TALN 2010, Montréal, 19-23 juillet 2010

Recherche contextuelle d'équivalents en banque de terminologie

Caroline Barrière
ITI-CNR, Gatineau, Canada
caroline.barriere@nrc-cnrc.gc.ca

Résumé Notre recherche démontre que l'utilisation du contenu d'un texte à traduire permet de mieux cibler dans une banque de terminologie les équivalents terminologiques pertinents à ce texte. Une banque de terminologie a comme particularité qu'elle catégorise ses entrées (fiches) en leur assignant un ou des domaines provenant d'une liste de domaines préétablie. La stratégie ici présentée repose sur l'utilisation de cette information sur les domaines. Un algorithme a été développé pour l'assignation automatique d'un profil de domaines à un texte. Celui-ci est combiné à un algorithme d'appariement entre les domaines d'un terme présent dans la banque de terminologie et le profil de domaines du texte. Pour notre expérimentation, des résumés bilingues (français et anglais) provenant de huit revues scientifiques nous fournissent un ensemble de 1130 paires d'équivalents terminologiques et le Grand Dictionnaire Terminologique (Office Québécois de la Langue Française) nous sert de ressource terminologique. Sur notre ensemble, nous démontrons une réduction de 75% du rang moyen de l'équivalent correct en comparaison avec un choix au hasard.

Abstract Our research shows the usefulness of taking into account the context of a term within a text to be translated to better find an appropriate term equivalent for it in a term bank. A term bank has the particularity of categorising its records by assigning them one or more domains from a pre-established list of domains. The strategy presented here uses this domain information. An algorithm has been developed to automatically assign a domain profile to a source text. It is then combined with another algorithm which finds a match between a term's domains (as found in the term bank) and the text's domain profile. For our experimentation, bilingual abstracts (French-English) from eight scientific journals provide 1130 pairs of term equivalents. The Grand Dictionnaire Terminologique (Office Québécois de la Langue Française) is used as a terminological ressource. On our data set, we show a reduction of 75% in the average rank of the correct equivalent, in comparison to a random choice.

Mots-clés : recherche contextuelle, équivalents terminologiques, banque de terminologie, désambiguïsation par domaine

Keywords: contextual search, term equivalents, term bank, domain-based disambiguation

1 Introduction

Une banque de terminologie contient un ensemble de fiches terminologiques. Chaque fiche représente un concept (une notion) et à ce concept est associé un ou plusieurs termes dans une ou plusieurs langues. Une banque de terminologie est souvent multilingue. Le Grand Dictionnaire Terminologique (GDT) de l'Office Québécois de la Langue Française[1] qui nous servira pour notre expérimentation a comme langues de base le français et l'anglais. Ainsi, nous dirons que pour un concept, il existe une fiche terminologique contenant des termes équivalents en français et anglais servant à nommer ce concept dans chacune des langues.

Les banques de terminologie ont pour but de répertorier les concepts appartenant à divers domaines de spécialité (mécanique automobile, finances, biologie moléculaire, etc), et ainsi, elles catégorisent leurs entrées (fiches terminologiques) en leur assignant un ou des domaines provenant d'une liste de domaines préétablie. Ces banques sont utilisées par les rédacteurs, journalistes, enseignants ou autres langagiers à la recherche de termes justes et les traducteurs à la recherche d'équivalents corrects.

Même si l'organisation sous-jacente de la banque de terminologie est onomasiologique (par concept), la recherche dans cette banque ne peut se faire que par terme. Un terme étant souvent polysémique, le traducteur doit alors consulter les diverses fiches correspondant à ce terme pour découvrir laquelle exprime la notion recherchée. Cette recherche est essentielle car les différentes notions sont souvent liées à des équivalents différents dans les autres langues. Par exemple, la fiche pour *stem* dans le domaine de l'emballage et conditionnement donnera *tige de commande* comme équivalent, tandis que celle dans le domaine de la linguistique donnera *racine*, ou encore celle dans le domaine de l'aéronautique donnera *étrave*.

