

Validation de la notion de similarité textuelle dans un cadre multilingue

Romaric Besançon¹, Martin Rajman¹

¹ EPFL – DI-LIA – INR (Ecublens) – CH-1015 Lausanne – Suisse

Abstract

We propose in this contribution a method to validate the relevance of the notion of vector based semantic similarity, using a multilingual framework. The goal is to verify that vector based semantic similarities can be reliably transferred from one language to another. More precisely, the idea is to test whether the relative positions of documents in a vector space associated with a given source language are close to the ones of their translations in the vector space associated with the target language. The experiments have been conducted using the standard Vector Space model and the DSIR model. The method used for this test can also be applied to the automated evaluation of the performance of machine translation systems or to multilingual information retrieval.

Résumé

Dans cet article, nous proposons une approche permettant d'utiliser un corpus multilingue aligné pour la validation de la pertinence de la notion de similarité sémantique à base de représentation vectorielle des documents. L'idée est de vérifier, à l'aide de tests statistiques, que les positionnements relatifs des documents dans l'espace vectoriel associé à la langue source sont proches de ceux de leur traduction dans l'espace vectoriel associé à la langue cible. Les expériences présentées ici reposent sur deux modèles de représentation vectorielle des documents : le modèle vectoriel standard et le modèle DSIR. La méthodologie développée pour cette validation peut également être appliquée à l'évaluation automatique de la performance de systèmes de traduction automatique ou dans le cadre de la recherche documentaire multilingue.

Keywords: Représentation vectorielle, similarité textuelle, approche multilingue

1. Introduction

La notion de similarité textuelle est très souvent utilisée dans les applications de Traitement de la langue destinées à l'exploitation de corpus textuels de grande taille. Par exemple, en Recherche Documentaire, les documents pertinents retournés par le moteur de recherche sont les plus proches de la requête selon une certaine mesure de similarité (Salton and McGill, 1983); de même, dans le cas de la structuration automatique de bases de données textuelles (classification non supervisée), les documents sont également regroupés en classes en fonction d'une mesure de similarité spécifique (Salton et al., 1975a).

La notion de similarité entre documents est par ailleurs fortement liée au choix de la méthode de représentation des textes. Dans le cas de collections de textes de grande taille, la représentation la plus utilisée est la représentation vectorielle (par exemple mise en œuvre dans des systèmes de recherche documentaire tels que SMART (Salton, 1971)), dans laquelle un document est représenté par un vecteur dans un espace vectoriel dont les dimensions sont associées à des unités linguistiques spécifiques (mot, *stems*, lemmes, etc.).

La représentation vectorielle des documents, fondée sur un modèle simple et facile à mettre en

œuvre, a déjà prouvé sa pertinence et son efficacité dans le cadre d'évaluations liées à diverses applications (recherche documentaire, classification, etc.). Dans cet article, nous proposons une nouvelle méthode de validation indépendante de toute application. L'idée est d'utiliser un corpus multilingue aligné pour valider la pertinence de la représentation en montrant que les positionnements relatifs des représentations des documents dans l'espace vectoriel associé à une langue source sont proches de ceux de leurs traductions dans l'espace vectoriel associé à la langue cible. En d'autres termes, on cherche à vérifier que la similarité entre les représentations vectorielles de deux documents quelconques dans l'espace source est proche de la similarité entre les représentations vectorielles de leur traduction dans l'espace cible.

La suite de cette contribution s'articule de la façon suivante : dans la section 2, nous présentons les deux modèles de représentation vectorielle utilisées pour nos expériences, à savoir le modèle vectoriel standard et le modèle DSIR qui intègre plus d'information sémantique au moyen de co-occurrences ; la méthode utilisée pour effectuer les tests de validation de la notion de similarité est décrite dans la section 3 ; les données utilisées pour les tests, ainsi que les résultats obtenus, sont présentés dans la section 4 ; finalement, dans la section 5, nous décrivons une des applications possibles de la méthode utilisée, l'évaluation automatique des performances des systèmes de traduction automatique.

2. Représentation vectorielle des documents

Dans les deux modèles vectoriels considérés (le modèle vectoriel standard et le modèle DSIR), l'espace de représentation des documents est un espace vectoriel dont chaque dimension est associée à une unité linguistique particulière, appelée ci-après *terme d'indexation*. Nous détaillons dans les sous-sections suivantes les choix qui ont été faits dans nos expériences pour ce qui est de la définition des termes d'indexation et de la construction des représentations dans les deux modèles.

