits (Robitaille et al., 2006). De plus, cette méthode est facilement adaptable : les règles peuvent être modifiées ou ajoutées par simple édition d'un fichier externe.

3.3 Phase 3 : Génération d'une liste de candidats cibles pour chacun des termes composés sources

3.3.1 Modifications de l'approche classique

i) Modification des unités lexicales par variation morphologique

Supposons que l'un des termes sources extraits en Phase 2 soit *production annuelle*, lemmatisé en *production annuel,* et que dans le dictionnaire bilingue, *production* soit traduit par *production* et *output*, et *annuel* soit traduit par *yearly* et *annual*. Alors l'approche compositionnelle consiste en phase 3 à combiner les traductions : production *yearly, yearly production, production annual, annual production, output yearly, yearly output, output annual, annual output*. Et la phase 4 à sélectionner les termes qui apparaissent dans la liste de termes composés cibles extraite en Phase 2 : *yearly production* et *annual production*. Le seul fait qu'ils apparaissent dans le texte source étant discriminant.

Mais pour le terme *biologie tumorale* qui est traduit en anglais par *tumor biology*, le substantif *tumor* n'est pas la traduction de l'adjectif *tumorale*, mais du substantif dérivé de la même famille : *tumeur*.

Cette difficulté peut être résolue par variation graphique, morphologique, lexicale, ou syntaxique des éléments constituant les termes composés (Morin & Daille, 2009). Nous traitons les dérivations lexicales par l'application de règles de dérivation. Dans le cas de notre exemple, la lemmatisation ayant transformé *biologie tumorale* en *biologie tumoral*, la règle : *oral* → *eur* permet de transformer l'adjectif *tumoral* en le nom *tumeur* qui sera bien traduit par *tumor*. Dans le prototype, les règles sont consignées dans un fichier texte pour chaque langue. L'utilisateur peut ajouter ou modifier ces règles par l'édition de ces fichiers.

ii) Combinaison

Ordre de combinaison

Les unités lexicales des mots composés sources sont combinées pour chacune des variations générées précédemment. Pour une variation fixée, un mot composé ABC de longueur N, présente alors N! combinaisons de longueur N. Voici une illustration :

ABC → **C**AB, A**C**B, AB**C** ; **C**BA, B**C**A, BA**C** (N = 3, 3! = 6)

L'ensemble de ces permutations sont construites par récurrence sur N, illustrée ici par le positionnement de 'C' sur les différentes permutations de AB.

Profondeur de recherche

Dans la récurrence, nous gardons les sous-permutations de dimension N-p, et nommons p : « profondeur de recherche » de combinaison d'unités lexicales. Notons que pour une profondeur p, le nombre de combinaisons est $A^{p}_{N} = (N)\,!/(N\text{-}p)\,!$

Cela permet, après traduction, de générer des termes candidats de longueur différente. C'est une réponse possible au problème de fertilité (Morin & Daille, 2009).

3.3.2 Réduction du nombre de permutations

Chacune des combinaisons sources est traduite. Nous utilisons simplement un lexique bilingue pour effectuer le transfert de la langue source à la langue cible.

Pour réduire le nombre de combinaisons traduite, nous cherchons **avant traduction** l'ensemble des traductions de variantes d'unités lexicales trouvées en (ii) dans le dictionnaire. Les variantes qui ne possèdent pas de traduction, ou dont les traductions n'apparaissent pas parmi les composants des termes cibles extraits en Phase 2 sont éliminées avant combinaison, réduisant de façon importante cette complexité.

3.4 Phase 4 : Projection des candidats cibles sur la liste des Termes Composés Cibles extraits.

La sélection des meilleurs candidats cibles construits en phase 3 se fait par « projection ».

4 Expériences et Résultats

4.1 Expérience 1 : Wikipédia, français anglais, 935 termes candidats

Corpus

Nous avons téléchargé manuellement 13 articles Wikipédia français (source) et 15 anglais (cible) liés au thème de l'énergie éolienne. Après l'application d'un filtre retirant les métadonnées, les balises, et les parties textuelles liés à la navigation de Wikipédia, nous avons obtenu 36077 tokens français et 39761 tokens anglais. L'analyse de TreeTagger nous a permis d'identifier les lemmes et catégories grammaticales.

