Enrichissement d'un lexique bilingue par analogie

Philippe LANGLAIS, Alexandre PATRY
Université de Montréal
CP. 6128 succursale centre-ville
{felipe,patryale}@iro.umontreal.ca

Résumé. La présence de mots inconnus dans les applications langagières représente un défi de taille bien connu auquel n'échappe pas la traduction automatique. Les systèmes professionnels de traduction offrent à cet effet à leurs utilisateurs la possibilité d'enrichir un lexique de base avec de nouvelles entrées. Récemment, Stroppa et Yvon (2005) démontraient l'intérêt du raisonnement par analogie pour l'analyse morphologique d'une langue. Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adaptée au problème de la traduction d'entrées lexicales inconnues.

Abstract. Unknown words are a well-known hindrance to natural language applications. In particular, they drastically impact machine translation quality. An easy way out commercial translation systems usually offer their users is the possibility to add unknown words and their translations into a dedicated lexicon. Recently, Stroppa et Yvon (2005) shown how analogical learning alone deals nicely with morphology in different languages. In this study we show that analogical learning offers as well an elegant and efficient solution to the problem of identifying potential translations of unknown words.

Mots-clés : analogie formelle, enrichissement de lexiques bilingues, traduction automatique.

Keywords: formal analogy, bilingual lexicon projection, machine translation.

1 Introduction

Le raisonnement par analogie est un principe bien connu en sciences cognitives et en intelligence artificielle (Gentner *et al.*, 2001). L'aptitude à raisonner par analogie a longtemps fait l'objet de questions dans les tests SAT (Scholastic Assessment Test) aux États-Unis[1]. Turney et Littman (2005) décrivent une approche basée sur le modèle de l'espace vectoriel populaire en recherche d'information qui permet de répondre à 47% de 374 questions typiquement posées dans ces tests.

On trouve dans (Lepage, 2003) un traitement particulièrement riche du rôle de l'analogie dans la langue, tant d'un point de vue formel, algorithmique qu'historique. L'auteur décrit différentes expériences, notamment en analyse morphologique, qui attestent du bien-fondé applicatif de

[1] Les tests SAT introduits en 1926 aux États-unis incluaient des tests analogiques qui ont été retirés en 2005 (http://www.collegeboard.com/about/newstat/newstat.html).

l'analogie. Des principes dégagés dans ce travail, Lepage et Denoual (2005) présentaient récemment le système ALEPH, un système de traduction par l'exemple entièrement basé sur le principe de résolution d'analogies formelles. Une *analogie formelle* met en relation quatre entités, ce que l'on dénote $[\texttt{A} : \texttt{B} = \texttt{C} : \texttt{D}]$ et qui se lit: "A est à B ce que C est à D". Ce système d'une élégance remarquable[2] montrait des performances état-de-l'art dans les tâches partagées organisées en marge des ateliers IWSLT[3].

Stroppa et Yvon (2005) proposent une formalisation algébrique à la fois concise et accessible de l'analogie formelle et décrivent les fondements théoriques de l'apprentissage analogique. Ils démontrent expérimentalement l'élégance et la puissance de cette approche dans deux tâches d'étiquetage morphologique ; la première consistait à étiqueter morpho-syntaxiquement à l'aide d'un jeu d'étiquettes fines les mots inconnus d'une langue ; la seconde visait à prédire l'arbre d'analyse morphologique d'un lemme inconnu (lire également (Yvon *et al.*, 2004)).

D'autres auteurs se sont intéressés ces dernières années au potentiel applicatif des analogies formelles. Claveau et L'Homme (2005) montraient notamment qu'un type particulier d'analogies formelles très simples à calculer permettait de structurer les termes d'un domaine. Moreau et Claveau (2006) montrent également le bénéfice du raisonnement par analogie pour l'extension de requêtes dans un système monolingue de recherche d'information.

Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adéquate au problème concret de la traduction d'entrées lexicales inconnues. Nous rappelons en section 2 le principe général de l'apprentissage analogique. Nous présentons en section 3 comment il peut être appliqué à la tâche d'enrichissement d'un lexique bilingue. Nous évaluons notre approche en section 4 en la comparant à deux systèmes de base. Nous montrons que notre approche permet de traduire automatiquement 60% des mots inconnus d'une application. Nous dressons en section 5 un bilan de ce travail, et proposons des perspectives de recherche.