2.1. Choix des termes d'indexation

Une série de pré-traitements automatiques a été mise en œuvre pour extraire l'ensemble des termes d'indexation, noté T , à partir du corpus bilingue utilisé (ce corpus est décrit plus en détail dans la section 4) :

- le corpus bilingue a d'abord été analysé par un lemmatiseur construit à l'aide de la boîte à outils logicielle SYLEX (Constant, 1995) : chaque mot a ainsi été associé à sa catégorie morpho-syntaxique et à son lemme ;
- deux lexiques ont ensuite été extraits (un pour chaque langue), contenant les lemmes des mots pleins (noms, verbes et adjectifs) ;
- enfin, un filtrage fréquentiel des termes d'indexation a été effectué sur la base de leur *fréquence en documents* (*i.e.* le nombre de documents différents contenant un terme donné) ; plus précisément, l'ensemble d'indexation T retenu a été l'ensemble des éléments des lexiques dont la fréquence en documents est comprise dans l'intervalle de fréquence en documents $\left[\frac{|C|}{100}, \frac{|C|}{10} \right]$, où $|C|$ est le nombre de documents dans le corpus. Cet intervalle est usuellement considéré comme adéquat pour fournir des termes avec un bon pouvoir discriminant (Salton et al., 1975b).

2.2. Représentation des documents dans le cadre du modèle vectoriel standard

Dans le cadre du modèle vectoriel standard (VS), chaque document d est représenté au moyen d'un vecteur $d^{VS} = (d_1^{VS}, \dots, d_{|T|}^{VS})$, appelé *profil lexical*, dans lequel la j^e composante d_j^{VS} représente le poids (ou importance), dans le document d , du terme d'indexation t_j associé à la j^e dimension de l'espace vectoriel. D'une façon générale, la mesure d'importance utilisée est le plus souvent une fonction de la fréquence du terme dans le document et intègre de plus une pondération locale, une pondération globale et un facteur de normalisation par rapport à la longueur du document. La fonction choisie pour nos expériences correspond au schéma de pondération *ltn* de SMART (Salton and Buckley, 1988; Singhal, 1997) :

$$d_j^{VS} = w_j = \text{idf}_j \times (1 + \log(tf_j)) \quad (1)$$

où tf_j est la fréquence du terme dans le document et $\text{idf}_j = \log \frac{1}{df_j}$ est le facteur de fréquence en documents inverse, df_j étant la fréquence en documents du terme. Dans cette expression, le facteur de pondération locale est $(1 + \log(tf_j))$ et le facteur de pondération globale est idf_j . Ce dernier facteur permet d'accorder un poids plus important aux termes qui apparaissent moins fréquemment dans la collection et sont donc plus utiles pour la discrimination. Aucun facteur de normalisation n'est intégré directement. Toutefois, une normalisation implicite est effectuée par l'utilisation de la mesure de similarité du cosinus qui est indépendante de la norme.

2.3. Représentation des documents dans le cadre du modèle DSIR

Le modèle DSIR est un modèle vectoriel permettant d'intégrer des informations sémantiques supplémentaires par l'utilisation de co-occurrences (Rajman and Bonnet, 1992; Rajman et al., 2000; Besançon, 2001).

Dans le cadre de ce modèle, les unités linguistiques u_i considérées sont représentées par un vecteur $c_i = (c_{i1}, \dots, c_{i|T|})$, appelé *profil de co-occurrence*, dont chaque composante c_{ij} est la fréquence de co-occurrence de l'unité linguistique u_i avec un terme d'indexation t_j . Un document d est alors représenté comme la somme pondérée des profils de co-occurrence des unités linguistiques qu'il contient, c'est-à-dire par un vecteur $d^{DS} = (d_1^{DS}, \dots, d_{|T|}^{DS})$ où chaque d_j^{DS} est défini par :

$$d_j^{DS} = \sum_{u_i \in d} w_i c_{ij}$$

où la pondération w_i est celle définie par l'équation (1) ci-dessus.

