Lexiques de corpus comparables et recherche d'information multilingue

Frederik Cailliau[1] Ariane Cavet[1] Clément de Groc[2,3] Claude de Loupy[2]
(1) Sinequa, 12 rue d'Athènes, 75009 Paris
(2) Syllabs, 53 bis rue Sedaine, 75011 Paris
(3) LIMSI-CNRS, BP 133, 91403 Orsay CEDEX
{cailliau,cavet}@sinequa.com, {cdegroc,loupy}@syllabs.com

RÉSUMÉ

Nous évaluons l'utilité de trois lexiques bilingues dans un cadre de recherche interlingue français vers anglais sur le corpus CLEF. Le premier correspond à un dictionnaire qui couvre le corpus, alors que les deux autres ont été construits automatiquement à partir des sous-ensembles français et anglais de CLEF, en les considérant comme des corpus comparables. L'un contient des mots simples, alors que le deuxième ne contient que des termes complexes. Les lexiques sont intégrés dans des interfaces différentes dont les performances de recherche interlingue sont évaluées par 5 utilisateurs sur 15 thèmes de recherche CLEF. Les meilleurs résultats sont obtenus en intégrant le lexique de mots simples généré à partir des corpus comparables dans une interface proposant les cinq « meilleures » traductions pour chaque mot de la requête.

ABSTRACT

Lexicons from Comparable Corpora for Multilingual Information Retrieval

We evaluate the utility of three bilingual lexicons for English-to-French crosslingual search on the CLEF corpus. The first one is a kind of dictionary whose content covers the corpus. The other two have been automatically built on the French and English subparts of the CLEF corpus, by considering them as comparable corpora. One is made of simple words, the other one of complex words. The lexicons are integrated in different interfaces whose crosslingual search performances are evaluated by 5 users on 15 topics of CLEF. The best results are given with the interface having the simple-words lexicon generated on comparable corpora and proposing 5 translations for each query term.

MOTS-CLÉS : recherche d'information multilingue, corpus comparables, lexiques multilingues
KEYWORDS : multilingual information retrieval, comparable corpora, multilingual lexicons

1 La recherche multilingue

La recherche d'information multilingue (CLIR, Cross-Language Information Retrieval) consiste à trouver des documents pertinents dans une collection multilingue à partir de requêtes formulées dans une seule langue. Trois approches permettent de faire de la recherche multilingue : la traduction de la requête, la traduction des documents, ou bien la combinaison des deux. Nous nous sommes concentrés sur la traduction de la requête.

Dans cet article nous mettons à l'épreuve trois lexiques bilingues dans un contexte de recherche d'information interlingue : trouver des documents pertinents en anglais à partir de requêtes posées en français en utilisant le corpus CLEF 2000-2002. Deux des

lexiques, l'un avec des termes simples, l'autre avec des termes complexes, ont été construits à partir de corpus comparables (Déjean et Gaussier, 2002) prenant comme source les sous-ensembles français et anglais de CLEF. Ce corpus est constitué d'articles journalistiques parus en 1994. Nous indiquons d'abord quelques travaux liés, présentons ensuite le prototype construit pour l'évaluation, les principes de l'évaluation menée et les résultats.

2 Lexiques de traduction et mise en correspondance

Trois lexiques de traduction français vers anglais sont utilisés. Le premier lexique, appelé *GT*, a été construit en soumettant tous les mots du corpus CLEF 2003-2004 (qui est une extension des corpus CLEF 2000-2002) au dictionnaire en ligne Google Dictionary. Il contient 27 446 entrées totalisant 73 027 traductions en anglais qui correspondent à 32 298 mots uniques.

Le deuxième lexique, appelé *A*, a été construit selon la méthode décrite dans (Li et Gaussier, 2010). Dans cet article, les auteurs définissent une mesure de comparabilité et une stratégie permettant d'améliorer itérativement la qualité d'un corpus comparable (CLEF 2003-2004, sans la partie CLEF 2000-2002) en intégrant une sélection de documents issus d'un second corpus (Wikipedia). Enfin, une approche standard pour l'extraction de terminologies bilingues (Fung et Yee, 1998) appliquée à ce corpus enrichi permet d'extraire les 1 000 traductions les plus probables pour les 947 mots composant les thèmes de recherche du corpus CLEF[1]. Chaque traduction est alors munie d'un score, ce qui nous permet de présenter les n meilleures traductions à l'utilisateur (dans notre cas les 5 ou 10 premières comme nous verrons plus tard).

