Similarité de second ordre pour l'exploration de bases textuelles multilingues

Nikola Tulechki, Ludovic Tanguy

Similarité de second ordre pour l'exploration de bases textuelles multilingues

Tulechki Nikola[1,2] Tanguy Ludovic[1]
(1) CLLE-ERSS : CNRS et Université de Toulouse 2, 5 allées Antonio Machado, 31058 Toulouse CEDEX 9
(2) Conseil en Facteurs Humains, 4 impasse Montcabrier, 31500 Toulouse
{tanguy,tulechki}@univ-tlse2.fr

RÉSUMÉ

Cet article décrit l'utilisation de la technique de *similarité de second ordre* pour l'identification de textes semblables au sein d'une base de rapports d'incidents aéronautiques mélangeant les langues française et anglaise. L'objectif du système est, pour un document donné, de retrouver des documents au contenu similaire quelle que soit leur langue. Nous utilisons un corpus bilingue aligné de rapports d'accidents aéronautiques pour construire des paires de pivots et indexons les documents avec des vecteurs de similarités, tels que chaque coordonnée correspond au score de similarité entre un document dans une langue donnée et la partie du pivot de la même langue. Nous évaluons les performances du système sur un volumineux corpus de rapports d'incidents aéronautiques pour lesquels nous disposons de traductions. Les résultats sont prometteurs et valident la technique.

ABSTRACT

Second order similarity for exploring multilingual textual databases

This paper describes the use of *second order similarities* for identifying similar texts inside a corpus of aviation incident reports written in both French and English. We use a second bilingual corpus to construct pairs of reference documents and map each target document to a vector so each coordinate represents a similarity score between this document and the part of the reference corpus written in the same language. We evaluate the system using a large corpus of translated incident reports. The results are promising and validate the approach.

MOTS-CLÉS : similarité de second ordre, multilingue, ESA.

KEYWORDS: second order similarity, multilingual, ESA.

1 Introduction et contexte applicatif

Dans toute industrie à risque, le retour d'expérience (REX) occupe une place capitale dans les mécanismes de gestion de la sûreté. Des politiques de recueil, d'analyse et de stockage sont mises en place afin de garder une trace de tout événement qui s'écarte de la norme, de tout incident ou accident qui survient lors des opérations. Les informations ainsi recueillies servent ensuite de support aux experts de sûreté pour mettre à jour les règles et les procédures d'exploitation en les adaptant à un contexte en perpétuelle évolution.

L'aviation civile est sans doute le secteur dans lequel les politiques de recueil sont les plus avancées et il n'est pas rare que les bases de REX regroupent plusieurs centaines de milliers de rapports.

Les stratégies d'exploitation actuelles, basées sur la codification manuelle de chaque rapport s'avèrent insuffisantes, à cause d'un codage souvent incomplet et hétérogène (Tulechki et Tanguy, 2012). De ce fait, proposer aux experts des outils facilitant l'accès à l'information contenue dans la partie textuelle des rapports est devenu capitale (Tulechki, 2011). Plus précisément encore, l'un des moyens privilégiés d'exploitation de ce type de base par des experts consiste à partir d'un événement particulier et à rechercher des cas similaires afin de faire émerger de nouveaux risques non encore identifiés (et codés).

Cependant, compte tenu du caractère intrinsèquement international de l'activité, les informations dans les bases sont souvent écrites dans des langues différentes, ce qui complique considérablement leur exploitation de manière outillée. Notre objectif est donc de concevoir un système capable de calculer la similarité textuelle entre deux textes, quelle que soit la langue dans laquelle ils sont écrits. Afin que le traitement de plusieurs langues soit possible les textes doivent d'abord être ramenés à une représentation commune. Traditionnellement ceci implique l'utilisation de techniques de traduction automatique (TA). Dans notre cas la TA n'est pas envisageable puisque que les systèmes de TA disponibles ne sont pas adaptés aux particularités stylistiques du langage technique de l'aviation. Pour ces raisons nous nous sommes tournés vers la *similarité de second ordre*, qui pour une implémentation multilingue ne nécessite pas d'autres ressources qu'un corpus aligné servant d'intermédiaire (Claveau, 2012).