Dans le cadre du modèle DSIR pur, les termes explicitement présents dans le document ne sont donc pris en compte que par le biais de leur profil de co-occurrence. Pour tenir compte plus directement de la présence d'un terme dans un document, un modèle DSIR hybride intégrant à la fois les occurrences et les co-occurrences des termes dans le document a également été proposé (Rungsawang, 1997; Rajman et al., 2000). Dans ce modèle, un document est représenté par un vecteur de composantes d_j^{DS} définies par :

$$d_j^{DS} = \alpha w_j + (1 - \alpha) \sum_{u_i \in d} w_i c_{ij} \quad (2)$$

où α ($0 \leq \alpha \leq 1$) représente le facteur d'hybridation permettant de contrôler l'importance relative dans l'hybridation du modèle DSIR par rapport au modèle VS standard.

3. Méthode de validation

L'idée de la méthode de validation proposée ici pour vérifier la pertinence de la mesure de similarité dérivée d'une représentation vectorielle des documents repose sur l'hypothèse d'invariance suivante : on considère qu'une similarité entre représentations est d'autant plus pertinente que sa valeur calculée pour les représentations de deux documents quelconques dans une langue donnée est proche de celle obtenue pour les représentations des traductions de ces documents dans une autre langue.

Plus précisément, on cherche à tester dans quelle mesure le vecteur des similarités d'un document par rapport à un ensemble de documents de référence dans une langue source est proche du vecteur des similarités de sa traduction par rapport à l'ensemble des traductions des documents de référence dans une langue cible.

À cet effet, considérons un corpus bilingue aligné composé des corpus \mathcal{C} et \mathcal{C}' , respectivement dans les langues L et L' . La procédure de test fonctionne alors de la façon suivante :

- on extrait du corpus aligné d'origine un corpus de test et un corpus de référence ;
- dans chacune des deux langues, on détermine les positionnements des documents du corpus de test par rapport aux documents du corpus de référence (le positionnement d'un document étant caractérisé l'ensemble de ses similarités avec les documents du corpus de référence) ;
- on compare les positionnements ainsi obtenus pour les deux langues.

Construction des corpus de test et de référence On extrait aléatoirement du corpus bilingue disponibles un ensemble de n documents alignés ($n = 500$ dans nos expériences). Ces documents constituent le corpus de test, les $N = |\mathcal{C}| - n$ documents restants formant le corpus de référence.

On note :

- d_i les n documents du corpus de test TEST- L dans la langue source L ;
- d'_i les n documents du corpus de test TEST- L' dans la langue cible L' ;
- D_i les N documents du corpus de référence REF- L dans la langue source L ;
- D'_i les N documents du corpus de référence REF- L' dans la langue cible L' .

La construction des corpus de test et de référence est synthétisée par la figure 1.

Représentation des documents Tous les documents, ceux des corpus de test comme ceux des corpus de référence, sont représentés dans l'espace vectoriel associé à leur langue. On note que deux documents de langue différente ne sont donc pas directement comparables car ils ne sont pas représentés dans le même espace.

À chaque document d_i (resp. d'_i) du corpus de test TEST- L (resp. TEST- L'), on associe alors un vecteur des similarités dont la j^e composante est la similarité de ce document avec le j^e document du corpus de référence REF- L (resp. REF- L').