Le troisième lexique, appelé *MT*, a été construit avec la méthode détaillée dans (Morin et Daille, 2010). Il résulte d'une extraction terminologique de termes complexes, puis d'un alignement interlingue dans les corpus comparables CLEF 2000-2002, et contient 64 556 entrées totalisant 68 956 traductions anglaises. Les traductions correspondent à 28 795 mots uniques.

Les entrées des lexiques *GT* et *A* sont des lemmes, tandis que les entrées de MT sont des formes, ce qui explique son nombre d'entrées élevé. En ignorant les accents, la casse, en tenant compte de l'existence de termes complexes et en procédant si nécessaire à une lemmatisation, nous maximisons les chances d'obtenir une correspondance entre les mots de la requête et ceux des lexiques.

3 Interface d'évaluation

Le prototype a été construit sur le moteur de recherche de Sinequa. L'interface d'évaluation utilise l'interface classique du produit composée d'une case de recherche, d'une liste de résultats ordonnés par pertinence et des « facettes » à gauche de la liste des résultats. Les facettes sont des groupes nominaux et des noms de personnes, de lieux et d'entreprises, qui ont été extraits des documents et servent à la navigation.

[1] Pour appliquer cette approche à des thèmes de recherche inconnus, il serait nécessaire de calculer en amont les traductions possibles pour l'ensemble des mots du corpus.

Pour l'évaluation, seuls des documents en anglais ont été indexés. Les utilisateurs ont posé leurs requêtes en français, sauf sur l'interface *baseline*, et la recherche multilingue ne s'active qu'après une première requête. Elle propose des traductions pour chaque mot du lexique de traduction sélectionné. Dans notre expérience, l'utilisateur est obligé de cocher les traductions voulues pour trouver des documents en anglais. Suivant les travaux de Pirkola *et al.* (2003), entre autres, nous avons structuré la requête pour traiter les traductions comme des synonymes pour le calcul de la pertinence.

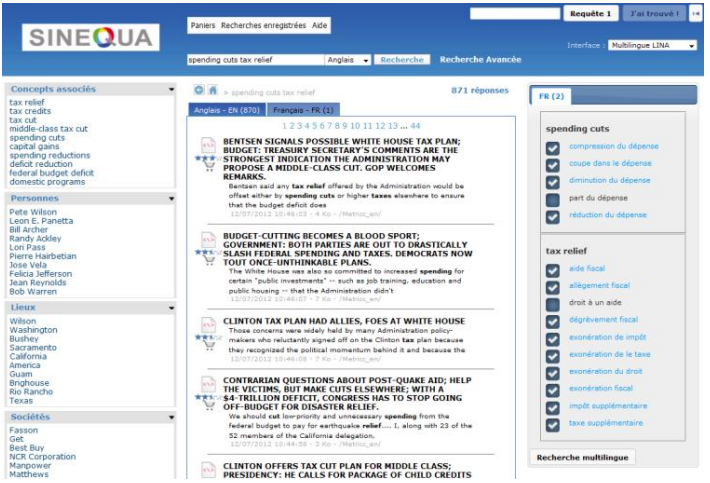

FIGURE 1 – Interface de requêtage multilingue, exemple anglais vers français.

4 Principes d'évaluation

L'expérience présentée ici a pour but d'évaluer l'apport de suggestions de traductions sur le temps de recherche, dans un contexte interlingue. Nous avons pour cela repris les principes des expériences menées par Crestan & Loupy (2004) dans un contexte monolingue. Il a alors été montré que des systèmes d'aide à la recherche peuvent considérablement améliorer l'efficacité des utilisateurs pour une recherche monolingue.

Les 5 interfaces testées sont décrites dans la table 1.

Nº	Nom	Description
1	B	Baseline : aucun lexique, l'utilisateur formule ses requêtes en langue cible
2	GT	Lexique de mots simples traduits issus d'un dictionnaire en ligne
3	A5	Lexique des 5 meilleures traductions issues des corpus comparables
4	A10	Lexique des 10 meilleures traductions issues des corpus comparables
5	MT	Lexique n'ayant que des termes complexes issus du corpus comparable

TABLE 1 – Description des interfaces.

Ces 5 interfaces sont testées par 5 utilisateurs devant trouver des documents pertinents dans la langue cible en effectuant des recherches sur 15 thèmes. Nous avons réparti ces thèmes de recherche en 5 groupes de 3, nommés G1 à G5 et les avons distribués sur les couples (interface, utilisateur) comme indiqué dans la table 2.

