Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire

Hai-Hieu Vu, Jeanne Villaneau, Farida Saïd, Pierre-François Marteau

▶ **To cite this version:**

Hai-Hieu Vu, Jeanne Villaneau, Farida Saïd, Pierre-François Marteau. Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire. TALN 2015, Jun 2015, Caen, France. hal-01167929

HAL Id: hal-01167929
https://hal.science/hal-01167929

Submitted on 25 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire

Hai-Hieu Vu[1] Jeanne Villaneau[1] Farida Saïd[2] Pierre-François Marteau[1]
(1) Université de Bretagne Sud, laboratoire IRISA
(2) Université de Bretagne Sud, laboratoire LMBA
hai-hieu.vu, jeanne.villaneau, farida.said, pierre-francois.marteau@univ-ubs.fr

Résumé. Cet article présente une méthode pour mesurer la similarité sémantique entre phrases qui utilise Wikipédia comme unique ressource linguistique et qui est, de ce fait, utilisable pour un grand nombre de langues. Basée sur une représentation vectorielle, elle utilise une indexation aléatoire pour réduire la dimension des espaces manipulés. En outre, elle inclut une technique de calcul des vecteurs de termes qui corrige les défauts engendrés par l'utilisation d'un corpus aussi général que Wikipédia. Le système a été évalué sur les données de SemEval 2014 en anglais avec des résultats très encourageants, au-dessus du niveau moyen des systèmes en compétition. Il a également été testé sur un ensemble de paires de phrases en français, à partir de ressources que nous avons construites et qui seront mises à la libre disposition de la communauté scientifique.

Abstract.
Semantic similarity between sentences based on Wikipedia and Random Indexing.
This paper proposes a semantic similarity measure for sentence comparison based on the exploitation of Wikipedia as the only language resource. Such similarity measure is therefore usable for a wide range of languages, basically those covered by Wikipedia. Random Indexing is used to cope with the great dimensionality and the spareness of the data vectorial representations. Furthermore, a statistical weight function is used to reduce the noise generated by the use of a multi domain corpus such as Wikipedia. This semantic similarity measure has been evaluated on SemEval 2014 dataset for English language leading to very promising results, basically above the average level of the competing systems that exploit Wikipédia in conjunction with other sources of semantic information. It has been also evaluated on a set of pairs of sentences in French that we have build specifically for the task, and made freely available for the research community.

Mots-clés : Similarité sémantique, Indexation aléatoire, Wikipédia, Relation sémantique.

Keywords: Semantic Textual Similarity, Random indexing, Wikipédia, Semantic Relatedness.

1 Introduction

Mesurer la similarité entre deux phrases (ou textes courts) consiste à évaluer jusqu'à quel point le sens de ces phrases est proche. Cette tâche (STS : Semantic Textual Similarity) est souvent utilisée dans plusieurs domaines importants du Traitement Automatique des Langues (TAL), parmi lesquels on peut citer la recherche d'informations (Balasubramanian *et al.*, 2007), la catégorisation de textes (Ko *et al.*, 2002), le résumé de texte (Erkan & Radev, 2004), la traduction automatique, etc. Longtemps considérée comme une sous-tâche dans les domaines cités, la STS fait depuis quelques années l'objet d'un intérêt croissant. Depuis 2012, la tâche STS de SemEval confronte les résultats de différents systèmes, presque tous consacrés à la langue anglaise. La version 2014 de Semeval a cependant proposé une évaluation des systèmes sur des phrases en espagnol, à laquelle 9 équipes ont participé (Agirre *et al.*, 2014).

La similarité lexicale constitue une première approche pour mesurer la similarité entre deux textes (Hirao *et al.*, 2005; Lin, 2004). Cependant, elle ne tient compte, ni des relations sémantiques entre les mots ou groupes de mots d'un même texte, ni de la similarité sémantique entre les mots des deux textes (synonymie, paraphrase, etc.). Pour pallier ce manque et suivant le principe selon lequel les mots qui apparaissent dans un même contexte ont potentiellement une similarité sémantique importante, les systèmes récents se fondent sur des études statistiques de gros corpus de la langue qui permettent de prendre en compte ces contextes. Les meilleurs systèmes de la tâche STS de SemEval2014 utilisent des ressources

linguistiques qui ne sont disponibles que pour la langue anglaise en y incluant, outre des corpus de très grande taille, des corpus de paraphrases, le WordNet, etc. (Kashyap *et al.*, 2014; Sultan *et al.*, 2014). Il est également intéressant de constater que les systèmes qui sont arrivés en tête dans le challenge en langue espagnole de SemEval ont utilisé un système réalisé pour l'Anglais, en transformant les phrases données en espagnol en leur équivalent anglais (Chavez *et al.*, 2014; Kashyap *et al.*, 2014).

Pour les langues moins bien dotées en ressources linguistiques que ne l'est la langue anglaise, Wikipédia représente un corpus très intéressant en raison de sa taille croissante et de son caractère encyclopédique qui assure une couverture très générale de presque tous les domaines. Wikipédia représente donc une énorme ressource multilingue pour le traitement automatique de la langue naturelle (TAL), qui est exploitée de différentes façons, et en particulier pour définir des relations sémantiques entre termes et entre textes (cf. section 2).