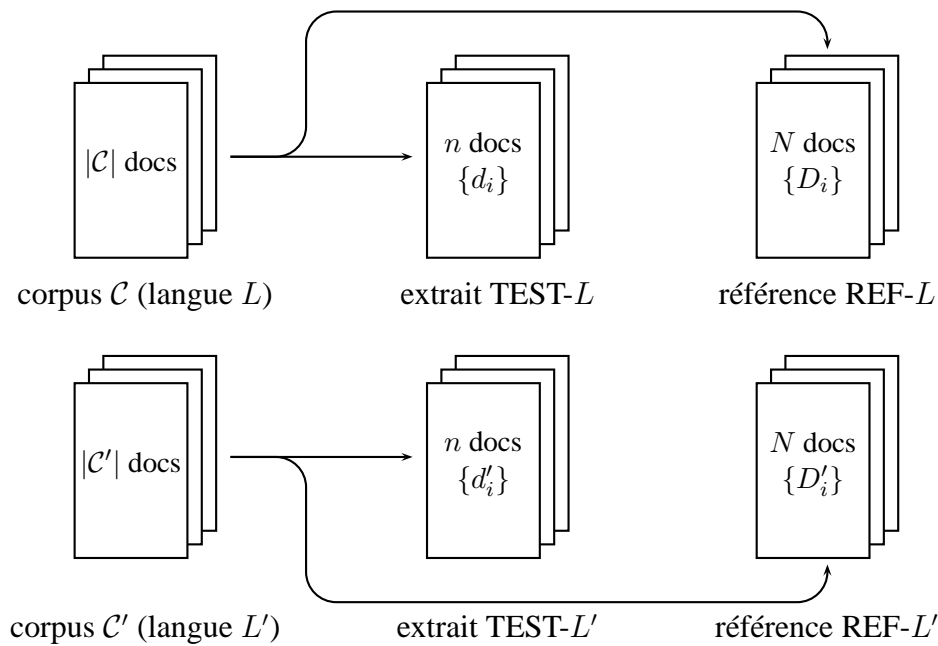


FIG. 1 – *Corpus de test et de référence*

Soit $V_s(d_i)$ (resp. $V_s(d'_i)$) ce vecteur de similarités, on a :

$$V_s(d_i) = (\delta(d_i, D_1), \dots, \delta(d_i, D_N))$$

$$V_s(d'_i) = (\delta(d'_i, D'_1), \dots, \delta(d'_i, D'_N))$$

où δ représente la similarité textuelle utilisée. Pour nos expériences, nous avons choisi la similarité du cosinus, définie, pour deux documents d et d' , par :

$$\delta_{cos}(d, d') = \frac{d \cdot d'}{\|d\| \|d'\|}$$

Le vecteur de similarités caractérise le positionnement relatif du document considéré par rapport aux documents du corpus de référence associé. Comme les corpus de référence sont alignés, les vecteurs de similarités ainsi obtenus sont, par construction, comparables entre eux, indépendamment de la langue. Ils peuvent de ce fait être utilisés pour tester l'invariance par traduction du positionnement d'un document par rapport au corpus de référence.

Notons que l'approche utilisée pour représenter des documents de langues différentes par rapport à un corpus de référence aligné est proche de certaines méthodes proposées pour la recherche documentaire multilingue, comme le modèle vectoriel généralisé (Carbonell et al., 1997; Yang et al., 1998) ou le modèle *Latent Semantic Indexing* (LSI) dans le cadre de la recherche documentaire multilingue (Dumais et al., 1996; Littman and Jiang, 1998).

Test d'invariance Pour chaque document d_i du corpus de test dans la langue L , on s'intéresse désormais à la similarité entre le vecteur de similarités $V_s(d_i)$ qui lui est associé et chacun des vecteurs de similarités associés aux n documents du corpus de test en langue L' . Dans nos expériences, ces similarités entre vecteurs de similarités sont également évaluées par la mesure du cosinus. L'idée, illustrée par la figure 2, est que si l'invariance par traduction est

effectivement vérifiée, alors la similarité entre $V_s(d_i)$ et $V_s(d'_i)$ devrait être significativement plus importante que chacune des similarités entre $V_s(d_i)$ et $V_s(d'_j)$ avec $j \neq i$.