Les 15 thèmes de recherche sont ainsi testés sur chaque interface, tout en garantissant qu'aucun utilisateur ne traite deux fois un même thème. Chaque utilisateur traite l'ensemble des thèmes et teste toutes les interfaces.

	U1	**U2**	**U3**	**U4**	**U5**
B	G1	G5	G4	G3	G2
A5	G2	G1	G5	G4	G3
A10	G3	G2	G1	G5	G4
GT	G4	G3	G2	G1	G5
MT	G5	G4	G3	G2	G1

TABLE 2 – Répartition des thèmes de recherche sur les couples (interface, utilisateur).

Les éléments évalués sont les suivants :

1. le temps mis par les utilisateurs pour accéder au premier document pertinent ;
2. le nombre de documents pertinents visualisés pendant un temps donné ;
3. le nombre de documents non pertinents visualisés pendant le même temps.

Le premier élément mesure l'impact des interfaces sur le temps d'une recherche informationnelle. Plus ce temps est bas, plus l'interface est performante. Le deuxième mesure l'impact sur le temps d'une recherche documentaire. Plus le nombre de documents est élevé, plus l'interface est performante. Le dernier mesure la perturbation causée par de mauvais résultats. Plus ce nombre de documents est bas, plus l'interface est performante.

Nous avons sélectionné les thèmes de recherche parmi les thèmes CLEF disponibles. Afin de permettre une évaluation de l'apport des termes complexes, nous avons choisi aléatoirement 15 thèmes parmi les 36 qui contenaient des termes présents dans le lexique bilingue de termes complexes. En voici les 5 premiers :

1 emeutes pendant un match de football à dublin
2 victimes d'avalanches
3 nouveau premier ministre portugais
4 nouveaux partis politiques
5 dommages à la couche d'ozone

Aucune autre indication n'était fournie aux utilisateurs qui devaient trouver des documents pertinents en y passant exactement 5 min.

5 Résultats

Les résultats bruts donnés dans cette section seront interprétés dans la section suivante. Les meilleurs résultats sont indiqués en gras tandis que l'aide sur laquelle nous orienterons l'argumentation est indiquée en rouge.

5.1 Temps pour le premier document pertinent

La table 3 présente les temps d'accès minimum, maximum et moyen au premier document pertinent selon les interfaces, où on voit que la *baseline* offre les temps d'accès minimum et moyen les plus faibles.

	B	GT	A5	A10	MT
Min	**23**	41	32	26	29
Max	229	246	234	300	**194**
Moy	**87,8**	135,8	99,2	123,8	102,4

TABLE 3 – Temps d'accès au premier document pertinent

5.2 Nombre de documents pertinents retrouvés

La table 4 donne le nombre moyen de documents pertinents trouvés selon les différentes interfaces. L'interface *A5*, proposant 5 candidats termes simples en traduction à la requête, obtient des résultats nettement supérieurs aux autres interfaces (52% de plus que la *baseline*). On peut également constater cela sur la courbe temporelle. On constate cependant que, si *A5* est effectivement l'interface produisant le plus de documents à terme, elle est largement distancée au début de la recherche par la *baseline*.

Interface	Pertinents
B	25
GT	23
A5	38
A10	26
MT	27

TABLE 4 – Nombre de documents pertinents récupérés par interface.

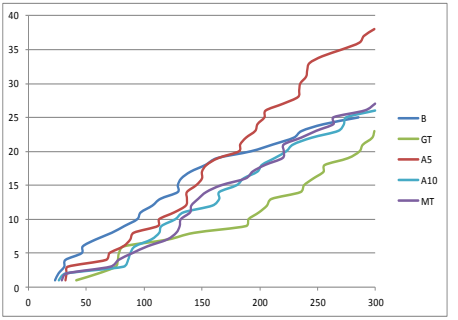

FIGURE 2 - Courbe temporelle (*s*) des documents pertinents retrouvés.

5.3 Nombre de documents non pertinents visualisés

La table 5 donne le nombre de documents non pertinents visualisés, à côté du nombre de documents pertinents trouvés.

Interface	Pertinents	Non pertinents
B	25	52
GT	23	31
A5	38	26
A10	26	33
MT	27	28

TABLE 5 - Nombre de documents non pertinents visualisés.