2 Raisonnement analogique

2.1 Apprentissage analogique

L'approche mise en place dans cette étude pour l'enrichissement de lexiques bilingues s'inscrit dans le cadre théorique de l'apprentissage par analogie proposé dans (Stroppa & Yvon, 2005). Un ensemble d'apprentissage $\mathcal{L} = \{L_1, \ldots, L_N\}$ est composé de N observations. Un ensemble de traits calculés sur une observation incomplète X définit un espace d'entrée. La tâche d'inférence consiste à prédire les traits manquants de X qui définissent à leur tour un espace de sortie. On désigne par $I(X)$ et $O(X)$ les projections respectives dans l'espace d'entrée et de sortie de l'observation X. La procédure d'inférence met en œuvre trois étapes:

1. construire $\mathcal{E}_{\mathcal{I}}(X) = \{(A, B, C) \in \mathcal{L}^3 \,|\, [\texttt{I(A)} : \texttt{I(B)} = \texttt{I(C)} : \texttt{I(X)}]\}$, l'ensemble de triplets analogiques de X, également dénommés *stems* dans la suite
2. construire $\mathcal{E}_{\mathcal{O}}(X) = \{Y \,|\, [\texttt{O(A)} : \texttt{O(B)} = \texttt{O(C)} : \texttt{Y}], \forall (A, B, C) \in \mathcal{E}_{\mathcal{I}}(X)\}$ l'ensemble des solutions trouvées aux équations analogiques formées par projection des triplets analogiques de X dans l'espace de sortie.
3. choisir $O(X)$ parmi les éléments de $\mathcal{E}_{\mathcal{O}}(X)$

[2]Ce système ne fait appel à aucune distance ni à aucun seuil pour rapprocher différents exemples.
[3]http://www.slc.atr.jp

Cette procédure d'inférence partage les avantages et les inconvénients de l'approche des k plus proches voisins (k-ppv). Il s'agit en effet d'une approche d'apprentissage passive qui n'effectue aucune généralisation à partir du corpus d'entraînement qui doit donc être conservé. Contrairement au k-ppv, l'étape 1 de recherche des exemples "proches" ne requiert pas la définition d'une distance entre deux exemples mais émerge du seul principe de *commutation linguistique* (Lepage, 2003). Cette pureté a un coût: la recherche d'exemples proches est une opération de complexité cubique en N, le nombre d'exemples dans $\mathcal{L}$, alors qu'elle est seulement linéaire en N dans le cas des k-ppv. Dans de nombreuses applications incluant celle présentée dans cette étude, cette recherche est trop coûteuse pour être effectuée au complet et des heuristiques doivent être appliquées pour réduire l'espace de recherche (voir section 3.2).

Le succès de l'approche repose presque entièrement sur le concept d'équation analogique que nous décrivons ci-après ainsi que sur l'hypothèse qu'il existe une correspondance entre les analogies construites sur l'espace d'entrée et leur projection dans l'espace de sortie.

2.2 Équation analogique

Différents niveaux paradigmatiques peuvent unir quatre objets en relation analogique. Ainsi des relations d'ordre sémantique comme celles utilisées dans les tests SAT peuvent être décrites: [hache : bûcheron = roulette : dentiste] et [aluminium : metal = novel : book]. Des analogies dites formelles avec lesquelles nous travaillons dans cette étude dénotent des relations graphiques entre les formes mises en relation. [fournit : fleurit = fournie : fleurie] et [abandoning : abandonment = amending : amendment] sont deux exemples (l'un en français, l'autre en anglais) de relations de nature morphologique. Le lecteur intéressé trouvera dans (Lepage, 2003) de nombreux exemples de relations analogiques dans des langues très différentes. L'équation analogique [A : B = C : ?] dénote l'ensemble des formes qui sont en relation analogique avec le triplet (ou stem) $\langle \mathtt{A}, \mathtt{B}, \mathtt{C} \rangle$:

$$[\mathtt{A} : \mathtt{B} = \mathtt{C} : \ ?] = \{X \mid [\mathtt{A} : \mathtt{B} = \mathtt{C} : \mathtt{X}]\}$$

Stroppa et Yvon (2005) montrent qu'il est possible de calculer les solutions d'une équation analogique formelle à l'aide d'un transducteur à états finis. Cette approche généralise l'algorithme proposé initialement par Lepage (1998) qui réside dans la synchronisation de deux tables d'éditions: l'une entre A et B, l'autre entre A et C. Intuitivement, cet algorithme compose dans le bon ordre les sous-séquences de B et de C qui ne sont pas dans A.