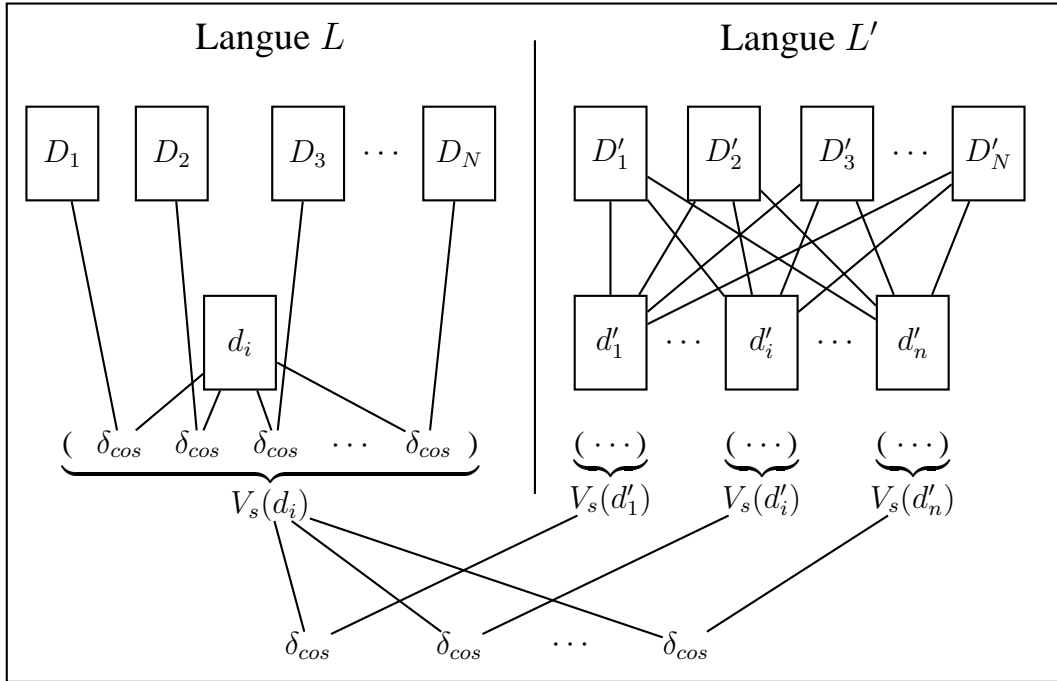


FIG. 2 – Méthode pour la validation

Pour tester cette hypothèse, on compte le nombre de documents d'_j ($j \neq i$) dans TEST- L' pour lesquels la similarité $\delta_{cos}(V_s(d_i), V_s(d'_j))$ est plus petite que la similarité $\delta_{cos}(V_s(d_i), V_s(d'_i))$.

Soit alors f_i la proportion de ces documents dans TEST- L' , on a :

$$f_i = \frac{1}{n-1} |\{d'_k \in \text{TEST-}L' | k \neq i \text{ et } \delta_{cos}(V_s(d_i), V_s(d'_k)) < \delta_{cos}(V_s(d_i), V_s(d'_i))\}|$$

Pour vérifier que peu de documents en langue L' sont plus similaires à un document d_i que sa traduction (selon la mesure choisie), on réalise alors un test statistique sur la proportion f_i . Plus précisément, on cherche à vérifier que la proportion f_i est significativement plus grande qu'une valeur seuil p_0 fixée *a priori*. Il s'agit donc du test d'hypothèse :

$$\begin{cases} H_0 : f_i = p_0 \\ H_1 : f_i > p_0 \end{cases}$$

Pour un échantillon assez grand (dans notre cas, $n = 500$), on rejette l'hypothèse H_0 au niveau α si (Grais, 1986, p. 261) :

$$\frac{f_i - p_0}{\sqrt{p_0(1 - p_0)}} \times \sqrt{n} > t_\alpha$$

où t_α est la valeur de la variable normale centrée réduite pour un seuil d'erreur α . Cela revient à rejeter l'hypothèse H_0 si :

$$f_i > p_0 + t_\alpha \sqrt{\frac{p_0(1 - p_0)}{n}} \quad (3)$$

4. Données et Résultats

4.1. Données

Les documents utilisés sont extraits du corpus *JOC* contenant les transcriptions des sessions de questions et de réponses du Parlement Européen, publiées en cinq langues dans le Journal Officiel de la Communauté Européenne.

Les langues retenues pour le test ont été le français et l'anglais et, pour ces deux langues, les tests ont été fait dans les deux sens :

- *fr-en* : { L =français, L' =anglais }
- *en-fr* : { L =anglais, L' =français }.

Le corpus complet contient 6729 documents alignés dans les deux langues. Parmi ces documents, nous avons sélectionné de façon aléatoire un corpus de test de $n = 500$ documents, les $N = 6229$ documents restants constituant donc les corpus de référence.

Les tailles des corpus en termes de nombre de mots, les tailles des lexiques (lemmes des noms, verbes et adjectifs) pour les corpus anglais et français, ainsi que les tailles des ensembles d'indexation (avec [70, 700] comme intervalle de fréquence pour le filtre fréquentiel) sont données dans la table 1.