Les interfaces aux banques terminologies varient, mais en général, lorsqu'un traducteur cherche l'équivalent d'un terme, il peut consulter la banque de terminologie en regardant de façon exhaustive toutes les fiches correspondantes au terme, ou il peut parcourir d'abord la liste des domaines répertoriés pour ce terme, choisir un domaine et ensuite consulter les fiches de ce domaine. Parcourir cette liste de domaine peut prendre du temps pour retrouver ceux qui conviennent. Le terme *stem* fera partie de 41 domaines, aussi variés que aéronautique, agriculture, horlogerie, cosmétologie, manutention et stockage, équipement ménager, loisir, linguistique, marine, plomberie, emballage et conditionnement, sports, imprimerie ou tabac.

Idéalement, le traducteur devrait pouvoir préciser à l'avance le domaine de son texte et réduire ainsi la portion adéquate de la banque de terminologie. Malheureusement les frontières entre les domaines sont floues (par exemple, environnement et agriculture, métallurgie et mines), et comme plusieurs notions recoupent plusieurs domaines, l'énumération à l'avance de tous les domaines couverts par un texte à traduire serait presque impossible.

Le but applicatif de notre projet est donc d'aider le traducteur à faire plus rapidement la recherche de fiches terminologiques en filtrant l'ensemble des fiches répertoriées dans la banque pour lui proposer les fiches les plus pertinentes à son texte. Nos buts technologiques sont alors (1) établir automatiquement un profil de domaines pour un texte (ou une portion de

[1] Le Grand Dictionnaire Terminologique est accessible gratuitement à http://www.granddictionnaire.com, et il est aussi accessible par le site officiel de l'Office québécois de la langue française (http://www.oqlf.gouv.gc.ca).

celui-ci) en choisissant parmi les domaines de la banque de terminologie[2], et (2) utiliser ce profil pour apparier un terme présent dans le texte à sa fiche terminologique et ainsi retrouver son équivalent.

Notre tâche est similaire à une tâche de désambiguïsation par sens mais adaptée à la langue de spécialité et au contenu d'une banque terminologique qui est différent du contenu d'un dictionnaire. Ainsi, quoique la désambiguïsation par sens ait été largement étudiée dans la communauté en linguistique computationnelle, ayant même donné lieu à des compétitions internationales (*Senseval-3* Mihalcea and Edmonds 2004, *SemEval-1* Agirre et al. 2007), la désambiguïsation automatique de termes spécialisés n'a pas beaucoup été étudiée, et nous ne retrouvons pas de travaux sur ce sujet. Cela est peut-être dû à une conception fausse que les termes spécialisés sont monosémiques. Il est vrai que les termes composés sont souvent monosémiques, mais la langue de spécialité est aussi remplie de termes simples ayant divers sens dans divers domaines, en plus parfois d'avoir aussi un sens répandu dans la langue générale. Le terme *stress* en anglais par exemple est très répandu et le plus souvent traduit par *stress* en français, mais ce terme prend soudainement un sens très particulier dans le domaine de la construction et devra se traduire par *fatigue*. Barrière (2007) présente une comparaison entre les ressources lexicographiques et terminologiques en particulier quant à leur utilité pour la désambiguïsation du sens.

La majeure partie des travaux en désambiguïsation sémantique utilise la ressource WordNet (de par sa large couverture et sa disponibilité). Ces travaux proposent aussi l'utilisation du contexte d'usage des mots pour aider à les désambiguïser, mais leur point de comparaison repose sur les définitions des divers sens répertoriés (algorithmes de type *Lesk*, référant à Lesk 1986). Malheureusement, même si une ressource terminologique peut contenir de courtes définitions, cette information n'est pas obligatoire, contrairement à un dictionnaire, rendant inutilisable tous ces algorithmes développés pour WordNet. En banque de terminologie, c'est plutôt l'assignation d'un ou plusieurs domaines qui est l'information minimale d'une fiche.

