Identification d'erreurs de traduction dans un dictionnaire de recherche d'informations translingue et traduction de mots composés à l'aide du World Wide Web

Hubert Naets, Gregory Grefenstette

CORIA 05
France - Grenoble - 9-11 mars 2005

Identification d'erreurs de traduction dans un dictionnaire de recherche d'informations translingue et traduction de mots composés à l'aide du World Wide Web

Hubert Naets — Gregory Grefenstette

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M)
Commissariat à l'Énergie Atomique
Bat. 38-1 ; 18, rue du Panorama ; BP 6
92265 Fontenay aux Roses Cedex ; France

naetsh@zoe.cea.fr
gregory.grefenstette@cea.fr

RÉSUMÉ. *La recherche d'informations translingue sur des textes non parallèles nécessite une phase de traduction entre une requête dans une langue source et un document dans une langue cible. Afin d'obtenir les mêmes performances que dans le cas d'une requête monolingue sur un document dans la même langue que cette requête, il est nécessaire de trouver les bonnes traductions pour tous les termes de la requête en langue source.*

Malheureusement, les dictionnaires de traduction disponibles ne contiennent pas les traductions exactes d'un grand nombre de mots composés qui peuvent être présents dans une requête. Les systèmes de recherche translingues utilisent des dictionnaires de traduction construits statistiquement ou manuellement. Afin de traduire un mot composé, beaucoup de ces systèmes génèrent toutes les traductions possibles mot à mot et vérifient la présence de ces traductions dans la base de donnée cible. La qualité de la recherche augmente lorsque il est possible d'utiliser des traductions de mots composés préalablement validées.

Il reste cependant deux problèmes encore non résolus avec cette méthode consistant à générer et à valider toutes les traductions : (1) Si la traduction exacte d'un élément d'un mot composé ne figure pas dans le dictionnaire de traduction, la traduction qui sera validée par cette méthode ne sera pas la meilleure traduction. (2) Si la bonne traduction ne comprend pas le même nombre d'éléments que le mot composé source, la meilleure traduction ne sera pas non plus générée.

Dans cet article, nous proposons deux méthodes pour identifier ces situations.

ABSTRACT. *Cross-language information retrieval over non parallel text requires a translation phase between a source language query and a target language document. In order to achieve the same performance as a monolingual target language query, good translations for all terms*

in a source language query must be found.
Unfortunately, available translation dictionaries do not contain exact translations for many multiword terms that can be found in a query. Cross language retrieval systems use statistically or manually built translation dictionaries to perform translation, and in order to translate a multiword term, many systems generate possible word-to-word translations and verify the existence of the translations in the target database. When validated translations of multiword structures are used, retrieval improves.
But there are two unsolved problems with the generate-and-validate method: (1) if the proper translation for one word in the multiword term is not in the translation dictionary the translation that will be validated by the method will not be the best translation, and (2) if the multiword term in the source is not translated by the same number of nonstop words in the target language, then the best translation will not be generated.
In this paper, we present two methods for recognizing when these situations arise.

MOTS-CLÉS : *Recherche d'information translingue, traduction de requêtes, dictionnaires bilingues, mots composés.*

KEYWORDS: *Cross-language information retrieval, query translation, bilingual dictionaries, compound words, multiword terms.*

1. Introduction

Un problème récurrent et toujours non résolu de la reformulation multilingue est celui des mots composés ne se traduisant pas par le même nombre de mots dans une langue cible que dans une langue source. Ainsi une personne utilisant un moteur de recherche translingue devrait trouver des documents anglais dans lesquels figure par exemple le mot « plunge » s'il lance comme requête « chute brutale des cours du pétrole ». Il devrait aussi trouver des documents en anglais parlant de « blacklist » en réponse à une requête sur « liste noire ». De la même façon, « vie privée » devrait correspondre à « privacy » et « clair de lune » à « moonlight ». Toutefois, beaucoup de systèmes de recherche d'information translingue utilisent une technique consistant à combiner toutes les traductions de chaque élément d'un mot composé afin de constituer un ensemble de traductions candidates pour ce mot composé [QU 02], méthode qui fonctionne d'ailleurs relativement bien dans les cas de traductions compositionnelles [GRE 99]. Ces systèmes utilisent pour ce faire des dictionnaires électroniques multilingues réalisés manuellement ou dérivés de corpus bilingues alignés. C'est la raison pour laquelle un « petit déjeuner » (« breakfast ») peut devenir, au cours d'une reformulation en anglais, entre autres traductions, un « light lunch » et la « pleine nuit » (« middle of the night ») se transformer en une « full night ».