	Corpus français	Corpus anglais
Nombres de mots dans le corpus	1160877	1053945
Taille du lexique (nombre de termes)	25322	24469
Taille de l'ensemble d'indexation	1062	1102

TAB. 1 – Taille des corpus, des lexiques et des ensembles d'indexation

4.2. Résultats

Pour le test statistique, nous avons choisi $p_0 = 0.9$. L'hypothèse considérée est donc que la proportion de documents ayant un positionnement relatif meilleur que celui de la traduction est inférieur à 10%. Le taux d'erreur α a été quant à lui fixé à $\alpha = 2.5\%$, ce qui correspond à un $t_\alpha = 1.96$. Avec ces paramètres, l'équation (3) est donc :

$$f_i > 0.9 + 1.96 \sqrt{\frac{0.09}{500}} = 0.9263 \quad (4)$$

ce qui revient à accepter l'hypothèse H_0 s'il y a plus de 37 documents (sur les 500) qui sont « meilleurs » que la traduction elle-même.

Lors de nos expériences, nous avons donc produit les valeurs pour les grandeurs suivantes :

- le nombre N_R de documents (parmi les 500) pour lesquels l'hypothèse H_0 est acceptée, *i.e.* les documents pour lesquels l'hypothèse d'invariance n'est pas vérifiée ;

- le nombre N_0 de documents (parmi les 500) pour lesquels la traduction est *le* document dont le positionnement relatif est le plus semblable à celui du document original ($f_i = 0$), *i.e.* les documents pour lesquels l'hypothèse d'invariance est parfaitement vérifiée.

La table 2 présente les résultats obtenus, d'une part, avec une représentation vectorielle standard, en utilisant le schéma de pondération *ltn* de SMART et, d'autre part, avec une représentation vectorielle hybride du modèle DSIR, pour diverses valeurs pour le coefficient d'hybridation α (la valeur $\alpha = 1$ correspond au modèle vectoriel standard).

Toutes les valeurs indiquées sont les moyennes obtenues sur 30 corpus de test extraits de manière aléatoire et indépendante.

α	<i>en-fr</i>		<i>fr-en</i>	
	N_R	N_0	N_R	N_0
1 (VS)	2.03 (0.41%)	476.47 (95.3%)	2.9 (0.58%)	473.7 (94.74%)
0.9	1.67 (0.33%)	476.3 (95.26%)	2.57 (0.51%)	474.4 (94.88%)
0.8	1.47 (0.29%)	477.5 (95.5%)	1.97 (0.39%)	475.77 (95.15%)
0.7	1.23 (0.25%)	478.67 (95.73%)	1.73 (0.35%)	476.63 (95.33%)
0.6	0.9 (0.18%)	478.33 (95.67%)	1.4 (0.28%)	476.7 (95.34%)
0.5	1 (0.2%)	474.6 (94.92%)	1.33 (0.27%)	473.37 (94.67%)
0.4	4.9 (0.98%)	449.83 (89.97%)	3.2 (0.64%)	452.1 (90.42%)
0	352.43 (70.49%)	40.8 (8.16%)	344.87 (68.97%)	52.1 (10.42%)

TAB. 2 – Résultats de la validation de la similarité

Au vu des résultats obtenus, l'hypothèse d'une proportion inférieure à 10% des documents peut donc être acceptée pour, en moyenne, plus de 99% des documents (99.4% pour le sens *fr-en*, 99.6% pour le sens *en-fr*) dans le cas de la représentation vectorielle standard. De façon plus précise, les résultats montrent que, pour environ 95% des documents, la traduction est effectivement *le* document le meilleur (selon la mesure utilisée) parmi les 500 documents testés.

Ces résultats permettent donc incontestablement de postuler que le positionnement relatif des documents les uns par rapport aux autres dans l'espace vectoriel est stable lors du passage d'une langue à l'autre. Cela est en soi une confirmation intéressante de la pertinence des mesures de similarité entre documents fondées sur une représentation vectorielle simple des documents.

Notons au passage que ce résultat n'était pas évident, en particulier parce que les lexiques ne sont pas alignés. En effet, Les deux lexiques (anglais et français) sont choisis selon la même procédure, en fonction des fréquences en documents des unités linguistiques sélectionnées, mais chacun de façon indépendante en ne prenant en compte que le corpus dans la langue considérée. De ce fait, le terme associé à la i^e dimension d'un espace vectoriel n'a aucune raison de correspondre à la traduction du terme associé à la i^e dimension de l'autre espace vectoriel. Autrement dit, bien que la représentation des documents dans chaque langue ne dépende que du corpus associé à cette langue, la méthode de représentation permet la représentation des documents par des positions relatives stables d'une langue à l'autre.