Les travaux de Gliozzo et al. (2004) utilisent les « subject codes » ou domaines ajoutés à WordNet (Magnini et Cavaglià 2000) pour déterminer le domaine d'un texte dans lequel apparaît un mot ambigu. Ce domaine du texte sert alors de caractéristique additionnelle (plutôt qu'unique pour nous) pour la désambiguïsation du mot, en surplus des critères habituels (comparaison aux définitions) tel que mentionné ci-haut. Ces domaines ajoutés à certaines entrées de WordNet sont peu souvent utilisés comme critères de désambiguïsation, mais plutôt comme système de catégorisation des mots à un niveau de granularité moins précis. Par exemple, le terme *bank* a 10 sens, mais n'a que 5 domaines possibles car économie regroupe 4 sens, géographie regroupe 2 sens, architecture regroupe 2 sens, transport regroupe 1 sens et il reste un sens général. Dans les travaux de Suarez et Palomar (2006) par exemple, il s'agit de comparer des algorithmes de désambiguïsation en produisant des résultats pour deux niveaux de granularité, soit fine avec les sens et plus grossière avec les domaines.

La ressource utilisée pour le travail présent est le Grand Dictionnaire Terminologique de l'Office Québécois de la Langue Française (GDT)[3]. Le tableau 1 montre des exemples de

[2] Il reste à valider si ce premier but serait aussi un but applicatif et non-seulement un intermédiaire algorithmique pour le but ultime d'aider le traducteur à être plus productif en lui proposant le bon équivalent. Un logiciel tel LogiTerm (http://www.terminotix.com) suggère une « analyse de domaine » dans sa version Entreprise, ce qui porte à croire que cette tâche en elle-même a une valeur ajoutée.

[3] Nous remercions l'Office Québécois de la Langue Française pour l'accès gratuit à la base de données du Grand Dictionnaire Terminologique qu'il a accordé au Groupe de Technologies Langagières Interactives pour des buts de recherche.

termes anglais extraits d'une revue scientifique traitant de biologie avec des statistiques sur leur présence dans le GDT. Pour chaque terme anglais, on peut voir des exemples d'équivalents français, le nombre de fiches, de domaines et d'équivalents présents. Le nombre d'équivalents n'est pas identique au nombre de fiches car deux notions différentes (deux fiches) peuvent avoir le même terme équivalent. Le nombre de domaines n'est pas non plus identique au nombre de fiches car une fiche peut être assignée à deux domaines, ou deux fiches pourraient appartenir à un seul domaine (dans le cas d'un terme polysémique même à l'intérieur d'un domaine). Les relations entre fiches, domaines et équivalents ne sont pas de type un-à-un mais plutôt de type plusieurs-à-plusieurs rendant le problème plus complexe.

Terme anglais	Exemples de termes français (GDT)	Nombre de fiches	Nombre de domaines	Nombre d'équivalents
stress	contrainte, effort, fatigue, sollicitation, stress	46	36	16
activity	activité, exercice, mouvementation	19	17	3
feather	plume, auréole, barbule, clavette, fouet	17	29	12
membrane	membrane, feuillet de parchemin, paroi étanche	8	8	4
calcium	calcium	5	6	1
infrared spectroscopy	spectroscopie infrarouge, spectroscopie de l'infrarouge	3	6	2
fibrosis	fibrose	2	2	1
heat shock	choc thermique	1	2	1
haloperidol	halopéridol	1	1	1

Tableau 1 – Exemples de termes d'une revue retrouvés dans le GDT

Ainsi, pour aider le traducteur, l'idéal serait de lui proposer la bonne fiche correspondant à son texte, ou du moins lui montrer les fiches en ordre de pertinence. Dans notre expérimentation, pour pouvoir réellement évaluer notre succès à cette tâche, nous aurions besoin de textes annotés sémantiquement, ce qui serait long et coûteux à obtenir. Nous avons par contre accès à des paires d'équivalents en contexte (voir description à la section 2) et nous pourrons ainsi évaluer notre capacité à retrouver le bon équivalent, sachant que ceci n'est qu'une approximation (tout de même raisonnable) de la tâche.

La section 2 présente l'ensemble des données utilisé, la section 3 présente l'algorithme de désambiguïsation utilisé, la section 4 présente les résultats obtenus et une analyse des résultats et des difficultés, et finalement la section 5 nous permet de conclure et d'envisager des travaux futurs.