Extraction de termes monolingues candidats

L'extraction monolingue de candidats sources français donne 846 termes composés français et 935 termes composés anglais de longueur comprise entre 2 et 5 unités lexicales, comme *autonomie énergétique* en français, ou encore *vertical axis wind turbine* en anglais. En voici quelques caractéristiques de description statistique :

	Termes initiaux	**Freq min.**	**Freq Max.**	**Freq. Moy.**	**Freq. = 1**	**Freq. = 2**	**Freq 3-5**	**Freq 6-10**	**Freq > 10**
FR	846	1	62	1,5	703 (83%)	94 (11%)	34 (4%)	9 (1%)	6 (0,7%)
EN	935	1	56	1,4	772 (83%)	121 (13%)	35 (4%)	0 (0%)	7 (0,7%)

TABLE 1 – Répartition statistique des mots composés monolingues candidats – Wikipedia - Éoliennes – FR - EN

Alignement de Termes

Nous utilisons le dictionnaire ELRA français anglais contenant 103.190 entrées françaises correspondant à 238.742 traductions. L'alignement compositionnel entre les 846 termes sources français et les 935 termes cibles anglais est réalisé aux profondeurs 0 (longueur N), 1 (longueur >=N-1, et 2. (>=N-2). L'alignement des termes a été évalué

manuellement. Une erreur est de type A si la traduction est partielle, de type B si la traduction est complètement fausse. Les résultats sont rassemblés ici :

Profondeur	Nb alignés	Nb Erreurs A	Nb Erreurs B	Rappel	Précision	Score F
0	79	1 (1,2%)	5 (6,3%)	9,3%	92,4% (74)	5,9
1	94	2 (2,5%)	5 (6,3%)	11%	91,1% (73)	5,1
2	144	49 (34%)	12 (8,3%)	17%	57,6% (83)	3,8

TABLE 2 – Statistiques d'alignement bilingue suivant la profondeur de recombinaison – Wikipedia - Éoliennes – FR - EN

On enregistre un taux de précision TOP1 de 92,4% en profondeur 0 et de 91,1 en profondeur 1, pour une légère amélioration du rappel. Le traitement de la fertilité est amélioré par la prise en compte de traductions anglaises possédant les mêmes lexèmes que le français, sans préposition. Ex : *production de électricité | electricity production*. Ce taux chute pour la profondeur 2. Voici la répartition des termes alignés suivant leur nombre d'occurrence dans le corpus, pour la profondeur 1 :

	Termes alignés	Freq min.	Freq Max.	Freq. Moy.	Freq. = 1	Freq. = 2	Freq 3-5	Freq 6-10	Freq > 10
FR	94	1	62	4,4	53 (56%)	12 (13%)	12 (13%)	7 (7%)	10 (11%)
EN	94	1	23	2,5	63 (67%)	13 (13%)	11 (12%)	0 (0%)	7 (7%)

TABLE 3 – Répartition des mots composés alignés – Wikipedia - Éoliennes – FR - EN

Plus de la moitié de l'effectif sont des hapax, et deux tiers sont de fréquence inférieure ou égale à 2. Cela montre que la méthode compositionnelle est très adaptée à l'alignement de termes de très basse fréquence. On note cependant, en comparant ce tableau au tableau de répartition de l'ensemble de l'effectif, que les mots composés de plus hautes fréquences ont proportionnellement une tendance à mieux s'aligner que ceux de basse fréquence : le quartile des mots composés alignés de fréquence supérieure à 10 est de 11% pour le français, alors que seulement 0,7% de l'ensemble des mots composés français extraits en Phase 2 ont une telle fréquence. Il est intéressant de noter que pour ce corpus, contrairement au corpus suivant, l'application de règles morphologiques ne fourni que peu de résultats : 3 termes 94.

4.2 Expérience 2 : Corpus médical, français anglais

Le module d'alignement de termes étant indépendant du module d'extraction de termes du corpus, nous avons pu tester l'alignement en profondeur 0 d'une liste de termes déjà extraite, issue d'un corpus médical spécialisé sur le cancer du sein de 3483 termes composés français avec une liste de 6642 termes composés anglais. Aussi, nous ne nous intéressons ici qu'à l'analyse de l'alignement de termes composés.