Dans ce travail, nous avons implémenté une variante de cet algorithme qui calcule premièrement les deux tables d'édition[4] (opération de complexité quadratique avec la longueur, comptée en caractères, des chaînes en présence), puis qui synchronise ensuite les deux tables (opération de complexité linéaire) pour tout chemin d'édition minimal de chaque table. Comme le nombre de chemins d'édition de coût minimal peut être exponentiel, nous considérons au plus les M premiers chemins de chaque table (dans nos expériences, M a été fixé expérimentalement à la valeur non critique de 20) et un total de M^2 paires de chemins est donc au plus synchronisé.

Il est important de noter qu'une équation peut admettre zéro, une ou plusieurs solutions qui ne sont pas nécessairement des formes légitimes de la langue étudiée. L'équation [fournir : fourniront = courir : ?] a par exemple pour solution couriront.

[4]Les coûts des opérations sont unitaires à l'exception de l'opération d'insertion qui est de coût nul, précisément car les caractères de B et C "insérés" sont ceux que nous désirons conserver dans les solutions.

3 Application à l'enrichissement d'un lexique bilingue

Les principes décrits dans la section précédente peuvent être appliqués au problème de l'enrichissement d'un lexique bilingue. Cette opération qui consiste à étendre un lexique existant à de nouvelles entrées présente de nombreux intérêts pratiques, notamment dans le cas de paires de langues faiblement dotées. La couverture d'un lexique, aussi grande soit-elle, n'est pas garante de son utilité. Ainsi, même pour une paire de langues largement dotée comme le français et l'anglais, n'existe-t-il pas de lexique bilingue couvrant les termes de tous les domaines de spécialité. Ceci justifie l'intérêt de travaux visant à apprendre automatiquement à traduire les termes spécifiques comme ceux du domaine médical (Claveau & Zweigenbaum, 2005).

3.1 Approche

Notre approche peut-être illustrée sur un exemple simple. Pour chercher la traduction du mot inconnu `futilité`, nous identifions des relations analogiques dans la langue source comme: [`activités : activité = futilités : futilité`]. Nous projetons (par une opération définie plus loin) ces relations en langue cible de manière à définir des équations analogiques (cibles) comme: [`actions : action = gimmicks : ?`] dont `gimmick` est une solution.

Formellement, nous disposons d'un corpus d'apprentissage $\mathcal{L} = \{\langle S_1, T_1\rangle, \ldots, \langle S_N, T_N\rangle\}$ qui réunit des paires de mots en relation de traduction. L'espace d'entrée est l'ensemble des mots de la langue source, l'espace de sortie celui des mots de la langue cible ; et on définit[5]:

$$\forall X \equiv \langle \mathtt{S}, \mathtt{T}\rangle,\ I(X) = S \text{ et } O(X) = T$$

L'enrichissement de $\mathcal{L}$ consiste pour toute forme source S inconnue de l'espace d'entrée à identifier les triplets analogiques sources qui entrent en équation analogique avec S:

$$\mathcal{E}_{\mathcal{I}}(S) = \{\langle \mathtt{i}, \mathtt{j}, \mathtt{k}\rangle, \in [1, N]^3 \mid S_i \neq S_j \neq S_k \text{ et } [S_i : S_j = S_k : S]\}$$

Chaque élément de $\mathcal{E}_{\mathcal{I}}(S)$ est ensuite projeté dans l'espace de sortie à l'aide de l'opérateur $proj_{\mathcal{L}}$ et les solutions calculées dans cet espace sont colligées dans $\mathcal{E}_{\mathcal{O}}(S)$:

$$\mathcal{E}_{\mathcal{O}}(S) = \bigcup_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle \in \mathcal{E}_{\mathcal{I}}(S)} \mathcal{E}_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle}(S)$$

où:

$$\mathcal{E}_{\langle \mathtt{i},\mathtt{j},\mathtt{k}\rangle}(S) = \{T \mid [U : V = W : T], \forall (U, V, W) \in (proj_{\mathcal{L}}(S_i) \times proj_{\mathcal{L}}(S_j) \times proj_{\mathcal{L}}(S_k))\}$$

Le mécanisme de projection que nous utilisons consiste à simplement à retourner pour une entrée source S du lexique bilingue les associations cibles (ou traductions) qui lui correspondent (une entrée S possède potentiellement plusieurs traductions dans $\mathcal{L}$):

$$proj_{\mathcal{L}}(S) = \{T \mid \langle S, T\rangle \in \mathcal{L}\}$$

[5]Par exemple, pour l'observation $X = \langle \mathtt{déjà}, \mathtt{already}\rangle$, $I(X) = \mathtt{déjà}$ et $O(X) = \mathtt{already}$.