Le système présenté dans cet article (WikiRI) repose sur un modèle vectoriel, ou Vector Space Models (VSM). Le principe consiste à construire un espace vectoriel de grande dimension, dans lequel un mot est représenté par un vecteur unique qui rend compte de ses contextes d'occurrence. Plus précisément, le modèle utilisé est celui des GVSM (Generalized Vector Space Model), où les documents sont utilisés comme base de l'espace. Les termes y sont représentés comme des vecteurs dans la base des concepts définis à partir des articles de Wikipédia. Pour remédier aux problèmes posés par le nombre d'articles présents dans Wikipédia et sa constante augmentation, nous proposons une représentation vectorielle de la sémantique des termes qui utilise le Random Indexing (RI) (cf. section 3). Par ailleurs, WikiRI introduit des modifications dans les calculs des vecteurs de termes pour corriger le bruit engendré par l'utilisation d'une ressource linguistique aussi encyclopédique que Wikipédia : elles sont détaillées dans la section 4.

Nous avons effectué les expérimentations et les évaluations sur des ensembles de données en français (Sensim-french [1] que nous avons construites et sur les données de SemEval 2014 pour l'anglais (SemEval-2014 Task 10 [2]). Elles indiquent des résultats intéressants qui sont décrits dans la section 5.

2 Wikipédia en tant que ressource linguistique

Actuellement disponible dans 288 langues, Wikipédia est le plus grand référentiel de connaissances générales sur le Web. Les statistiques officielles de Wikipédia en date du 12/12/2014 font état d'un nombre d'articles en langue anglaise de 4 668 468 et de 1 569 491 articles pour la langue française.

- **Structure du réseau** : Si l'on ne tient pas compte de la direction des liens entre articles, le graphe de Wikipédia est presque entièrement connecté : 98.5% des articles sont liés les uns aux autres. En tenant compte de la direction des liens, on retrouve la structure en noeud papillon du Web : des composantes denses fortement connectées sont liées entre elles par des liens unidirectionnels. La zone centrale (SCC) - pour strongly connected component - est composée

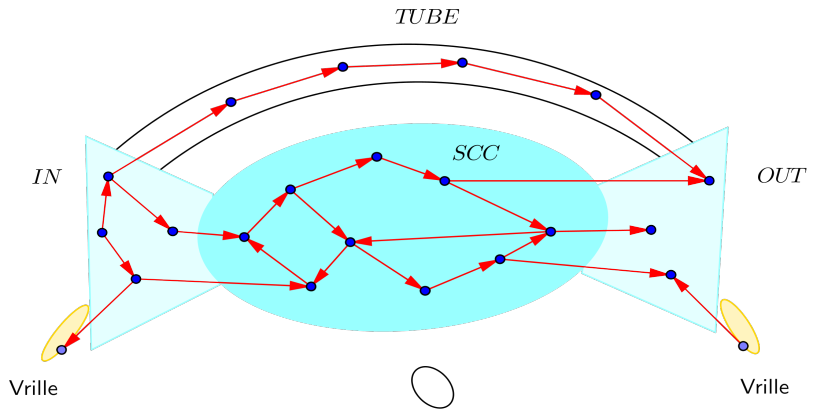

FIGURE 1 – Structure en noeud-papillon de Wikipédia

d'articles fortement liés entre eux : deux articles quelconques de cette zone peuvent toujours être liés par un chemin

1. http ://vuhaihieu-001-site1.smarterasp.net
2. http ://alt.qcri.org/semeval2014/task10/

direct ou indirect. La zone (IN), de taille plus réduite, est composée d'articles qui permettent d'accéder aux articles de la zone (SCC), mais qui ne sont pas accessibles depuis cette zone. La zone (OUT), de taille équivalente à (IN) est composée à l'inverse d'articles qui sont accessibles depuis la zone (SCC), mais qui n'y renvoient pas. Les tubes sont des zones de taille plus réduites, qui relient directement les articles de la zone (IN) aux articles de la zone (OUT), sans passer par la zone (SCC). Les vrilles sont des zones atypiques qui relient des articles isolés de l'ensemble, soit à la zone (OUT), soit à la zone (IN).
Plus de 2/3 des articles de Wikipédia appartiennent au large noyau (SCC) et un signe de maturité de Wikipédia est la bonne stabilité dans le temps des différentes composantes ; ce qui serait actuellement le cas du Wikipédia anglais.