Cette technique de combinaison des traductions des parties d'un mot composé pose également le problème de la validation ou de l'invalidation des traductions des mots composés. En combinant toutes les traductions possibles pour chaque partie d'un mot composé, un moteur de recherche translingue est en effet amené à générer des requêtes cibles qui n'intéressent pas l'utilisateur ou qui ramènent des informations non désirées. Ainsi, une personne émettant une requête à propos de « recettes fiscales » préférera trouver des « fiscal revenues » plutôt que des « fiscal recipes ». S'il est bien connu que la base de données à laquelle la requête est appliquée contribuera pour une grande part à désambiguïser la requête dans la mesure où — si l'on poursuit sur notre exemple —, il semble moins probable de trouver « fiscal » dans une recette de cuisine et « recipes » dans un document comptable, une telle chose n'est néanmoins pas à exclure. Par ailleurs, la multiplication de requêtes non motivées (comme « fiscal recipes » par exemple) contribue à ralentir inutilement les recherches de documents dans une base de données.

Le problème dee la traduction de mots composés dans le domaine de la recherche d'information a déjà été traité, entre autres, au moyen d'approches basées sur des exemples à l'aide de patrons bilingues [KAT 94], au sein de systèmes de traduction automatique basée sur des grammaires d'unification [BOU 92] ou encore à l'aide de modèles de traduction statistiques ([FUJ 99b] et [FUJ 01]).Nous en proposons ici une approche assez simple utilisant le World Wide Web comme un immense corpus d'exemples attestés.

À défaut de pouvoir systématiquement déterminer la traduction correcte d'un mot composé dans les cas qui nous occupent — qu'il s'agisse d'une reformulation multilingue ne comportant pas le même nombre d'éléments que l'unité linguistique source,

ou que la traduction correcte d'un des éléments ne figure pas dans le dictionnaire de reformulation multilingue —, nous traitons ici ici de deux méthodes permettant d'identifier ces situations.

Le but des expérimentations présentées ci-après est donc de savoir si une traduction correcte existe parmi l'ensemble des traductions candidates pour un mot composé et, si elle existe, quelle est-elle.

Dans la suite, nous présentons successivement ces deux méthodes et le résultat de leur combinaison, avant d'en discuter les avantages et les inconvénients.

2. La méthode des proportions

2.1. *L'hypothèse de départ*

L'hypothèse de départ est la suivante : il est possible de valider ou d'invalider un certain nombre de traductions candidates de mots composés en prenant en compte le rapport existant entre la fréquence d'un mot composé source et la fréquence de chacune de ses traductions candidates sur le Web. Intuitivement, on s'attend à ce que les traductions correctes aient, vis-à-vis de leur mot composé source, le même rapport que le nombre total de pages web de la langue cible vis-à-vis du nombre total de pages web de la langue source. En d'autres termes, le rapport entre le nombre de pages en français et en anglais devrait être le même qu'entre un mot composé en français et sa traduction correcte en anglais.

En mars 2004, le nombre de mots en français présents sur le Web était de 13 648 627 000 contre 145 959 354 000 en anglais, ce qui correspond à un ratio de 10,69[1] (en février 2000, ce ratio était de 21,4 [GRE 00]). Il est ainsi possible d'estimer qu'un moteur de recherche devrait trouver approximativement, lors d'une requête, 10 à 20 fois moins de pages pour un mot composé en français que pour sa traduction correcte en anglais, sauf en ce qui concerne certains faits de société, certaines réalités culturelles ou géographiques propres aux mondes francophone ou anglophone. À titre d'exemple, dans Google, la requête *déchets radioactifs* renvoie 75 000 pages, tandis que sa traduction, *radioactive waste*, en renvoie 1 060 000, ce qui correspond à un ratio anglais – français de 14,1.