3.2 Implémentation

Trouver l'ensemble des triplets analogiques de S est une opération trop coûteuse en temps (cubique avec la taille de l'espace d'entrée). Nous utilisons deux techniques pour réduire cette complexité. La première consiste à utiliser les équations analogiques en mode génératif: plutôt que de vérifier tous les triplets $\langle S_i, S_j, S_k \rangle$ entretenant une relation analogique avec S, nous cherchons les solutions à $[S_j : S_i = S :?]$. Il s'agit d'une méthode exacte qui repose sur la propriété (Lepage, 2003):

$$[\mathtt{A} : \mathtt{B} = \mathtt{C} : \mathtt{D}] \equiv [\mathtt{B} : \mathtt{A} = \mathtt{D} : \mathtt{C}]$$

Cette méthode, qui réduit la construction de $\mathcal{E}_\mathcal{I}(S)$ à une opération de complexité quadratique, est encore trop coûteuse. Nous appliquons donc une seconde méthode, cette fois-ci heuristique, qui consiste à ne calculer les équations analogiques que sur les seuls mots proches de S ; formellement, nous construisons $\mathcal{E}_\mathcal{I}(S)$ selon (étape 1):

$$\mathcal{E}_\mathcal{I}(S) = \{U \mid [A : B = S : U], \forall A \in v_\delta(S) \text{ et } B \in v_\beta(A)\}$$

où $v_\gamma(A)$ est une fonction de voisinage d'une lexie A de la forme:

$$v_\gamma(A) = \{B \mid f(B, A) \leq \gamma\}$$

Dans cette étude, nous avons utilisé pour fonction f la distance d'édition (Levenshtein, 1966)[6].

Nous avons mentionné qu'une équation analogique peut générer plusieurs solutions, certaines n'étant pas des formes légitimes d'une langue. Aussi, l'étape 3 du processus d'inférence consiste dans notre cas à sélectionner les solutions analogiques les plus fréquemment générées et à ne retenir que celles qui sont présentes dans un (grand) lexique monolingue cible $\mathcal{V}$. Nous avons compilé à cet effet à partir de textes variés un lexique monolingue totalisant 466 439 formes différentes. Des exemples de traductions produites par analogie sont présentés en Table 1.

4 Expériences

Nous avons réalisé nos expériences dans le cadre de la campagne d'évaluation des systèmes de traduction qui s'est tenue lors de l'atelier WMT'06 (Koehn & Monz, 2006). Dans cette tâche, les corpus d'entraînement et de test étaient constitués de textes parlementaires européens. À l'insu des équipes participantes, les organisateurs ont ajouté aux 2 000 phrases du corpus de test (corpus `domaine` dans la suite), 1064 phrases[7] hors-domaine en provenance du site internet de `Project Syndicate` (`http://www.project-syndicate.com`), une organisation sans but lucratif qui distribue des articles de revue sur des thèmes variés (politique, économie, science, etc.). Ce corpus est baptisé `hors-domaine` dans la suite.

Nos expérimentations simulent une situation typique du développement d'un système de traduction basé sur l'exemple: nous disposons d'un corpus d'entraînement bilingue sur lequel est

[6] Le lecteur attentif aura noté que nous avons plus haut (section 2.1) souligné que contrairement à l'approche des K plus proches voisins, le raisonnement par analogie ne requiert pas de distance. La distance que nous utilisons ici n'est pas constitutive de l'approche (comme c'est le cas dans les k-ppv) mais répond seulement à des considérations pratiques: nous pourrions par exemple nous en affranchir en tirant aléatoirement des triplets dans l'espace d'entrée.