Alignés	classique	morpho	Erreurs A	Erreurs B	Rappel	Précision
808	702 (87%)	106 (13%)	2%	0,5%	24 %	97,5 %

TABLE 4 – Statistique d'alignement de mots composés – Cancer du sein – FR – EN

La précision est très forte, pour un rappel significatif. On note ici l'importance des variations morphologiques puisqu'elles engendrent 13% des solutions.

4.3 Expérience 3 : Corpus médical, français japonais

Nous avons testé l'alignement entre 23487 mots composés français et 26188 mots composés japonais extraits d'un corpus médical sur le diabète et l'obésité. Ces listes ne sont pas bien nettoyées : des dates et des extractions ne correspondant pas à des termes complets subsistent. Le petit dictionnaire généraliste utilisé est celui de Jean-Marc Desperrier[1] (18039 entrées françaises, 32444 traductions japonaises). Nous n'avons pas de données sur le degré de comparabilité de ce corpus. Les résultats confirment la bonne précision de la méthode, et l'importance des variations morphologiques.

Alignés	classique	morpho	Erreurs A	Erreurs B	Rappel	Précision
140	85 (61%)	55 (39%)	5%	3%	0,6%	92%

TABLE 5 – Statistique d'alignement de mots composés – Diabète - Obésité – FR – JP

5 Conclusions et perspectives

Cette étude montre la très bonne précision de la méthode compositionnelle, en particulier pour les hapax et les termes très peu fréquents. Nos résultats confirment ceux de (Morin & Daille, 2009) qui obtiennent des taux de précision de 88 % pour des termes composés de basse fréquence. Ainsi que ceux de (Robitaille et al., 2006) qui obtiennent des précisions comprises entre 49% et 92% sur une extraction français – japonais. Ces derniers mesurent un rappel élevé, calculé sur une population restreinte des termes français et japonais vérifiant dans chaque langue une cohérence forte avec une liste de graines bilingues (test de Jacquard $>=$ 0,01), alors que nous n'imposons pas de telle restriction. Les raisons générales du faible rappel ont été identifiées dans des travaux précédents : non-compositionnalité des termes composés et couverture des dictionnaires bilingues (Morin & Daille 2009).

Remerciements

Ce travail, qui s'inscrit dans le cadre du projet METRICC (www.metricc.com), a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009. Merci à Koichi Takeuchi pour sa coopération et à Emmanuel Morin et Béatrice

[1] http://dico.fj.free.fr/dico.php [15 janvier 2012]

Daille pour leurs conseils.

Références

BARONI, M., & BERNARDINI, S. BOOTCAT (2004). Bootstrapping corpora and terms from the web. Dans E. L. R. A. Elra (Éd.), *Proceedings of LREC* (Vol. 2004, p. 1313–1316). ELRA.

DAILLE, B. (2003). Conceptual Structuring through Term Variations. *Proceeding MWE '03 Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*.

DE GROC, C. (2011). Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. *Proceedings of the IEEEWICACM International Conferences on Web Intelligence*, 497–498.

DÉJEAN, H., & GAUSSIER, E. (2011). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables hal. *Lexicometra*.

FUNG, P. (1995). Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. *Proceedings of the Third Workshop on Very Large Corpora* (p. 173–183).

FUNG, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. *Parallel Text Processing* (p. 1–17). Springer.

LAROCHE, A., & LANGLAIS, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *COLING* (p. 617-625).

MORIN, E., & DAILLE, B. (2009). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*.

ROBITAILLE, X., SASAKI, Y., TONOIKE, M., SATO, S., & UTSURO, T. (2006). Compiling French-Japanese Terminologies from the Web. *EACL*.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, United Kingdom.

TAKEUCHI, K., KAGEURA, K., KOYAMA, T., DAILLE, B., & ROMARY, L. (2009). Pattern Based Term Extraction Using ACABIT System. *CoRR*.

YU, K., & TSUJII, J. (2009). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. *HLT-NAACL (Short Papers)* (p. 121-124).