2.2. *L'expérimentation*

Nous avons testé cette hypothèse sur des mots composés en français et sur leurs traductions candidates en anglais.

1. http ://www.infonortics.com/searchengines/sh04/slides/greffen.pdf

2.2.1. *La sélection des mots composés en français*

Nous avons pris pour point de départ un corpus d'articles provenant du journal français *Le Monde* de 1995 et comprenant un peu plus de 2 800 000 tokens. À l'aide du système *LIMA* (*LIC2M Multilingual Analyzer*) développé au *LIC2M*, nous avons étiqueté et désambiguïsé morphosyntaxiquement le corpus et en avons extrait toutes les successions « nom – adjectif » et « adjectif – nom ».

EXEMPLES : — despote éclairé — dessein européen — dessin animé — deuxième autorité — premier accident

Nous avons ensuite calculé l'information mutuelle (*mutual information*) de chacune de ces séquences en nous basant sur leur fréquence au sein du corpus afin d'établir une liste de cooccurrences « nom – adjectif » et « adjectif – nom ». Nous n'avons conservé que les séquences apparaissant fréquemment dans le corpus d'articles (plus de 5 fois) et ayant une information mutuelle élevée (cf. figure 1).

```
immense majorité
indemnité journalière
industrie automobile
ingénierie financière
instances disciplinaires
interdiction complète
investisseurs internationaux
jeune femme
joueurs professionnels
large consensus
lettre mensuelle
libre circulation
liste noire
long chemin
lourdes pertes
maillot jaune
maîtrise comptable
marchés financiers
marine française
matière grise
mauvaise humeur
médecins libéraux
```

Figure 1. *Séquences « adjectif – nom » et « nom – adjectif »*

Si deux mots, $m1$ et $m2$ ont respectivement une probabilité $P(m1)$ et $P(m2)$ dans un corpus, leur information mutuelle $MI(m1, m2)$ peut être définie de la façon suivante [CHU 90] :

$$MI(m1, m2) = \log_2 \frac{P(m1, m2)}{P(m1)\,P(m2)}$$

L'utilisation de l'information mutuelle a en outre permis d'éliminer un certain nombre de séquences contenant des adjectifs numéraux ainsi que quelques séquences mal étiquetées morphosyntaxiquement.

La méthode employée n'est pas dépendante du type de mots composés ; dans le cadre de cette première expérimentation, nous nous sommes limités aux séquences « nom – adjectif » et « adjectif – nom » : des séquences « nom – préposition – nom » ou « verbe – préposition – nom » auraient également pu être prises en compte. De même que le recours à l'information mutuelle n'est pas indispensable ; il permet d'isoler plus facilement les mots composés propres à un ou plusieurs domaines du corpus ou de la base de données.

2.2.2. *La traduction en anglais des mots composés*

Un dictionnaire de reformulation bilingue français – anglais a ensuite été utilisé pour générer toutes les combinaisons possibles de traductions candidates pour chaque mot composé. Ce dictionnaire, réalisé au CEA/LIC2M, comporte 116 819 traductions de mots simples et de mots composés étiquetés morphosyntaxiquement.

Pour chaque séquence « nom – adjectif » ou « adjectif – nom », toutes les traductions du nom présentes dans le dictionnaire de reformulation ont été combinées avec toutes celles de l'adjectif (cf. figure 2).

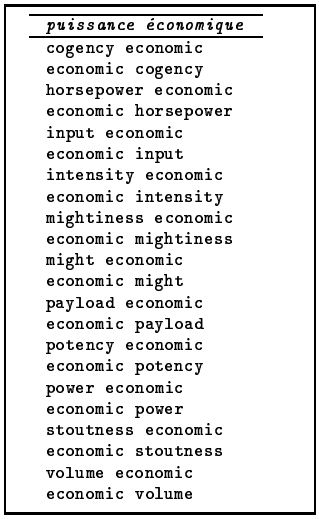

Figure 2. *Génération de toutes les traductions de « puissance économique » à l'aide du dictionnaire de reformulation*

Nous avons sélectionné aléatoirement 170 séquences sources (ainsi que toutes leurs traductions) parmi lesquelles nous avons éliminé manuellement une séquence mal étiquetée morpho-syntaxiquement en français (ce qui posait des problèmes de traduction), pour n'en garder que 169.