[7] 30 de ces phrases contenaient des problèmes d'encodage et ont été retirées de notre étude.

source	cand	nb	(candidat, fréquence)
anti-agricole	296	5	(anti-farm,5) (anti-agricultural,3) (anti-farming,3) (anti-rural,3) (anti-farmer,3)
concentrerait	2947	7	**(concentrat,11)** (concentrate,4) (summarized,4) (summarizing,4) (concentrating,3) (focuss,3) (focus,3)
écrirait	156	4	**(writs,1)** (write,1) (writes,1) **(writ,1)**
réintégrés	2686	18	(reinstated,20) (reintegrated,17) (re-integrated,13) (re-entered,10) (reincluded,8) (reinvolved,8) (reincorporated,8) (reinserted,7) (reinstate,7) (reintegrate,6) (reinstating,4) (accomplished,3) (rebuilt,3) (reinclude,3) (rejoined,3) (reverte,2) (reintegration,2) (reintegrating,2)
galette	218	1	(pancake,13)

TAB. 1 – Exemples de traductions obtenues par projection à partir de $\mathcal{L}_{100\,000}$. cand indique le nombre de solutions analogiques cibles générées, nb indique les traductions candidates retenues une fois validées par le lexique $\mathcal{V}$. Les traductions en gras sont clairement erronées.

appris un lexique bilingue (probabiliste dans notre cas). Nous disposons de plus d'une (grande) collection de textes en langue cible que nous avons utilisée ici pour compiler le lexique monolingue cible $\mathcal{V}$ (voir la section 3.2). Afin de bien analyser les limites de l'approche, nous nous sommes concentrés sur la paire de langues français-anglais qui nous est familière[8]. Notre but est de prédire des traductions anglaises de termes français inconnus du corpus d'entraînement. Nous avons éliminé de notre étude les formes numériques (nous pouvons les traiter de manière simple).

4.1 Évaluation automatique

Évaluer la qualité de différentes variantes de notre approche nécessite le parcours de plusieurs listes de traductions. En plus de s'avérer fastidieuse, cette entreprise s'avère délicate: beaucoup de traductions produites ne sont valides que dans certains contextes seulement. Il suffit pour cela de consulter les exemples de la Table 1 pour s'apercevoir de la difficulté de la tâche.

Nous avons donc, dans un premier temps, procédé à une évaluation automatique où un lexique bilingue de référence est utilisé. Ce lexique, dénommé $\mathcal{L}_{ref}$, est obtenu par entraînement sur le corpus au complet de WMT'06 (688 000 paires de phrases) d'un modèle statistique obtenu à l'aide de la trousse à outils GIZA++ (Och & Ney, 2000)[9]. De la même manière, nous avons entraîné des modèles lexicaux $\mathcal{L}_T$ sur différentes tranches du corpus d'entraînement ($T = 5\,000$, 10 000, 100 000, 200 000 et 500 000 paires de phrases). Nous avons alors traduit à l'aide du raisonnement analogique les mots français du corpus de test de WMT'06 qui n'étaient pas présents dans les lexiques $\mathcal{L}_T$ mais présents dans le lexique de référence $\mathcal{L}_{ref}$. Une traduction candidate est considérée correcte si elle est validée par $\mathcal{L}_{ref}$.

À des fins de comparaison, deux approches de base (*baseline*) ont été testées sur la même

[8]Des résultats similaires sont observés pour la paire de langues espagnol-anglais.

[9]En pratique, pour éliminer une partie du bruit d'un lexique appris automatiquement, nous le croisons (intersection) avec un lexique résultant de l'entraînement d'un modèle lexical entraîné dans la direction opposée (anglais-français versus français-anglais).

tâche dans les mêmes conditions. La première approche (BASE1) consiste à proposer comme traduction d'un mot source inconnu, les mots cibles les plus similaires (au sens de la distance d'édition). Cette approche marchera d'autant mieux que les langues sont proches (ex. *docteur* → *doctor*). La deuxième approche (BASE2) ressemble davantage à l'approche analogique et consiste à identifier les formes sources connues du lexique $\mathcal{L}_T$ qui sont proches du mot inconnu, puis à proposer leurs traductions telles qu'indiquées par ce lexique (ex. *demanda* → *demande* → *request*). Chacune de ces approches est testée selon deux variantes. La première ($_{id}$) propose le même nombre de traductions que l'a suggéré l'approche ANALOG dans les mêmes conditions (les approches sont donc directement comparables) ; la seconde ($_{10}$) propose dix traductions pour chaque mot inconnu.