D'autre part, il est également intéressant de noter que la représentation hybride du modèle DSIR permet d'améliorer encore la proportion de documents pour lesquels l'hypothèse d'invariance est acceptée, menant à près de 99.8% de documents pour lesquels l'hypothèse est vérifiée (cf.

figure 3). La meilleure valeur d'hybridation semble être entre 0.5 et 0.6. Par contre, la représentation par le seul modèle DSIR ($\alpha = 0$) donne de très mauvais résultats : ceci peut être expliqué par le fait que la représentation par les seuls profils de co-occurrences a l'inconvénient de « lisser » les représentations des documents : au contraire du modèle vectoriel standard dans lequel les vecteurs sont très creux et les différences très marquées, les vecteurs des représentations par le modèle DSIR sont des moyennes de profils de co-occurrences et ont des profils moins discriminants. Par contre, l'expérience confirme que les vecteurs DSIR contiennent des informations intéressantes pour la représentation qui sont capturées par l'utilisation du modèle hybride.

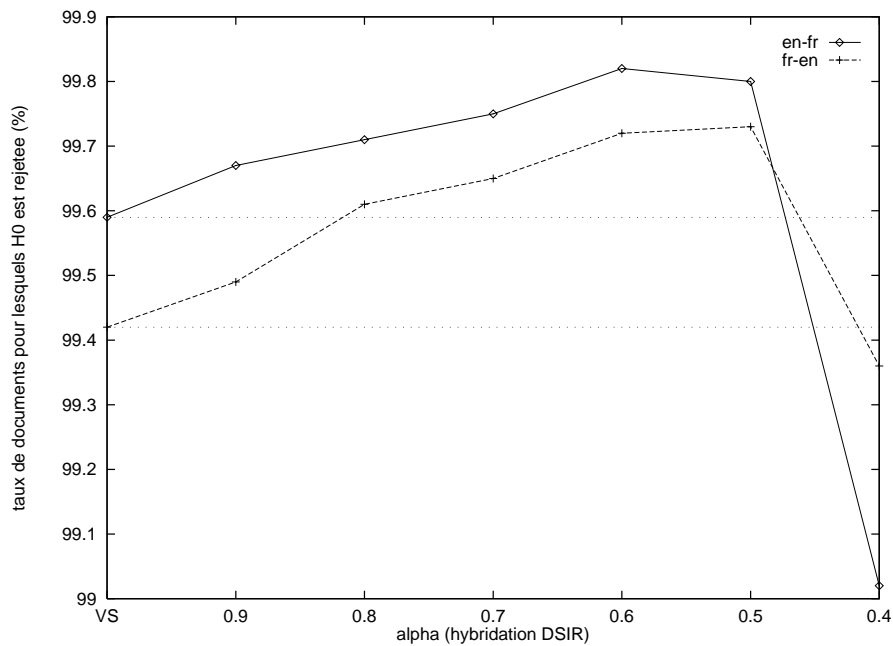


FIG. 3 – Effet de l'hybridation avec le modèle DSIR sur la pertinence de la similarité textuelle dans un cadre multilingue

5. Applications

La technique présentée pour valider la notion de similarité textuelle fondée sur une représentation vectorielle des documents peut être directement utilisée pour estimer la performance des systèmes de traduction automatique, plus précisément pour établir une comparaison entre différents systèmes de traduction.

Chaque système de traduction associe à un texte dans une langue source un texte traduit dans la langue cible. L'idée pour comparer la qualité sémantique des traductions (*i.e.* « est-ce que le texte traduit conserve l'information présente dans le texte source ? ») produite par les systèmes à évaluer est alors :

1. de calculer le vecteur de similarités caractérisant le positionnement du texte source par rapport à un corpus de référence (dans la langue source) ;
2. de calculer les vecteurs de similarités caractérisant les positionnements respectifs des traductions produites par les systèmes par rapport aux traductions des textes du corpus de référence ;

3. d'établir un classement selon la similarité entre ces vecteurs et le vecteur représentant le texte source.