Le nombre moyen de traductions candidates pour chacun de ces 169 mots composés en français est de 39,9 , avec un minimum de 2 traductions et un maximum de 728.

La moitié de ces 169 mots composés en français (84 exactement) possède au moins une traduction correcte parmi l'ensemble des traductions générées, alors que l'autre moitié (85) n'en possède aucune.

2.2.3. *Les traductions de référence*

Afin d'établir les traductions servant de référence au cours de notre évaluation (que nous considérons au long de ce document comme étant les « traductions correctes »), nous avons demandé à un locuteur natif anglophone de produire ce qu'il lui semblait être la meilleure traduction pour chaque mot composé en français, sans que ce locuteur ait pris connaissance au préalable des traductions candidates générées à partir du dictionnaire de reformulation. Par la suite, nous lui avons demandé de repréciser la traduction de 12 mots composés qui semblaient admettre plusieurs traductions correctes (« faible marge » admet ainsi deux traductions correctes dans notre jeu de tests : « slight margin » et « low margin »).

2.2.4. *L'interrogation du moteur de recherche*

L'étape suivante a consisté à interroger le moteur de recherche Google avec chacun des mots composés sources et chacune des traductions candidates successivement et de récupérer, pour chaque requête, le nombre de pages dans lesquelles figure l'expression (source et cible) recherchée. Par exemple, « puissance économique » (figure 3) est présent dans 26 000 pages indexées par Google. Si l'on prend sa première traduction potentielle, « cogency economic », on constate qu'elle n'apparaît que dans une seule page et que le rapport entre cette seule page et le nombre de pages ramenées par « puissance économique » est de 0,000038. Au contraire, avec 300 000 pages, «economic power» a une fréquence 11,538462 fois supérieure à celle du mot composé source, ce qui le positionne comme meilleure traduction possible, compte tenu de notre hypothèse de départ (rapport de 10 à 20 entre l'anglais et le français).

2.2.5. *Révision de l'hypothèse de départ*

Il est très vite apparu nécessaire de réviser l'hypothèse de départ dans la mesure où elle s'avère trop restrictive : en effet, la majorité des traductions correctes ramènent plus ou moins de pages web que l'hypothèse ne le prédit. Ainsi, par exemple, « long way » renvoie 165 fois plus de pages que « long chemin » ; à l'opposé, « communist candidate » n'est que 1,45 fois plus prolixe que « candidat communiste », tandis qu'« official announcement » se contente de retourner 6,29 fois plus de pages qu'« annonce officielle ». Plus généralement, seules 18,75 % des traductions correctes sont comprises dans un intervalle allant de 10 à 20. Le tableau 1 montre que la plupart des traductions correctes se situent au-deça ou au-delà de cet intervalle.

L'expérimentation a montré que dans le cas de notre échantillon de test, il fallait prendre un intervalle allant de 3 à l'infini afin d'obtenir les meilleurs résultats, la proportion de mauvaises traductions augmentant nettement en-dessous de 3.

traduction	fréq google	ratio EN-FR
puissance économique	*26000*	
cogency economic	1	0.000038
economic cogency	2	0.000077
horsepower economic	12	0.000462
economic horsepower	102	0.003923
input economic	379	0.014577
economic input	6660	0.256154
intensity economic	274	0.010538
economic intensity	186	0.007154
mightiness economic	0	0.000000
economic mightiness	2	0.000077
might economic	697	0.026808
economic might	28500	1.096154
payload economic	9	0.000346
economic payload	28	0.001077
potency economic	7	0.000269
economic potency	310	0.011923
power economic	21600	0.830769
economic power	300000	11.538462
stoutness economic	0	0.000000
economic stoutness	0	0.000000
volume economic	759	0.029192
economic volume	975	0.037500