T	5 000		10 000		50 000		100 000		200 000		500 000	
	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%
	domaine											
ANALOG	50.8	30.7	54.4	44.3	57.9	63.9	57.0	63.8	57.7	64.4	30.4	67.6
BASE1$_{id}$	31.6	30.7	32.3	44.3	24.7	63.9	20.3	63.8	20.9	64.4	8.7	67.6
BASE2$_{id}$	34.5	30.7	37.1	44.3	39.0	63.9	37.8	63.8	34.4	64.4	56.5	67.6
BASE1$_{10}$	26.7	100.0	28.3	100.0	23.9	100.0	20.0	100.0	16.6	100.0	11.8	100.0
BASE2$_{10}$	26.3	100.0	30.8	100.0	29.3	100.0	27.6	100.0	24.9	100.0	55.9	100.0
unk	[3 171 , 6.8]		[2 245 , 6.1]		[754 , 3.7]		[456 , 2.8]		[253 , 2.0]		[34 , 1.2]	
	hors-domaine											
ANALOG	52.4	28.9	54.4	42.4	51.7	68.0	53.6	73.4	55.3	79.2	43.9	86.8
BASE1$_{id}$	28.0	28.9	29.0	42.4	27.3	68.0	23.1	73.4	26.8	79.2	22.7	86.8
BASE2$_{id}$	32.9	28.9	35.0	42.4	32.5	68.0	35.9	73.4	40.8	79.2	59.1	86.8
BASE1$_{10}$	24.7	100.0	25.9	100.0	25.1	100.0	20.9	100.0	25.2	100.0	25.0	100.0
BASE2$_{10}$	21.7	100.0	26.4	100.0	27.2	100.0	29.4	100.0	33.6	100.0	57.9	100.0
unk	[2 270 , 6.2]		[1 701 , 5.5]		[621 , 3.2]		[402 , 2.4]		[226 , 1.7]		[76 , 1.3]	

TAB. 2 – Résultats de différentes méthodes d'extension de lexique en fonction de la taille du lexique à étendre (ligne du haut). Les lignes préfixées de *unk* indiquent le nombre de mots à traduire ainsi que le nombre moyen de traductions dans $\mathcal{L}_{ref}$.

Les résultats de cette évaluation automatique sont consignés en Table 2 en fonction de la taille du lexique utilisé pour la projection. Deux mesures sont rapportées: `p%` représente le pourcentage d'entrées inconnues ayant au moins une traduction valide ; `r%` indique le pourcentage de mots traduits. La première ligne de cette table indique par exemple que sur le corpus de test `domaine`, notre approche (ANALOG) propose au moins une traduction pour 30.7% des mots du corpus de test du domaine qui sont inconnus du lexique $\mathcal{L}_{5000}$. La moitié (50.8%) de ces entrées étendues contiennent une traduction valide (selon $\mathcal{L}_{ref}$).

La Table 2 appelle plusieurs commentaires. Il convient tout d'abord de garder à l'esprit que ce que nous mesurons ici est davantage l'aptitude d'une approche à reconstruire un lexique de référence à partir d'un lexique de base. Nous ne mesurons par exemple pas ici les traductions produites pour les mots inconnus du lexique de référence. Globalement, les performances de ANALOG augmentent avec la taille du lexique de base. Ceci est normal car plus ce lexique est grand, plus il contient de formes sources qui peuvent entrer en relation analogique avec un mot inconnu et plus les traductions de ces mots sont nombreuses, ce qui permet de créer davantage de relations analogiques dans la langue cible. Les mesures faites à l'aide du lexique $\mathcal{L}_{500\,000}$

sont certainement peu fiables: seulement 34 et 76 entrées sont en effet évaluées sur le jeu de test `domaine` et `hors-domaine` respectivement. On remarque également que les approches BASE1 et BASE2 sont inférieures en qualité à l'approche ANALOG (BASE1 est sans surprise la moins bonne des trois). Il est possible avec les approches de proximité d'obtenir un taux de traduction `r%` parfait, au prix d'une détérioration de la qualité des traduction produites, ce qui suggère qu'une combinaison des deux approches (comme par exemple écouter BASE2 lorsque ANALOG est silencieuse) améliorerait les performances globales.