Dans un premier temps nous présenterons les principes généraux d'approche par similarité de second ordre monolingue ainsi que son application dans des contextes multilingues. Ensuite nous détaillerons notre expérience sur un corpus spécialisé multilingue.

2 Similarité textuelle

2.1 Similarité de premier ordre

Calculer la similarité textuelle revient à attribuer un score représentant le degré de ressemblance entre deux textes en se basant sur leur taux de recouvrement lexical. Aujourd'hui encore le modèle vectoriel (Salton *et al.*, 1975) est le plus couramment utilisé. Le score de similarité est obtenu en calculant le recouvrement (généralement par une mesure de type cosinus) entre deux vecteurs dans un espace à n dimensions correspondant aux termes présents dans la collection. Compte tenu du fait que les documents sont rapprochés grâce aux termes qu'il partagent, cette approche est particulièrement sensible à la variation lexicale. Deux documents qui traitent du même sujet, mais y réfèrent avec des synonymes ne seront pas rapprochés par le calcul et les techniques existantes bien connues, visant à en assurer le rapprochement, reposent classiquement sur des ressources lexicales coûteuses à développer et à maintenir dans le cadre d'un domaine très spécialisé. Cette similarité est dite de *premier ordre* dans la suite de cet article.

2.2 Similarité de second ordre

2.2.1 Principe de base

De multiples techniques cherchant à représenter plus fidèlement les textes en fonction de leur contenu et à maîtriser les incohérences dues à la variation lexicale ont vu le jour. Une en particulier, mise au point par Gabrilovich et Markovitch (2007) consiste à calculer une similarité de premier ordre entre chaque document de la collection et un ensemble de n documents

pivots arbitraires extérieures à cette collection. Les scores forment par la suite un vecteur de *n* dimensions qui est utilisé pour représenter le document. La similarité est ensuite calculée de manière standard en comparant les vecteurs des documents dans ce nouvel espace (voir figure 1).

L'implémentation originelle, appelée ESA [1] a été évaluée sur un corpus de paires de textes sur lesquels un jugement de similarité avait été donné par des annotateurs humains. Le système atteint des performances supérieures à la fois à la similarité de premier ordre et aux techniques de réduction de dimensions comme la LSA/LSI.

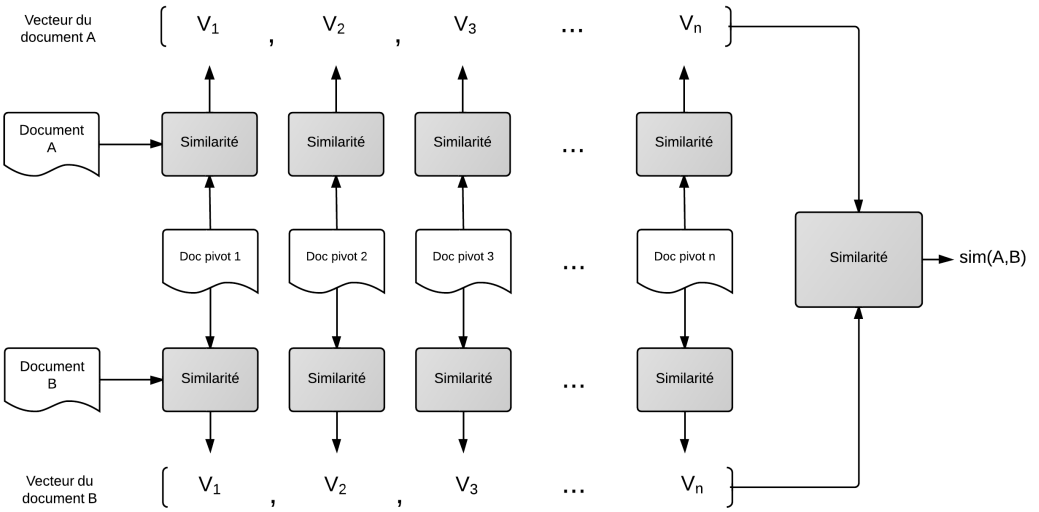

FIGURE 1 – Principe de la similarité de second ordre

On voit bien que le contenu des deux documents n'est pas indexé directement. Cette technique permet donc de traiter des documents en évitant de se baser sur le partage de termes.