Figure 3. *Rapport entre la fréquence de puissance économique et celle de ses traductions potentielles*

ratio	pourcentage de traductions correctes	pourcentage de traductions incorrectes les plus probables
< 1	18,75 %	56,38 %
1 – 3	16,7 %	17,02 %
3 – 5	11,45 %	4,25 %
5 – 10	18,75 %	2,12 %
10 – 20	18,75 %	6,38 %
20 – 50	8,3 %	4,26 %
50 – ∞	7,3 %	9,57 %

Tableau 1. *Répartition du ratio anglais – français sur l'ensemble des résultats corrects et pour les résultats incorrects renvoyant le plus de pages pour une traduction donnée*

2.3. *Les résultats*

Le test a porté sur la capacité de cette première méthode à identifier d'une part les cas où aucune traduction correcte ne figurait parmi l'ensemble des traductions générées et à déterminer d'autre part quelle était la meilleure des traductions correctes si au moins une traduction correcte avait été produite. Compte tenu de la modification de l'hypothèse, la meilleure traduction a été redéfinie comme étant la traduction renvoyant le plus de pages, à condition que ce nombre de pages soit au moins trois fois

supérieur au nombre de pages renvoyées par le mot composé source. Le tableau 2 se lit de la façon suivante : « Si la méthode indique avoir trouvé une bonne traduction ou aucune traduction (verticalement), alors cette affirmation est correcte ou fausse (horizontalement) dans X % des cas ». La dernière ligne (« précision globale ») correspond à la performance générale de la méthode, que celle-ci atteste de la découverte d'une traduction correcte ou qu'elle déclare qu'il n'en existe aucune.

	correct	faux
la bonne traduction	69,1 %	30,9 %
aucune bonne traduction	76,4 %	23,6 %
précision globale	72,9 %	27,1 %

Tableau 2. *Précison de la méthode des proportions*

On constate ainsi que dans près des trois quarts des cas, la méthode des proportions permet ou bien de déterminer correctement la bonne traduction ou bien d'établir qu'il n'en existe aucune parmi celles qui sont proposées. Si l'on s'intéresse plus particulièrement au cas où cette méthode affirme ne pas trouver de bonne traduction (52,4 % des cas), on remarque qu'elle ne se trompe que dans 23,6 % des cas. Les performances sont un peu moins bonnes si l'on s'arrête sur les cas où la méthode dit avoir trouvé la bonne solution (47,6 % des cas) puisque cela n'est vrai qu'à 69,1 %.

La méthode prouve ainsi sa capacité à identifier un grand nombre de cas où aucune traduction correcte n'existe parmi l'ensemble des traductions d'un mot composé générées à l'aide d'un dictionnaire de reformulation. Toutes les traductions ne comprenant pas le même nombre d'éléments que dans le mot composé source (cf. figure 4) se retrouvent ainsi dans cette catégorie.

mot composé source	traduction
bête noire	thorn in the side
chute brutale	plunge
classement général	ranking
denrées alimentaires	foodstuffs
interdiction complète	ban
liste noire	blacklist
petit dêjeuner	breakfast
pleine nuit	middle of the night
vie privée	privacy

Figure 4. *Mots composés en français et traductions non terme à terme correspondantes*

3. La méthode de la coprésence

3.1. *L'hypothèse de départ*

Il existe sur le World Wide Web un certain nombre de textes multilingues dont une partie est la traduction ou le résumé de l'autre, ou encore qui traitent du même sujet (par exemple les forums de discussion). Il serait donc possible d'utiliser ces pages comme autant de bitextes ou de multitextes et de valider ou d'invalider ainsi les différentes traductions d'un mot composé.

Au lieu de calculer le rapport entre le nombre de pages renvoyées par un mot composé et par ses traductions, on crée une requête où figurent à la fois le mot composé source et un des mots composés candidats. À la requête « "déchets radioactifs" + "radioactive waste" », le moteur de recherche Google renvoie ainsi 5 100 pages.