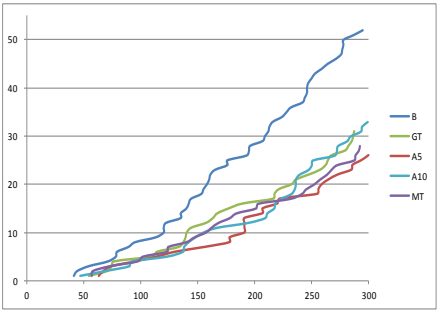

FIGURE 3 – Courbe temporelle (*s*) des documents non pertinents visualisés.

On constate que l'interface *A5* est la meilleure. La *baseline* conduit à visualiser 2 fois plus de documents non pertinents. La courbe temporelle est très marquée par la différence entre la *baseline* et les autres interfaces.

5.4 Précision moyenne

La table 6 présente la précision moyenne à 5 min. Sur les courbes, on voit une nette dominance de l'interface *A5* dès le début de la 2ème min. L'interface *MT*, qui ne propose que des termes complexes, arrive en 2ème position au bout des 5 min. On constate aussi que l'aide apportée par *GT* surpasse l'absence d'aide au bout des 5 min.

Type d'aide	Précision moyenne
B	0,2
GT	0,21
A5	0,42
A10	0,27
MT	0,29

TABLE 6 – Précision moyenne à 5 min.

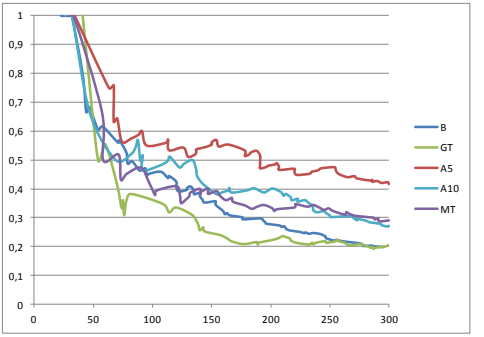

FIGURE 4 – Précision moyenne au cours du temps (*s*).

5.5 Nombre d'opérations effectuées

Le nombre d'opérations effectuées est une notion de moindre importance mais qui montre la facilité d'utilisation d'une interface. La table 7 indique le nombre moyen (ainsi que minimal et maximal) de requêtes posées. On y voit que le nombre de recherches minimal correspond à la *baseline B* mais que l'interface *A5* en est proche.

	B	**GT**	**A5**	**A10**	**MT**
Min	**1**	2	2	**1**	2
Max	**8**	17	**8**	21	18
Moy	**3,3**	4,5	3,9	5,1	6,9

TABLE 7 – Nombre moyen de requêtes effectuées pour un même thème.

5.6 Comparaison des utilisateurs

La table 8 montre le nombre de documents pertinents trouvés en 5 min. va de 0 à 8 documents selon les thèmes. Un même document peut être validé par *n* utilisateurs.

Aucun utilisateur n'a trouvé de document pertinent pour le thème 14 (*démission du secrétaire général de l'otan*). Pour 9 thèmes sur 15, il y a au moins un couple (utilisateur, interface) qui n'a retrouvé aucun document pertinent. Les thèmes sont donc difficiles.

Req.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
U1	2	0	1	6	7	0	5	1	0	2	0	3	4	0	0	31
U2	1	3	0	3	6	2	4	2	0	0	0	1	1	0	0	23
U3	0	2	0	1	2	0	3	1	0	0	0	2	1	0	0	12
U4	1	2	2	6	8	2	7	2	0	3	5	4	2	0	1	45
U5	0	2	0	3	6	1	5	2	1	1	0	3	4	0	0	28
Total	4	9	3	19	29	5	24	8	1	6	5	13	12	0	1	139

TABLE 8 – Nombre de documents pertinents trouvés par thème, par utilisateur.

Dans la table 9 nous présentons le nombre de documents pertinents trouvés selon les utilisateurs et les interfaces. La grande disparité entre les utilisateurs est sans doute due au fait qu'aucune instruction ne leur a été donnée quant à l'évaluation de pertinence.

On peut calculer le rappel si on considère que le nombre maximal de documents pertinents trouvés par un utilisateur pour un thème est la référence de ce qu'il fallait trouver. Le rappel par couple (utilisateur, interface) permet de retrouver les interfaces les plus pertinentes pour chaque utilisateur. Les résultats sont présentés dans la table 10, où on voit que 3 utilisateurs sur 5 ont été plus performants avec l'interface *A5*. 2 utilisateurs sur 5 (dont 1 à égalité avec *A5*) ont été plus performants avec l'interface *MT* alors que celle-ci ne propose que des relations entre termes complexes.