2.2.2 Le choix des pivots

Originellement l'ESA utilise des articles de Wikipédia comme documents pivots. Ses auteurs insistent sur l'apport en terme de connaissances de leur choix et l'importance du fait que l'espace ainsi construit est déterminé par rapport aux "concepts naturels" définis par les rédacteurs de l'encyclopédie. Le caractère "explicite" [2] permet en effet que chacune des dimensions soit directement interprétable. Il s'en est suivi qu'une partie considérable de la recherche dans ce domaine s'est centrée sur les stratégies d'exploitation de la catégorisation de Wikipédia afin de construire des pivots en concaténant des articles en fonction de leur place dans la hiérarchie.

Cependant Claveau (2012) a démontré que la similarité de second ordre peut être efficace sans obligatoirement se baser sur une ressource structurée. En utilisant des textes tout-venants comme

1. Explicit Semantic Analysis
2. Les auteurs ont sans doute choisi cette dénomination pour se différencier des "concepts implicites" formés par les méthodes de réduction de dimensions.

pivots, il a évalué la technique sur des tâches de RI et de fouille de texte en obtenant à chaque fois des résultats encourageants.

La question du choix des pivots pour le traitement des textes d'un domaine spécialisé ne s'est pas encore posée dans la littérature. Néanmoins, il semble évident que compte tenu du fonctionnement de la similarité de second ordre, utiliser des pivots issus du même domaine est préférable. Des pivots inadaptés aux documents traités peuvent à la fois engendrer du bruit et du silence ; indexer un rapport d'accident aéronautique en utilisant sa similarité (ou plutôt sa différence) avec l'article Wikipédia sur Walt Disney ne semble guère distinctif. Pire encore, un terme spécifique contenu dans les documents mais absent des pivots sera perdu à jamais du point de vue du calcul.

2.3 Application inter-langue

L'adaptation de la similarité de second ordre à un contexte multilingue est relativement simple. L'espace dans lequel sont représentés les documents étant indépendant [3] de la langue, tout document peut y être représenté. Pour cela il suffit d'utiliser comme pivots des paires de documents traduits dans plusieurs langues afin de pouvoir calculer les similarités de premier ordre avec la partie de la collection écrite dans la même langue que le document (voir figure 2).