2 Définition du problème et ensemble de données

Nous avons accès à huit revues scientifiques contenant chacune un ensemble d'articles[4]. Chaque article possède un résumé en anglais et un résumé en français tel que traduit par la rédaction. Chaque article contient aussi de 2 à 5 mots-clés donnés par les auteurs et traduits par la rédaction. L'alignement de ces mots-clés dans les fichiers est suffisamment cohérent pour faire l'extraction automatique des paires correspondantes. Il y a quelques anomalies (inversions, oublis), mais elles sont rares (moins de 1%). Ainsi, la constitution de notre ensemble de données, soit l'ensemble des paires de termes équivalents, peut se faire

[4] Merci à l'ICIST (Institut Canadien de l'Information Scientifique et Technique) pour l'obtention du contenu de ces revues.

entièrement automatiquement à partir des listes de mots-clés des auteurs. De cette liste, nous utilisons uniquement les mots-clés se retrouvant dans le résumé[5].

Le tableau 2 montre sur l'ensemble des revues, le nombre d'articles associés, et le nombre de paires de termes équivalents extraits. Nous voyons que le GDT ne sera utile que pour une petite portion des données. Nous pouvons voir que des 10944 paires extraites de l'ensemble des revues scientifiques, seulement 13.6% sont présentes dans le GDT et contiennent plusieurs équivalents. Les termes qui ne possèdent qu'une seule fiche ne sont pas des candidats intéressants pour notre tâche de désambiguïsation. De plus, même si plusieurs fiches sont présentes dans le GDT, parfois l'équivalent désiré (tel que cité dans l'article scientifique) ne s'y retrouve pas. Notre pourcentage d'intérêt est alors réduit à 10.3% pour un total de 1130 paires de termes polysémiques dont l'équivalent désiré est répertorié et qui serviront donc à notre expérimentation.

Revue Scientifique	Nombre de résumés	Nb de paires de termes	% paires ne se retrouvant pas dans le GDT	% paires n'ayant qu'un seul équivalent dans le GDT	% paires ayant plus d'un équivalent dans le GDT
Biochemistry and cell biology	472	736	64.9%	24.0%	11.0%
Canadian Geotechnical Journal	618	1047	54.8%	17.6%	28.6%
Canadian Journal of Botany	1002	1925	72.9%	15.6%	11.5%
Canadian journal of chemistry	1341	1846	73.1%	19.4%	7.4%
Canadian journal of civil engineering	608	1239	53.1%	17.8%	29.1%
Canadian journal of microbiology	852	1456	71.0%	19.4%	9.6%
Canadian journal of physiology and pharmacology	846	1504	63.5%	26.4%	10.1%
Genome	854	1191	76.2%	15.5%	8.2%
TOTAL	6593	10944	67.3%	19.2%	13.6%

Tableau 2 – Présence des termes dans le GDT

3 Recherche contextuelle de fiche terminologique

Le processus se divise en 2 étapes, soit (1) déterminer le profil de domaines associés au texte, et (2) en fonction de ce profil, déterminer la fiche correspondant au terme en contexte.

Etape 1 - Assignation d'un profil de domaines à un texte

1) Initialiser une liste de domaines (AllDomains)
2) Extraire un ensemble de termes du contexte (T_1 à T_N)
3) Pour chacun des termes T_x
    a. Chercher les domaines du GDT correspondant aux fiches répertoriées pour ce terme (D_1 à D_Z)
    b. Pour chacun de ces domaines D_y
        i. Assigner une valeur inversement proportionnelle au nombre de domaines que le terme couvre : $DValue(D_y) = 1 / nbDomaines(T_x) + 0.5$
        ii. Si le domaine se retrouve dans la liste AllDomains, alors incrémenter sa valeur dans la liste de $DValue(D_y)$
        iii. Sinon, alors ajouter le domaine à la liste AllDomains avec sa valeur $DValue(D_y)$
4) Trier en ordre décroissant de valeurs (DValue) la liste AllDomains

[5] Les mots-clés étant donnés par les auteurs, ils peuvent être absents du résumé ou s'y retrouver sous une forme différente.