	U1	U2	U3	U4	U5	Total
B	3	1	2	9	10	**25**
GT	5	6	3	5	4	**23**
A5	13	4	1	12	8	**38**
A10	6	11	2	3	4	**26**
MT	4	1	4	16	2	**27**
Total	**31**	**23**	**12**	**45**	**28**	

TABLE 9 - Documents pertinents par interface et utilisateur.

	B	A5	A10	GT	MT
U1	50%	**63%**	40%	47%	50%
U2	13%	50%	**75%**	52%	8%
U3	17%	13%	22%	14%	**31%**
U4	67%	**100%**	75%	72%	**100%**
U5	58%	**90%**	36%	50%	22%
	204%	**315%**	249%	236%	212%

TABLE 10 - Interface la plus performante par utilisateur.

6 Analyse des résultats

Globalement, nous constatons que la présence d'une aide à la traduction ralentit l'accès au premier document pertinent. Cela est parfaitement logique car, dans le cas où une aide est proposée, l'utilisateur commence par regarder les traductions proposées par l'interface et sélectionne celles qui lui paraissent pertinentes avant de lancer la requête. Lorsqu'il n'y a pas d'aide, les utilisateurs lancent directement la requête en anglais.

En revanche, on peut voir sur la figure 2 que l'interface *A5* donne de meilleurs résultats après 2,5 min en moyenne et que les autres interfaces ont une pente qui devrait les amener à dépasser la *baseline* après 5 min. Dès le départ, beaucoup de documents pertinents et non pertinents sont visualisés par l'interface *baseline*. Cela se voit dans les courbes de précision moyenne selon le temps. La dominance d'*A5* est claire dès la 2e min.

Parmi toutes les interfaces proposant de l'aide à l'utilisateur pour la traduction des requêtes, l'*A5* semble de loin la meilleure. En particulier, l'aide provenant du

dictionnaire en ligne est peu performante en termes de documents pertinents retrouvés. Notons cependant que la pente de la courbe en fin de temps est très fortement croissante et il est possible que de meilleurs résultats soient apparus si l'expérience avait été prolongée au-delà de 5 min. Mais il semble peu probable que cette courbe puisse rejoindre celle d'*A5*. Notons aussi la mauvaise performance de *A10* par rapport à *A5* qui ne peut s'expliquer que par des contraintes ergonomiques, comme un effort de lecture et sélection des traductions trop intensif. Enfin, si l'on considère la performance par utilisateur, on se rend compte que l'interface *A5* est la plus performante pour 3 utilisateurs sur 5. Les résultats obtenus par MT sont étonnamment bons compte tenu du fait qu'il ne propose de traductions que si la requête posée contient des mots composés contenus dans ce lexique.

7 Conclusion

Notre évaluation montre que la meilleure interface intègre le lexique généré à partir des corpus comparables et propose 5 traductions pour chaque mot de la requête. Le résultat est surprenant, car ce lexique n'a pas été généré sur les données du corpus d'évaluation et n'avait pas été ressenti comme un lexique de bonne qualité linguistique. Néanmoins, les corpus CLEF 2000-2002 et 2003-2004 étant très semblables, ce lexique correspond mieux au corpus d'évaluation que le lexique générique issu du dictionnaire en ligne. Ce résultat, ainsi que la bonne performance des termes complexes démontrent l'efficacité des lexiques construits à partir de corpus comparables pour la recherche d'information multilingue interactive. Il serait intéressant de confirmer ces résultats à l'aide d'une évaluation de plus grande ampleur.

Remerciements

Ces travaux ont été exécutés dans le cadre du projet ANR MéTRICC (ANR-08-CONT-013). Nous remercions le LIG et le LINA de nous avoir fourni les lexiques de traduction.

Références

CRESTAN, E., LOUPY, C. de (2004). Browsing Help for a Faster Retrieval. In *Proceedings of COLING 2004*. Genève, Suisse, pages 576-582.

DÉJEAN H., GAUSSIER, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. In *Lexicometrica*, n° spécial 2002, 22 pages.

FUNG, P., YEE, L.Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of COLING-ACL 1998*. Montreal, pages 414-420.

LI, B., GAUSSIER, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proc. of COLING 2010*. ACL, pages 644-652.

MORIN, E., DAILLE, B. (2010). Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation* (LRE). Vol. 44, 1-2. Springer, pages 79-95.

PIRKOLA, A., PUOLAMÄKI, D., JÄRVELIN, K. (2003). Applying query structuring in cross-language retrieval. In *Inf. Process. Manage.* 39, 3 (May 2003), pages 391-402.