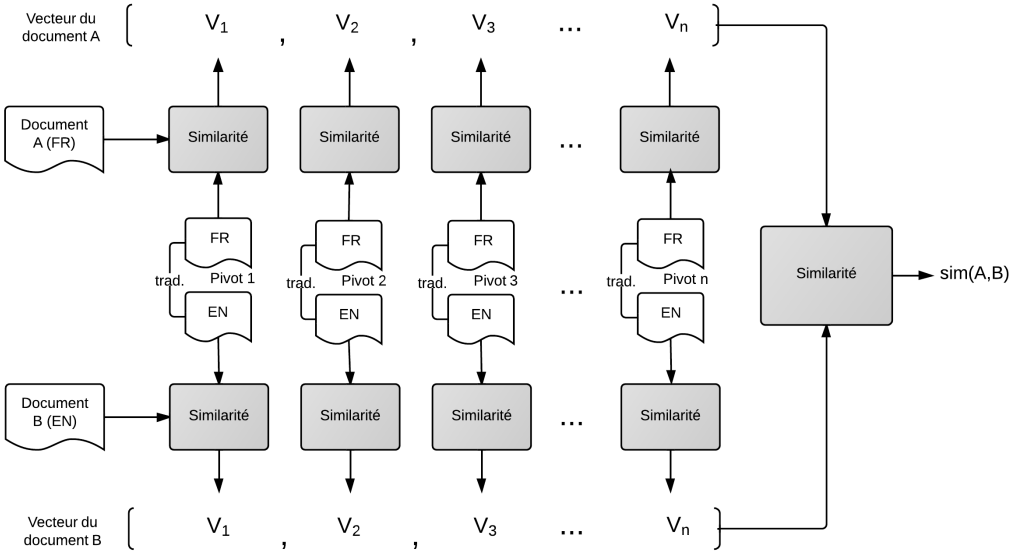

FIGURE 2 – Similarité de second ordre inter-langue

Sorg et Cimiano (2012) ont appliqué l'ESA à plusieurs langues. Pour cela ils construisent des ensembles de pivots en exploitant les liens de traduction présents dans Wikipédia. Si le *n*-ème pivot correspond au concept *Hôpital*, la *n*-ème coordonnée des vecteurs des documents en anglais

3. Par comparaison à l'espace des termes pour la similarité de premier ordre.

correspondra à la similarité entre le document et l'article *Hospital* de la version anglaise de l'encyclopédie ainsi que la même coordonnée d'un document en allemand correspondra à la similarité entre le document et l'article *Krankenhaus* de la version allemande. Les documents sont ainsi représentés dans le même espace et une similarité peut être calculée à la fois entre documents d'une même langue et de langues différentes.

Afin d'évaluer le système les auteurs utilisent un corpus parallèle de documents législatifs traduits dans plusieurs langues et la tâche de *recherche de partenaire* [4]. Étant donné un document dans une langue donnée, la tâche consiste à retrouver ses traductions (partenaires) parmi les documents de la base. Comme mesure, les auteurs utilisent le *rappel au rang k* (R@k), qui consiste à chercher le partenaire parmi les *k* documents les plus similaires retournés par le système. Un R@10 de 1 signifie que pour tout document, sa traduction se trouve dans les 10 premiers documents. Ce score repose sur l'hypothèse qu'un système performant doit maximiser la similarité entre un document et sa traduction. Lors de l'évaluation de leur système, Sorg et Cimiano (2012) atteignent un R@10 variant entre 0,27 et 0,51.

Une méthode similaire à été également utilisée pour la clusterisation de documents multilingues (Kiran Kumar *et al.*, 2011), toujours en utilisant la Wikipédia comme corpus pivot.

3 Application à un domaine spécialisé

Notre système s'inspire des travaux cités précédemment afin d'adapter la technique à un corpus de rapports d'incidents aéronautiques écrits en français et en anglais. Nous utilisons deux corpus distincts :

Pour les pivots, nous utilisons un corpus de rapports d'accidents du Bureau de la Sécurité des Transports du Canada [5]. Ces documents longs de plusieurs pages existent systématiquement en anglais et en français et décrivent de façon exhaustive l'analyse d'un accident aéronautique. Afin d'obtenir un nombre suffisant de pivots, nous les découpons en paragraphes [6] que nous alignons entre les deux langues en nous basant sur l'isomorphie de leurs structures HTML. Ce découpage permet d'obtenir 10032 paires de pivots à partir de 390 paires de documents.

A des fins d'évaluation, nous utilisons un second corpus de rapports d'incidents issu de la base CADORS [7] qui contient des rapports volontairement soumis aux autorités de régulation de l'aviation canadienne. Ces documents d'une centaine de mots en moyenne résument un incident aéronautique. Ils sont très semblables aux textes des autorités de contrôle françaises auxquelles notre système est destiné. Compte tenu de la réglementation canadienne, comme pour le corpus des pivots, les rapports québécois sont systématiquement traduits et nous pouvons donc procéder à une évaluation par la tâche de *recherche de partenaire*. Au total le corpus d'évaluation comporte 9217 documents bilingues comme ceux présentés en exemple en table 1.