Nous avons évalué que pour la moitié des entrées non traduites, c'est une absence de relation analogique identifiée dans l'espace cible qui aboutit à l'absence de candidat. En moyenne sur le lexique $\mathcal{L}_{100\,000}$, une entrée inconnue entre en relation analogique 988 fois du côté source, ce qui génère en moyenne 52 formes sources qui appartiennent au lexique de projection $\mathcal{L}_T$. Du côté cible, une moyenne de 99 solutions analogiques sont proposées (par forme inconnue source) ; une moyenne de 5 d'entre-elles seulement sont validées par $\mathcal{V}$ et donc considérées ici.

4.2 Évaluation manuelle

Une inspection des traductions proposées révèle certains problèmes dont cette évaluation ne rend pas compte. En particulier, certaines entrées reçoivent une traduction correcte alors que le lexique de référence est erroné ou incomplet. C'est par exemple le cas des exemples de la Figure 1. Par exemple, `circumventing` et `fellow` sont des traductions légitimes de `coutournant` et `concitoyen` respectivement. Sur les 20 premières entrées lexicales considérées erronées par notre procédure d'évaluation, 12 contenaient des traductions valides et 4 des entrées étaient mal traduites dans le lexique de référence.

contournant (*49 candidats*)
`analog` ⋄ (circumventing,55) (undermining,20) (evading,19) (circumvented,17) (overturning,16) (circumvent,15) (circumvention,15) (bypass,13) (evade,13) (skirt,12) $\mathcal{L}_{ref}$ ⋄ **skirting**, **bypassing**, by-pass, overcoming
concitoyen (*24 candidats*)
`analog` ⋄ (citizens,26) (fellow,26) (**fellow-citizens**,26) (people,26) (citizen,23) (fellow-citizen,21) (fellows,5) (peoples,3) (civils,3) (fellowship,2) $\mathcal{L}_{ref}$ ⋄ **fellow-citizens**

FIG. 1 – Les 10 meilleures traductions produites par `analog` à patir de $\mathcal{L}_{200\,000}$ pour deux mots inconnus et leurs traductions dans $\mathcal{L}_{ref}$. Les traductions en gras sont présentes dans la liste candidate et la liste de référence.

Nous avons donc entrepris une évaluation manuelle des traductions proposées par les deux approches ANALOG et BASE2 pour les 127 termes du corpus `domaine`[10] inconnus de $\mathcal{L}_{ref}$. Nous avons décidé, non sans arbitraire, d'identifier comme valide une traduction candidate dès lors qu'elle était synonyme d'une traduction possible du mot source (*citizen* était par exemple considérée comme une traduction acceptable de *concitoyen*).

75 (60%) des mots inconnus recevaient au moins une traduction valide avec la première méthode, contre 63 (50%) avec la seconde. Sur ces mots, 61 traductions (81%) étaient proposées

[10]Nous n'avons pas observé de différence notable dans l'évaluation automatique sur les mots du et hors-domaine.

en tête par ANALOG contre 22 (35%) pour BASE2. Des 52 mots n'ayant pas reçu par ANALOG de traduction satisfaisante, 38 (73%) n'ont en fait reçu aucune traduction. Ces mots sont en majorité des noms propres, des mots d'une autre langue (Latin, Grec ou Anglais) ainsi que des mots composés.

Nous concluons de cette évaluation manuelle que le raisonnement par analogie permet pour la paire de langue français-anglais de proposer une traduction valide pour 80% des entrées inconnues *simples* (c'est-à-dire excluant les noms propres, les mots d'emprunts, les mots composés et les données chiffrées) de notre jeu de test.

5 Discussion et perspectives

Nous avons montré que le raisonnement analogique permettait de proposer une traduction valide à 60% des mots inconnus d'un jeu de test particulier (soit 80% des mots simples) pour la paire de langues français-anglais. Nous menons actuellement des expériences sur d'autres paires de langues disponibles dans le cadre de la tâche partagée de WMT'06. Des taux similaires sont observés pour la paire de langue Espagnol-Anglais, alors qu'une perte d'environ 10% est observée sur la paire allemand-anglais.