Le classement ainsi obtenu permet d'estimer (de façon automatique) les performances relatives des systèmes de traduction automatique. Cette technique a été utilisée dans (Rajman and Hartley, 2001) et a montré que, si la technique est assez peu fiable pour un document particulier, elle est robuste pour le classement moyen (*i.e.* obtenu pour un ensemble des documents) et permet effectivement une évaluation des performances des systèmes de traduction relativement bien corrélée avec les mesures d'adéquation (*adequacy*) et d'informativité (*informativeness*) produites par des experts humains (Carroll, 1966).

Notons que d'autres applications de la méthode présentée peuvent être envisagées. En effet, la méthode de représentation des documents par un vecteur de similarités par rapport à un corpus de référence étant indépendante de la langue, elle peut par exemple être utilisée pour calculer des similarités entre documents de langue différente dans des applications destinées à l'exploitation de corpus multilingues, comme par exemple la recherche documentaire dans des corpus multilingues (Oard, 1997).

6. Conclusion

La notion de similarité textuelle étant fondamentale pour de nombreux systèmes de traitement de la langue, il est important qu'elle puisse être validée d'une façon indépendante d'une application particulière. Nous proposons ici une validation reposant sur l'invariance des positionnements respectifs des documents, représentés dans un espace vectoriel pour des langues différentes. Les résultats très positifs obtenus pour cette validation sont intéressants car ils confirment les hypothèses faites sur la stabilité et la pertinence de la similarité fondée sur une représentation vectorielle. La méthode de validation proposée peut également être utilisée pour comparer différents modèles de représentation : on a par exemple montré dans cet article que le modèle DSIR intégrant des co-occurrences dans la représentation vectorielle permet d'améliorer la pertinence de la similarité textuelle.

Toutefois, les résultats présentés ici sont des résultats quantitatifs globaux. Une étude plus approfondie devrait être entreprise, en particulier pour mettre en évidence de façon détaillée les propriétés des documents pour lesquels l'hypothèse d'invariance doit être rejetée, ainsi que l'apport de la représentation à l'aide du modèle DSIR pour ces documents.

D'autre part, la validation n'a été effectuée qu'avec un corpus bilingue anglais-français. Des tests avec d'autres langues devraient également être envisagés pour vérifier la validité générale des conclusions, en particulier pour des paires de langues plus fortement différentes que l'anglais et le français.

Enfin, la technique utilisée pour effectuer la validation trouve également des applications pour d'autres tâches comme la comparaison automatique des performances de différents systèmes de traduction automatique ou la recherche documentaire multilingue.

Références

- Besançon R. (2001). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes*. PhD thesis, École Polytechnique Fédérale de Lausanne.

- Carbonell J. G., Yang Y., Frederking R. E., Brown R. D., Geng Y., and Lee D. (1997). Translingual information retrieval: A comparative evaluation. In *IJCAI (1)*, pages 708–715.
- Carroll J. (1966). An experiment in evaluating the quality of translations. In Pierce. J. editor, *Language and machines. Report by ALPAC. NASNRC*, pages 67–75.
- Constant P. (1995). *Manuel de développement SYLEX-BASE*. INGÉNIA-LN, Paris, France.
- Dumais S., Landauer T., and Littman M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pages 16–23.
- Grais B. (1986). *Méthodes Statistiques*. Dunod.
- Littman M. and Jiang F. (1998). A comparison of two corpus-bases methods for translingual information retrieval. Technical Report CS-1998-11, Department of Computer Science, Duke University, Durham, North Carolina 27708-0129.
- Oard D. W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*.
- Rajman M., Besançon R., and Chappelier J.-C. (2000). Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2).
- Rajman M. and Bonnet A. (1992). Corpora-base linguistics: new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.
- Rajman M. and Hartley T. (2001). Automatically predicting mt systems rankings compatible with fluency, adequacy and informativeness scores. In *MT evaluation workshop, MT Summit VIII*, Siantiago de Compostela, Spain.
- Rungsawang A. (1997). *Recherche Documentaire à base de sémantique distributionnelle*. PhD thesis, ENST, Paris.
- Salton G. editor (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.
- Salton G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.
- Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill.
- Salton G., Wong A., and Yang C. S. (1975a). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Salton G., Yang C. S., and Yu C. T. (1975b). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.
- Singhal A. (1997). *Term Weighting Revisited*. PhD thesis, Department of Computer Science, Cornell University.
- Yang Y., Carbonell J. G., Brown R. D., and Frederking R. E. (1998). Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence*, 103(1–2):323–345.