On peut supposer que la traduction figurant dans la requête qui renvoie le plus grand nombre de pages est la meilleure traduction ; de même qu'on peut émettre l'hypothèse qu'une traduction liée à une requête ne renvoyant aucune page est fausse.

Cette méthode est fortement tributaire du nombre de pages bilingues ou multilingues présentes sur le web pour les deux langues considérées et est donc peu exploitable dans le cas de langues faiblement représentées sur internet.

3.2. *L'expérimentation*

Nous sommes partis du même corpus de 169 mots composés et de leurs traductions candidates que celui que nous avons utilisé dans le but de tester la méthode des proportions. Pour chaque traduction candidate, nous avons créé une requête « "mot composé" + "traduction potentielle" » dont nous nous sommes servi pour interroger Google. Ainsi, si l'on reprend l'exemple de «puissance économique» (figure 5), on constate que le moteur de recherche renvoie des pages dans deux cas seulement : « economic might » (5 pages) et « economic power » (94 pages), ce dernier constituant ainsi la meilleure traduction et correspondant à la traduction de référence.

3.3. *Les résultats*

La difficulté principale de cette seconde méthode tient au fait que, pour chaque requête, le moteur de recherche utilisé renvoie peu de pages — y compris pour les meilleures traductions — par rapport à la première méthode que nous avons proposée. Il s'avère ainsi impossible, dans un certain nombre de cas, de sélectionner la meilleure traduction, dans la mesure où une requête constituée d'une traduction moins bonne renvoie parfois une ou deux pages de plus que la meilleure traduction.

Ceci explique les assez mauvais résultats de la méthode de la coprésence lorsqu'il s'agit de déterminer la meilleure traduction (cf. tableau 3) : le taux d'erreur est en effet

traduction	fréq google
puissance économique	
cogency economic	0
economic cogency	0
horsepower economic	0
economic horsepower	0
input economic	0
economic input	0
intensity economic	0
economic intensity	0
mightiness economic	0
economic mightiness	0
might economic	0
economic might	5
payload economic	0
economic payload	0
potency economic	0
economic potency	0
power economic	0
economic power	94
stoutness economic	0
economic stoutness	0
volume economic	0
economic volume	0

Figure 5. *Nombre de pages renvoyées par Google lors d'une requête « "puissance économique" + "traduction candidate" »*

de 43,2 %. Les résultats sont meilleurs si l'on ne considère pas la traduction ramenant le plus de pages mais les n traductions ramenant au moins une page. La meilleure traduction figure parmi celles-ci dans 69,6 % des cas. Il est à noter que parmi les 30,4 % d'erreurs, se trouvent souvent des mots composés devant normalement figurer sous l'intitulé « aucune bonne traduction ».

Si l'on se penche sur les cas où cette seconde méthode n'identifie aucune bonne traduction, on constate que si ces cas sont peu nombreux (seulement 26,5 %), le taux d'erreur est faible (8,9 %).

	correct	faux
la bonne traduction	56,8 %	43,2 %
aucune bonne traduction	91,1 %	8,9 %
précison globale	65,9 %	34,1 %

Tableau 3. *Précison de la méthode de la coprésence*

4. La combinaison des deux méthodes

Nous avons enfin voulu savoir si le fait de combiner la première et la seconde méthode présentait un avantage par rapport à chaque méthode prise isolément. Nous

n'avons retenu que les cas où les deux méthodes valident ou invalident conjointement une même traduction.

Le tableau 4 montre que les résultats sont légèrement meilleurs, essentiellement lorsqu'aucune traduction correcte n'est trouvée.

	correct	faux
la bonne traduction	68,5 %	31,5 %
aucune bonne traduction	94,7 %	5,3 %
précision globale	74,5 %	25,5 %

Tableau 4. *Précison de la combinaison des deux méthodes*

Ces résultats doivent toutefois être relativisés dans la mesure où ils ne concernent que les cas où les deux méthodes fournissent la même « bonne » traduction (43,2 % des cas) ou s'accordent à indiquer qu'il n'en existe aucune parmi l'ensemble des traductions candidates pour un mot composé source (22,5 % des cas), ce qui correspond à 65,7 % des mots composés sources. Dans les autres cas (34,3 %), les résultats des deux méthodes divergent : il n'est pas possible de privilégier les résultats d'une méthode par rapport à l'autre si ce n'est lorsqu'on peut établir qu'une des traductions candidates possède exactement la même forme que le mot composé source, à l'exception des accents, auquel cas il faut choisir la méthode des proportions. Nous parlerons plus en détail de ce cas dans la section suivante.