Pour illustrer ce premier algorithme, supposons 4 termes extraits d'un texte, soient : *stand* (55 domaines), *stem* (41 domaines), *ramet* (5 domaines), et *stand development* (2 domaines). La contribution au poids d'un domaine par le mot *stand* sera de 0.518 (0.5 + 1/55), tandis que la contribution au poids d'un domaine par le mot *ramet* sera de 0.7 (0.5 + 1/2). Le tableau 3 montre le calcul du poids de divers domaines obtenus à partir de ces 4 différents termes en fonction de l'appartenance de chaque terme aux divers domaines.

	stand	stem	ramet	stand development	total
sylviculture	0.518	0.524	0.7	1.0	2.742
agriculture	0.518		0.7	1.0	2.218
forestry	0.518	0.524	0.7		1.742
botany		0.524	0.7		1.224
electric mixers	0.518	0.524			1.042
arboriculture			0.7		0.7
sea transport		0.524			0.524
dentistry	0.518				0.518

Tableau 3 – Exemples de contribution des domaines au profil

Cet algorithme dépend du nombre de termes extraits, et les résultats seront présentés à la section 4 en fonction de cette variable. L'algorithme de recherche de collocations fréquentes (décrite dans Smadja 1993) est utilisé pour l'extraction de termes. Le tableau 4 montre quelques exemples de termes extraits et de profils de domaines obtenus par l'algorithme ci-haut. Il faut préciser que plusieurs des termes extraits ne se retrouveront pas dans le GDT et ainsi ne pourront pas contribuer à la détermination du profil du domaine.

Termes extraits du contexte	Profil de domaines (top 5)
sedimentation, sediment, pH, Welland river, solid concentration	geology (1.69), paper industry (1.69), hydrology (1.12), water (1.12), bottoms and shores (1.1)
excavation, strain sections, plain strain, finite element, three-dimensional finite	physics (1.0), water (1.0), foundations (0.62), construction (0.62), mining industry (0.62)
hydraulic conductivity, sds, naphthalene, NaCl, flushing	chemistry (1.28), water (1.23), plastic additives (0.75), plastics (0.75), organic chemistry (0.75)

Tableau 4 – Exemples de profils de domaine

Les étapes d'extraction de termes et d'assignation de domaine ne sont pas évaluées en soi car elles nécessiteraient des évaluations humaines (subjectives et coûteuses). Nous préférons les évaluer indirectement par leur impact sur la tâche finale d'assignation d'un équivalent pertinent à un terme, tâche que nous pourrons évaluer automatiquement. Cette évaluation indirecte est aussi proposée par Gliozzo et al. 2004 auquel nous avons fait référence précédemment. En effet, bien que nous puissions faire un parallèle entre la tâche d'assignation de domaine à un texte et l'ensemble (très vaste) de travaux en catégorisation de documents, la plupart de ces travaux utilisent des approches de type apprentissage-machine supervisé, ayant accès à des ensembles de textes pré-assignés à des catégories, mais nos travaux sont plutôt de type non-supervisé n'ayant pas accès à des textes étiquetés et ne permettant pas d'évaluation directe. Ainsi, comme Gliozzo et al. 2004, nous ne faisons pas une évaluation directe mais mesurons indirectement le succès par la tâche de désambiguïsation (ici recherche d'équivalent) qui utilise les résultats de l'assignation de domaines.

Une fois le profil du texte déterminé (étape 1), l'algorithme suivant permet de déterminer le domaine approprié au terme.

Etape 2 - Assignation d'une fiche à un terme

1) Soit la liste des domaines provenant du profil du texte (AllDomains)
2) Établir la liste des domaines du terme (TermDomList) en parcourant toutes ses fiches dans le GDT
3) Pour chaque domaine de TermDomList
    a. Si le domaine est présent dans AllDomains
        i. Attribuer à ce domaine une valeur égale à sa valeur AllDomains
    b. Sinon
        i. Retrouver la similarité entre le domaine et chacun des domaines de AllDomains
        ii. Pondérer chacune des similarité par la valeur de chacun des domaines de AllDomains
        iii. Attribuer à ce domaine une valeur égale à la somme pondérée des similarité
4) Ordonnancer les valeurs des domaines dans TermDomList
5) Attribuer au terme le domaine ayant la valeur maximale
6) Récupérer la (ou les) fiches ayant ce domaine