4. *mate retrieval*
5. `http://www.bst-tsb.gc.ca/fra/rapports-reports/aviation/index.asp`
6. Nous avons choisi ce niveau de grain, afin d'obtenir suffisamment de pivots pour un bon fonctionnement du système.
7. Civil Aviation Daily Occurrence Reporting System. `http://wwwapps.tc.gc.ca/Saf-Sec-Sur/2/cadors-screaq/`

CRQ590M, a Beech A100 operated by Air Creebec as flight number CRQ590, was on an IFR MEDEVAC flight from Chibougamau/Chapais (CYMT) to Montréal/Trudeau (CYUL). At 1535Z, the crew was instructed to conduct a missed approach for Runway 06R due to the presence of C FFWJ, an Airbus A-320 operated by Air Canada as flight number ACA407, which was lined up for departure and which had a mechanical problem. CRQ590 eventually landed without incident at 1546Z.	CRQ590M, un Beech A100 exploité par Air Creebec sous l'indicatif de vol CRQ590, effectuait un vol d'évacuation médicale selon les règles de vol aux instruments (IFR) depuis Chibougamau / Chapais (CYMT) à destination de Montréal/Trudeau (CYUL). À 1535Z, l'équipage a reçu l'instruction d'interrompre son approche pour la piste 06 droite en raison de la présence de C-FFWJ, un Airbus A-320 exploité par Air Canada sous l'indicatif de vol ACA407 qui était aligné au départ et qui avait un problème mécanique. CRQ590 a finalement atterri sans encombre à 1546Z.

TABLE 1 – Exemple de rapport d'incident et sa traduction

4 Architecture du système

Prétraitements et normalisation

Nous utilisons pour le prétraitement des corpus (documents-pivots et corpus d'évaluation) des outils génériques disponibles pour le langage Perl. La segmentation est ainsi faite par un simple *tokeniseur* [8] basé sur des expressions régulières. Nous appliquons ensuite le raciniseur *Snowball* [9] et un anti-dictionnaire standard. Vu que les corpus sur lesquels nous travaillons sont souvent de mauvaise qualité, comportant de nombreux documents écrits entièrement en majuscules, nous normalisons la casse et supprimons les accents pour le français.

Pondération et calcul de similarité

Afin de prendre en compte l'importance relative des termes dans les documents nous utilisons un schéma de pondération proposé par Turney et Pantel (2010) : la *Positive Pointwise Mutual Information* pour la similarité entre les documents et les pivots. Les vecteurs de second ordre ne sont pas pondérés : tous les documents-pivots ont un poids identique pour le calcul de la similarité (basé sur une mesure cosinus).

Élagage

Contrairement aux vecteurs de premier ordre, très creux par définition, les vecteurs de second ordre sont systématiquement pleins. Ceci alourdit considérablement le calcul et pour cette raison nous appliquons un seuil minimum arbitraire de 0,05 et ramenons tout score inférieur à ce seuil à zéro. Cette opération laisse des vecteurs de second ordre relativement creux avec en moyenne 45 valeurs non-nulles (sur 10000) par document.

5 Évaluation

Afin d'évaluer le système, nous avons appliqué la tâche de *recherche de partenaire* au corpus issu de la base CADORS cité ci-dessus.

8. `http://search.cpan.org/~dami/Search-Tokenizer-1.01/lib/Search/Tokenizer.pm`
9. `http://search.cpan.org/~creamyg/Lingua-Stem-Snowball-0.952/lib/Lingua/Stem/Snowball.pm`

Lors de nos premiers tests, nous avons trouvé que pour certains rapports, le partenaire (*i.e.* sa traduction) se trouvait très loin dans la liste des résultats, dans certains cas à un rang supérieur à 500. Nous avons regardé les rapports en question et nous nous sommes aperçus que le corpus contenait des séries de rapports très similaires, au point de poser la question des limites de l'intérêt de l'analyse de similarité pour certains textes. En effet, à cause de la nature réglementaire du signalement d'incidents aéronautiques, certains problèmes courants sont systématiquement rapportés *via* des textes standardisés selon un schéma commun [10]. Il apparaît clairement que les seules différences entre les documents de ces séries sont des codes, des nombres et éventuellement des noms de villes à priori absents des pivots et dont l'impact sur la similarité est nul. Retrouver la traduction au sein de la série repose par contre uniquement sur ces éléments, ce qui explique le problème rencontré. Si notre méthode est inadaptée à ces cas particuliers, ils peuvent être traités par des méthodes de surface simples.