5. Discussion

Quels sont les avantages de ces méthodes consistant à utiliser le Web pour valider ou invalider des traductions candidates lors de la reformulation multilingue de mots composés ?

Ces techniques permettent en premier lieu d'identifier les erreurs dans un dictionnaire de traduction multilingue. Ainsi que tous les résultats l'ont montré, chaque méthode permet, avec un plus ou moins grand taux de réussite (76,4 % pour la méthode des proportions et 91,1 % pour la méthode de la coprésence), de déterminer s'il n'existe aucune traduction correcte pour un mot composé dans une langue source. Il est ainsi possible de compléter très facilement le dictionnaire multilingue de mots simples afin de suppléer à ses carences. De la même façon, il devient aisé de détecter les traductions ne comportant pas le même nombre d'unités que le mot composé source ou dont un élément a changé de catégorie morphosyntaxique (« allocations familiales » (adjectif – nom) devient « family benefits » (nom – nom)).

Les deux techniques offrent ensuite la possibilité de créer semi-automatiquement — voire automatiquement à condition d'accepter des erreurs — des dictionnaires multilingues de mots composés. Il est par exemple possible d'extraire une liste de mots composés dans un corpus particulier, de générer l'ensemble des traductions candidates

pour chaque mot composé et, partant, d'utiliser l'une des méthodes décrites ci-dessus pour sélectionner les traductions correctes. Cela permet de limiter le nombre ou la taille des requêtes adressées à une base de données (pour rappel, l'un des 169 mots composés utilisés dans notre test possédait 728 traductions candidates).

Dans le domaine de la veille, ces techniques peuvent assister l'utilisateur dans la création d'un profil multilingue en sélectionnant les meilleures traductions pour chaque mot composé.

En outre, cette approche peut facilement être combinée à d'autres : il est ainsi possible de l'utiliser conjointement avec un modèle de traduction statistique intégrant les mots composés.

Ces méthodes ne sont néanmoins pas exemptes de problèmes.

La première difficulté est liée à la morphologie flexionnelle. Si la génération de traductions candidates ne pose aucun problème en chinois vu que cette langue ne connaît pas à proprement parler de morphologie, il n'en va déjà plus de même dans des langues à faible morphologie comme l'anglais ; la situation devient assez délicate dans le cas de langues synthétiques. Car, contrairement à un corpus strictement défini, le Web présente cet inconvénient de ne pouvoir être lemmatisé dans sa totalité, du moins jusque dans un futur plus ou moins proche. Il devient ainsi nécessaire de fléchir chaque traduction candidate.

Un second problème concerne les faits de société, les réalités culturelles, géographiques, environnementales propres à chaque langue ou à certaines régions employant une langue particulière. Ainsi, si la « cranberry » est bien connue en Amérique du Nord, la « canneberge » est nettement moins répandue dans le monde francophone (excepté pour 6,5 millions de locuteurs francophones canadiens), ce qui se traduit sur le Web par 3 530 000 pages pour le mot anglais contre 14 600 pour son équivalent français, sous Google.

Enfin, une difficulté mineure liée à la proximité des formes entre deux langues survient dans le cas de la méthode de la coprésence. Parmi les traductions candidates du mot composé « conséquence immédiate », on trouve « consequence immediate ». La plupart des moteurs de recherche sur le Web (dont Google) ne sont pas sensibles à la casse ou aux accents. La requête « "conséquence immédiate" + "consequence immediate" » se ramène à une simple requête « "consequence immediate" », ce qui a pour effet de renvoyer un nombre important de pages web (17 200 ici). Il convient de détecter ces cas afin de les traiter séparément, par exemple en utilisant la première technique qui permet de se rendre compte que le rapport est de 1 entre le mot composé source et sa traduction candidate.