Afin d'être utilisée dans cette étape 2, la similarité entre deux domaines du GDT est pré-calculée sur le GDT entier pour tous ses domaines en fonction des termes que les domaines ont en commun, en utilisant la formule du coefficient Dice, soit : NbTermesEnCommun entre Domaine1 et Domaine 2 / NbTermes du Domaine1 + NbTermes du Domaine 2. Cette approche permettra d'établir automatique que le domaine psychology est plus près des domaines medecine (0.19), statistics (0.13), sociology (0.11), mathematics (0.01) et que des domaines photography (0.04), graphic design (0.03) ou finance (0.03). La ressource terminologique ne contenant ni définition, ni structure (les fiches ne sont pas liées entre elles dans une taxonomie), nous n'avons point d'autre information que le recoupement de termes pour établir les distances entre les domaines. Le GDT possède une organisation hiérarchique de ses domaines en 2 niveaux (domaines généraux et domaines spécifiques), mais l'utilisation de cette structure limiterait la transportabilité de notre méthode.

Pour illustrer l'étape 2, nous montrons au Tableau 5, à la première rangée, un profil de domaines. Ensuite chaque rangée subséquente montre la similarité entre un domaine possible pour un terme (aeronautics, agriculture, finance) et chacun des domaines du profil. Par exemple la similarité entre aeronautics et tobacco est de 0.0029. Ainsi, pour aeronautics, nous aurons une pondération de 0.0053*3.266 + 0.0029*2.55 + 0.0058*2.27 + 0.0045*1.57 + 0.0065*1.57 = 0.055, ce qui le rend moins probable que le domaine agriculture qui a obtenu un total de 3.396.

Profil de domaines pour un texte	agriculture (3.266)	tobacco (2.55)	forestry (2.27)	graphic design (1.57)	printing (1.57)	TOTAL
aeronautics	0.0053	0.0029	0.0058	0.0045	0.0065	0.055
agriculture	1.0	0.012	0.032	0.0078	0.0095	3.396
finance	0.0032	0.0029	0.0048	0.0039	0.0055	0.045

Tableau 5 – Exemple de calcul pour la pondération de 3 domaines

4 Analyse des résultats

Notre algorithme d'assignation d'équivalent est évalué sur l'ensemble de données décrites à la section 2. Tel que mentionné précédemment, nous ne pouvons en fait mesurer automatiquement si notre algorithme choisit la bonne fiche pour un terme car notre ensemble n'est pas étiqueté de cette façon. Nous pouvons plutôt approximer cette désambiguïsation par

l'attribution d'un bon (ou non) équivalent à un terme. Il s'agit d'une approximation car tel qu'indiqué à la section 1 (voir Tableau 1), un terme pourrait avoir plusieurs fiches mais un seul équivalent ou encore plusieurs équivalents et une seule fiche. Tout de même, pour faire une évaluation juste, nous comparons le rang moyen de l'équivalent correct tel qu'obtenu par notre algorithme au rang moyen de l'équivalent correct s'il était choisi au hasard.

Le tableau 6 montre un exemple de résultat pour une paire de termes provenant d'un article scientifique, soit le terme *layering* et son équivalent *marcottage* en contexte. Le tableau montre les termes extraits du contexte, le profil de domaines obtenu (résultat de l'algorithme 1) et les équivalents ordonnés (résultat de l'algorithme 2). Comme *marcottage* se retrouve en première place, il obtient le rang 1, et ce en comparaison avec le rang 5 si cet équivalent était choisi au hasard parmi les 10 candidats possibles.