Nous avons décidé de ne pas les prendre en considération en les identifiant en calculant pour chaque document la similarité moyenne des 100 premiers rapports similaires. Si cette moyenne dépassait 0,95, nous considérons que le rapport en question est un texte préformaté et l'excluons du corpus d'évaluation. Au total 823 paires de documents ont été exclues.

Le corpus final d'évaluation comporte donc 16788 documents monolingues, de façon à ce que la traduction de chacun soit aussi présente dans la base. Nous avons procédé à la tâche de recherche du partenaire pour la totalité du corpus et calculé le R@k pour les rangs 1, 10 et 100, séparément pour les documents en français et en anglais en ne prenant en compte que les documents retournés qui ne sont pas de la même langue que le document source. Les résultats sont résumés dans la table 2.

	FR	EN
R@1	0,43	0,45
R@10	0,71	0,74
R@100	0,90	0,94

TABLE 2 – Résultats de la recherche de partenaire

Comme nous pouvons le voir, les résultats sont encourageants et valident cette approche, au même niveau pour les deux langues. Dans plus de 40% des cas, la traduction est bien le document le plus similaire retourné par le système. Dans plus de 70% des cas, la traduction se situe dans les 10 documents les plus similaires.

6 Conclusion et perspectives

Nous avons présenté une approche permettant de calculer la similarité entre documents de langues différentes issues d'un domaine spécialisé que nous avons évaluée sur un grand corpus de documents réels, semblables aux documents auxquels le système est destiné. Cette expérience permet de valider la méthode et nous encourage à nous intéresser davantage aux particularités de ce type de calcul.

10. Les deux courts rapports ci-dessous exemplifient ce fait :
A :"*La station radio d'aéroport communautaire (CARS) de Waskaganish (CYKQ) n'a pas assuré les services de météo et de radio d'aérodrome entre 1300Z et 2100Z.*"
B :"*La station radio d'aéroport communautaire (CARS) d'Inukjuak (CYPH) n'a pas assuré les services de météo et de radio d'aérodrome entre 1130Z et 2130Z.*"

Puisque la méthode est basée sur une similarité classique de premier ordre (entre les documents-cibles et les pivots), il est logique que le paramétrage de cette dernière influence les performances. L'expérience présentée dans cet article utilise une chaîne de similarité basique, mais nous explorerons à l'avenir l'apport de traitements linguistiques plus sophistiqués en amont.

Le côté *explicite* de la méthode nous amènera surtout à nous intéresser de près aux documents pivots et à analyser plus précisément leur rôle dans le calcul final du score de similarité. Le fait que nous y avons facilement accès et que les pivots sont interprétables nous permettra de facilement tracer et comprendre les variations du comportement du système avec différents ensembles de documents pivots. Dans cette logique nous poursuivrons les recherches entamées dans Tulechki et Tanguy (2012) visant à identifier les *dimensions de similarité* entre des documents. Si, comme c'est le cas dans les données utilisées, les documents-pivots disposent d'un codage spécifique (méta-données, catégorisation externe, etc.), nous pourrons l'exploiter à la fois pour identifier ces dimensions, mais aussi pour restreindre les pivots en fonction de leurs caractéristiques, et ainsi orienter de façon interactive l'investigation en fonction des facettes exprimées par l'utilisateur.

Remerciements

Nous tenons à remercier Assaf Urieli de CLLE-ERSS d'avoir adapté son calculateur de similarité aux exigences particulières de cette expérience.

Références

Claveau, V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF. *In Actes de TALN*, pages 85–98, Grenoble.

Gabrilovich, E. et Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *In Proceedings of IJCAI*, pages 1606–1611, Hyderabad, India.

Kiran Kumar, N., Santosh, K. G. S. et Varma, V. (2011). Multilingual document clustering using wikipedia as external knowledge. *In Proceedings of IRFC*, pages 108–117.

Salton, G., Wong, A. et Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613—620.

Sorg, P. et Cimiano, P. (2012). Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74(0):26 – 45.

Tulechki, N. (2011). Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. *In Actes de RECITAL*, Montpellier.

Tulechki, N. et Tanguy, L. (2012). Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. *In Actes de TALN*, Grenoble.

Turney, P. D. et Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188.