6. Conclusion

Nous avons montré dans cet article que l'utilisation conjuguée du World Wide Web et d'un dictionnaire de reformulation multilingue permettait d'identifier les erreurs

de traduction de mots composés dans ce dictionnaire. Les deux techniques que nous avons présentées permettent en outre de déterminer les cas où une traduction terme à terme s'avère impossible. Dans près de 70 % des cas pour la méthode des proportions et 56 % des cas pour la méthode de la coprésence, ces techniques permettent en outre de sélectionner automatiquement la meilleure traduction pour peu que celle-ci ait été générée et fasse partie de la liste des traductions candidates. À l'exception des langues plus ou moins flexionnelles où il est nécessaire de faire appel à un logiciel de flexion, ces deux méthodes sont très faciles à mettre en œuvre dans le cadre de la recherche d'information translingue.

7. Bibliographie

[ALJ 01] ALJLAYL M., FRIEDER O., « Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation », *ACM Tenth Conference on Information and Knowledge Managemen (CIKM)*, Atlanta, Georgia, Novembre 2001.

[BAL 97] BALLESTEROS L., CROFT W. B., « Phrasal translation and query expansion techniques for cross-language information retrieval », *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997, p. 84–91.

[BOU 92] BOUILLON P., BOESEFELDT K., RUSSELL G., « Compound nouns in a unification-based MT system », *Proceedings of the third conference on Applied natural language processing*, Trento, Italie, 1992, p. 209-215.

[CHU 90] CHURCH K. W., HANKS P., « Word Association Norms, Mutual Information, and Lexicography », *Computational Linguistics*, vol. 16, n° 1, 1990.

[FED 02] FEDERICO M., BERTOLDI N., « Statistical Cross-Language Information Retrieval using N-best Query Translations », *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, août 2002, p. 167-174.

[FUJ 99a] FUJII A., ISHIKAWA T., « Cross-Language Information Retrieval for Technical Documents », *Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, p. 29-37.

[FUJ 99b] FUJII A., ISHIKAWA T., « Cross-Language Information Retrieval using Compound Word Translation », *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages (ICCPOL'99)*, mars 1999, p. 105-110.

[FUJ 01] FUJII A., ISHIKAWA T., « Japanese/English Cross-Language Information Retrieval : Exploration of Query Translation and Transliteration », *Computers and the Humanities*, vol. 35, n° 4, 2001, p. 389-420.

[GRE 98] GREFENSTETTE G., « Evaluating the adequacy of a multilingual transfer dictionary for the cross language information retrieval », *First International Conference on Language Resources and Evaluation*, Mai 1998, p. 755–758.

[GRE 99] GREFENSTETTE G., « The WWW as a resource for example-based MT tasks », *Proceedings of ASLIB'99 Translating and the Computer*, vol. 21, 1999.

[GRE 00] GREFENSTETTE G., NIOCHE J., « Estimation of English and non-English Language Use on the WWW », *Proceedings of RIAO'2000, Content-Based Multimedia Information Access*, Paris, 12–14 2000, p. 237–246.

[KAD 04] KADRI Y., NIE J.-Y., « Traduction de requêtes pour la recherche d'information translingue anglais-arabe », *JEP-TALN*, Fès, avril 2004.

[KAT 94] KATOH N., AIZAWA T., « Machine Translation of Sentences with Fixed Expressions », *Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany, 1994, p. 28 - 33.

[LEV 99] LEVOW G.-A., OARD D. W., « Evaluating lexicon coverage for cross-language information retrieval », *Workshop on Multilingual Information Processing and Asian Language Processing*, Novembre 1999, p. 69–74.

[QU 02] QU Y., GREFENSTETTE G., EVANS D. A., « Resolving Translation Ambiguity using Monolingual Corpora », *Working Notes for the CLEF 2002 Workshop*, 2002, p. 115–126.

[SAT 90] SATO S., NAGAO M., « Toward Memory-based Translation », *COLING-90*, vol. 3, 1990, p. 247–252.