En plus de prédire la traduction de mots inconnus, nous avons remarqué que cette technique peut-être utilisée pour enrichir les traductions d'entrées lexicales peu fréquentes dans le corpus d'entraînement. Ces entrées sont souvent mal traduites par les approches probabilistes. Nous avons également observé qu'il était envisageable d'appliquer `analog` à la traduction de séquences de mots. Nous planifions donc d'étudier de manière systématique l'impact de notre approche sur un système de traduction statistique basé sur les séquences de mots.

Toute analogie n'est pas bonne à faire ; aussi souhaitons-nous apprendre automatiquement à prédire la productivité d'une analogie. Ceci offrirait notamment une méthode de sélection plus efficace que la simple fréquence que nous avons considérée dans cette étude. Consulter le corpus d'entraînement lors d'une traduction n'est pas une approche satisfaisante. Nous souhaitons donc modéliser les régularités que recèlent les équations analogiques que nous formulons, dans la veine des travaux décrits par (Claveau & Zweigenbaum, 2005). Traduire un mot inconnu pourrait alors se faire par application de règles plutôt que par consultation d'exemples.

Plusieurs auteurs se sont intéressés à l'identification de traductions dans des corpus comparables, soit pour des mots simples (Fung & Yee, 1998; Rapp, 1999; Takaaki & Matsuo, 1999), soit pour des termes de spécialité (Morin & Daille, 2004). Les techniques proposées dans ces travaux peuvent être employées à l'enrichissement d'un lexique bilingue. Il convient cependant de souligner que contrairement à ces approches, les traductions que nous proposons émergent du seul principe de l'analogie. Nous ne sommes donc pas soumis au problème non trivial de l'acquisition de corpus dédiés (parallèles ou non) qui doivent contenir les mots que nous avons à traduire ainsi que leurs traductions.

Remerciements

Cette étude a largement profité de discussions que nous avons eues avec Nicolas Stroppa et François Yvon ainsi que du tutoriel donné à TALN'06 par Yves Lepage. Nous remercions les relecteurs anonymes pour la pertinence de leurs commentaires.

Références

CLAVEAU V. & L'HOMME M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie - utilisation comparée de ressources endogènes et exogènes. In *6ème rencontre de Terminologie et Intelligence Artificielle (TIA'05)*, Rouen, France.

CLAVEAU V. & ZWEIGENBAUM P. (2005). Automatic translation of biomedical terms by supervised transducer inference. In *10th Conference on Artificial Intelligence in Medicine (AIME'05)*, Aberdeen, Écosse.

FUNG P. & YEE L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of the 36th ACL*, p. 414–420, San Francisco, California.

GENTNER D., HOLYOAK K. J. & KONIKOV B. N. (2001). *The Analogical Mind*. Cambridge, MA: MIT Press.

KOEHN P. & MONZ C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, p. 102–121, New York City: Association for Computational Linguistics.

LEPAGE Y. (1998). Solving analogies on words: an algorithm. In *COLING-ACL*, p. 728–734.

LEPAGE Y. (2003). De l'analogie rendant compte de la commutation en linguistique. Mémoire d'Habilitation à diriger des recherches, Université Joseph Fourier, Grenoble I.

LEPAGE Y. & DENOUAL E. (2005). Aleph: an ebmt system based on the preservation of proportionnal analogies between sentences across languages. In *International Workshop on Statistical Language Translation (IWSLT)*, Pittsburgh, PA.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, **6**, 707–710.

MOREAU F. & CLAVEAU V. (2006). Extensions de requêtes par relations morpho-syntaxiques apprises automatiquement. In *3ème Conférence en Recherche d'Informations et Applications (CORIA'06)*, Lyon, France.

MORIN E. & DAILLE B. (2004). Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *TAL*, **45(3)**, 103–122.

OCH F. & NEY H. (2000). Improved statistical alignment models. In *38th annual meeting of the Association for Computational Linguistics (ACL'00)*, p. 440–447, Hongkong, China.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *37th annual meeting of the Association for Computational Linguistics (ACL'99)*, p. 519–526, College Park, Maryland.

STROPPA N. & YVON F. (2005). An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, p. 120–127, Ann Arbor, MI.

TAKAAKI T. & MATSUO Y. (1999). Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'99)*, p. 109–119, Chester, England.

TURNEY P. & LITTMAN M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning Journal*, **60**(1-3), 251–278.

YVON F., STROPPA N., DELHAY A. & MICLET L. (2004). *Solving analogical equations on words*. Rapport interne, École Nationale Supérieure des Télécommunications, Paris, France.