Paire de termes	layering - marcottage
Équivalents possibles dans le GDT	couchage, lité, marcottage , layering, pistes multicouches, dispersion, organisation en couches, superposition, feuilletage, vibration des strates
Contexte	... Clonal fragments in clearcut stands were large, predating the year of stand establishment, with many dead, old ramets, but many young stems. Ramet recruitment and lateral meristem formation were highest in clearcut stands, which contributed to replacement of older ramets lost to the disturbance. Clonal fragments in young stands were few and small, consisting of a few ramets and short decumbent stems. In maturing stands, clonal fragments were numerous but consisted of few ramets with extensive decumbent stem connections. No devil's club seedlings were observed in any of the stands sampled. Devil's club populations are maintained by prolific basal stem sprouting following disturbance and continual layering and clonal fragmentation throughout stand development.
Termes extraits	clonal fragments, stands, stand development, lateral meristem formation, ramets, stem, young stands, ramet recruitment, meristem formation, ...
Profil de domaines estimé	agriculture 3.27, tobacco 2.55, forestry 2.27, sylviculture 2.22, botany 1.75, graphic design 1.57, marine activities 1.57, household equipment 1.57, electric mixers 1.57, printing 1.57
Équivalents (en ordre décroissant de ressemblance de leur domaine au profil)	marcottage (agriculture : 0.682), marcottage (sylviculture : 0.443), lité (handling and storage : 0.021), superposition (art : 0.014), feuilletage (dairy industry : 0.009), organisation en couches (computer science : 0.008), vibration des strates (aeronautics : 0.008), pistes multicouches (audiovisual : 0.007), layering (finance : 0.006), superposition (music : 0.0), couchage (arboriculture : 0.0), pistes multicouches (digital post-production : 0.0)
Rang obtenu	1
Rang au hasard	5 (soit 10 équivalents / 2)

Tableau 6 – Exemple de résultat de notre algorithme

Revue Scientifique	Nb Paires	Rang si choix au hasard	5 termes	10 termes	15 termes	20 termes
Biochemistry and cell biology	59	6.91	1.41	1.39	1.36	1.31
Canadian Geotechnical Journal	220	9.12	2.52	2.46	2.44	2.45
Canadian Journal of Botany	167	7.56	2.41	2.14	2.16	2.22
Canadian journal of chemistry	118	6.88	1.80	1.73	1.79	1.74
Canadian journal of civil engineering	255	10.32	2.55	2.53	2.59	2.59
Canadian journal of microbiology	116	6.66	1.54	1.57	1.49	1.52
Canadian journal of physiology and pharmacology	125	6.48	1.72	1.55	1.59	1.64
Genome	70	7.02	1.57	1.58	1.63	1.67
MOYENNE - TOTAL	1130	8.13	2.12	2.05	2.06	2.07

Tableau 7 – Résultats détaillés par revue et globaux

Le tableau 7 montre les résultats agglomérés sur l'ensemble des termes. Les colonnes (de gauche à droite) indiquent : la revue scientifique, le nombre de paires d'équivalents tirés de cette revue, le rang moyen où se trouverait le bon équivalent si on prenait cet équivalent au hasard, et différents résultats du rang obtenu par nos algorithmes en fonction d'avoir extrait 5, 10, 15 ou 20 termes du contexte entourant le terme à l'étude. L'amélioration est tout de même marquante avec un rang moyen de 8 pour un choix au hasard et un rang moyen d'environ 2 pour un choix fait avec l'assignation automatique de domaine. Notre méthode fait une réduction de 75% du rang moyen (position du bon équivalent en supposant un accès linéaire). Bien que nous ayons mentionné à la section 3 une variation possible des résultats en fonction du nombre de termes extraits, le tableau 7 montre que ce n'est pas vraiment le cas.

5 Conclusions, discussion et travail futur

Le présent travail propose un algorithme en 2 étapes pour la recherche en base terminologique d'équivalents pertinents à des termes en contexte. Sur un ensemble de 1130 paires de termes équivalents, notre algorithme permet de réduire le rang moyen de l'équivalent pertinent de 75%. Par exemple, dans le cas d'un terme ayant 16 équivalents inscrits dans la banque de terminologie, le traducteur arriverait plus certainement au bon choix après avoir consulté 2 équivalents possibles plutôt que 8.

Bien que ce gain soit déjà important, diverses pistes pourront être explorées pour tenter d'améliorer les résultats obtenus. Par exemple, l'étendue du contexte pourrait être modifiée pour en mesurer l'impact, soit prendre une seule phrase, 2 phrases, le résumé en entier, voire même l'article complet comme contexte. Nous pouvons déjà anticiper que dans un contexte plus large, le choix de l'extracteur de termes et le nombre de termes extraits auront un plus grand impact. Une méthode pondérant le poids des termes extraits en fonction de leur distance au terme à l'étude serait intuitivement une bonne approche, mais cela reste à valider.

Plusieurs explorations seraient possibles pour la sous-tâche de similarité des domaines. Par exemple, plutôt que d'utiliser une mesure de recoupement des listes de termes à l'intérieur du GDT, une utilisation de ressources externes, voire même le WWW nous permettrait de (a) établir la similarité des domaines en fonction de leur information mutuelle tel que donnée par leur PMI-IR (Turney 2001), ou (b) augmenter le nombre de termes faisant partie de chaque domaine en utilisant des approches de types « topic signatures » (Agirre et al. 2001) qui recherchent des termes reliés à partir de définitions (ou dans notre cas de termes connus), permettant par la suite d'utiliser nos mesures de similarité de domaines caractérisés par de plus grands ensembles de termes.

Une limite de notre algorithme est de ne pouvoir différencier les équivalents provenant du même domaine. Il est possible que des notions rapprochées utilisent le même terme dans une langue mais soient différenciées dans une autre langue. Le terme *stem* est un excellent exemple car il possède 14 fiches dans le domaine de la botanique avec des équivalents variés en français tels *tige*, *perche*, *pied* ou *pédoncule* qui dépendent du type de plante ou fleur, ce que l'anglais ne différencie pas. L'état présent de notre approche ne permet pas cette différentiation car elle repose entièrement sur l'assignation du domaine qui lui ici est unique, soit botanique. Uniquement si la banque de terminologie incluait des définitions serait-il alors possible de développer de nouveaux algorithmes utilisant l'information provenant des définitions pour les combiner à nos algorithmes utilisant les domaines.

Les algorithmes développés ont l'avantage d'être transportables à d'autres banques de terminologie pour autant que celles-ci utilisent une catégorisation par domaine, ce que nous pouvons anticiper car cela est fondamental à la structure d'une base terminologique.

Du point de vue applicatif, notre travail futur sera la création d'une interface d'accès à une ressource terminologique qui présentera en ordre décroissant de pertinence les fiches associées à un terme. Cette interface pourra être mise à l'essai dans un contexte d'aide au traducteur pour la recherche d'équivalents appropriés à des termes provenant de textes à traduire.

Références

AGIRRE E., MARQUEZ L., WICENTOWSKI R. (2007) *Proceedings of the 4th International Workshop on Semantic Evaluations.*

AGIRRE E., ANSA O., MARTINEZ, D., HOVY E. (2001) Enriching WordNet concepts with topic signatures. *WordNet and Other Lexical Resources Workshop, NAACL 2001.*

BARRIÈRE, C. (2007). La désambiguïsation du sens en traitement automatique des langues : l'apport des ressources terminologiques et lexicographiques, dans L'Homme M.-C. et Vandaele S. (eds), *Lexicographie et Terminologie : Compatibilité des modèles et des méthodes*, 113-140.

GLIOZZO A., MAGNINI B., STRAPPARAVA C. (2004) Unsupervised Domain Relevance Estimation for Word Sense Disambiguation, *Proceedings of EMNLP 2004,* D. Lin and D. Wu (eds), Barcelona, Spain, pages 380-387, 2004.

LESK M. (1986) Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from and Ice Cream Cone, *ACM SIGDOC'86, The Fifth International Conference on Systems Documentation, Proceedings of ACM Press.*

MAGNINI B., CAVAGLIÀ G. (2000) Integrating subject field codes into WordNet. *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation,* Athens, Greece, June 2000.

MIHALCEA R., EDMONDS P. (2004) *Proceeding of Senseval-3: The third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.*

SEBASTIANI F. (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1): 1-47, 2002.

SMADJA F. (1993) Retrieving collocations from text: Xtract. *Computational Linguistics* 7(4): 143-177.

SUAREZ A., PALOMAR M. (2006) Word Sense vs. Word Domain Disambiguation: A Maximum Entropy approach, *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL,* pages 141-144, New York, June 2006.

TURNEY P.D. (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), Freiburg, Germany, pp. 491-502.