- **Nature sémantique des liens** : alors que dans les documents Web, un auteur peut arbitrairement lier une page à une autre, les liens dans Wikipédia indiquent une pertinence par rapport à un contexte local : un lien de la page A vers la page B indique que la page B est sémantiquement reliée au contenu, ou une partie du contenu de la page A.
- **Structure des liens** : les liens entrants dans Wikipédia ont tendance à se comporter comme les liens sortants (Jaap & Marijn, 2009) ; ce qui est consistant avec la nature sémantique des liens dans Wikipédia : si un lien de la page A vers la page B souligne une certaine pertinence de B alors il est vraisemblable que A soit également pertinent pour B.
- **Domaines couverts et qualité** : Wikipédia couvre des domaines de connaissance très variés, Arts, Géographie, Histoire, Science, Sports, Jeux... Dans le domaine des Sciences, cette encyclopédie collaborative s'avère aussi précise que l'"Encyclopedia Britannica" (Giles, 2005).
- **Evolution dans le temps** : la structure de Wikipédia et son évolution dans le temps sont régulièrement analysés (Voss, 2005; Buriol *et al.*, 2006; Capocci *et al.*, 2006; Nakayama *et al.*, 2008) et il s'avère qu'à l'instar du Web, cette encyclopédie se densifie au fil du temps aussi bien dans son contenu (nombre d'articles, longueur des articles) que dans sa structure en liens (nombre de liens entrants et sortants par article).
- **Référencement des articles** : chaque article (ou concept) de Wikipédia est référencé de manière unique par une adresse URL ; ce qui élimine tout risque d'ambiguïté.

Les caractéristiques précédentes et son multilinguisme font de Wikipédia un outil de choix pour le TAL qui ont d'ores et déjà donné lieu à des résultats intéressants (Gabrilovich & Markovitch, 2007; Hadj Taieb *et al.*, 2013; Strube & Ponzetto, 2006; Chan *et al.*, 2013). Cependant sa généralité, sa taille et son évolution permanente posent des problèmes de mise en œuvre, particulièrement pour les méthodes basées sur la vectorisation, étant donné la taille des espaces manipulés. Le Random Indexing est la solution que nous avons retenue pour pallier cette difficulté.

3 Random Indexing

Dans la méthodologie des VSM, un espace vectoriel de grande dimension est généré par la construction d'une matrice de co-occurrences F, dans laquelle chaque ligne F_w représente un unique mot et chaque colonne F_c représente un contexte c, typiquement un segment de plusieurs mots tel qu'un document, ou un autre mot. Dans les GVSM, ce sont les documents qui sont utilisés comme base de l'espace, pour répondre à la critique selon laquelle les mots ne constituent pas une base de vecteurs libres (Carbonell *et al.*, 1997).

Le modèle construit souffre de deux problèmes majeurs : la dimensionnalité et les données éparses. Lorsque le vocabulaire et le nombre de documents du corpus augmentent, la matrice de co-occurrence F entre termes et documents devient numériquement lourde à exploiter. Par ailleurs, une très grande proportion des mots n'apparaissent que dans un ensemble de documents très limité. Ainsi, dans une matrice de co-occurrence typique, 99% des entrées sont des zéros.

Pour pallier ces problèmes, diverses techniques de réduction de dimension peuvent être mises en œuvre, comme la décomposition en valeurs singulières (SVD) de la matrice F (Kumar, 2009). La nécessité de construire préalablement la matrice de co-occurrence entre termes et documents est un gros inconvénient lorsque l'on utilise des corpus en évolution constante tels que Wikipédia.

Une alternative aux techniques de réduction de dimension est le Random Indexing, basé sur le travail de Pentti Kanerva sur les représentations de données éparses (Kanerva, 1988; Kanerva *et al.*, 2000). Le Random Indexing procède d'abord par la représentation de chaque concept par un vecteur index de taille réduite, et ensuite le vecteur concept de chaque mot est calculé par sommation des vecteurs index de tous les concepts auxquels il est associé. Ainsi, l'ajout de nouveaux contextes n'implique pas une reconstruction complète de la matrice : il suffit de créer de nouveaux vecteurs index et d'adjoindre à la matrice les vecteurs colonnes correspondant aux nouveaux documents.

Les vecteurs index aléatoires sont choisis presque orthogonaux, ce qui conduit à une description approximative de l'espace contexte où les distances entre points sont approximativement préservées (William & Lindenstrauss, 1984). La description

qui suit du Random Indexing est faite à partir de celle qu'en a donnée Sahlgren (Sahlgren, 2005).

On alloue un vecteur index unique de longueur d à chaque contexte. Ces vecteurs sont constitués d'un grand nombre de 0 et d'un petit nombre de 1 et de -1. À chaque composante est allouée l'une de ces valeurs avec la probabilité suivante :

$$\begin{cases} +1 & \text{avec une probabilité } s/2 \\ 0 & \text{avec une probabilité } 1-s \\ -1 & \text{avec une probabilité } s/2 \end{cases}$$

où s désigne le nombre d'éléments non nuls. Le choix de s et d se fait en fonction du nombre de contextes à représenter. Pour chaque nouveau concept, un vecteur index est produit. Le vecteur contexte d'un terme est la somme des vecteurs index de tous les contextes dans lesquels ce terme apparaît.

Le vecteur contexte d'un terme qui apparaît dans chacun des contextes $c_1 = [1, 0, 0, -1]$ et $c_2 = [0, 1, 0, -1]$ serait $[1, 1, 0, -2]$. Si le contexte c_1 est rencontré de nouveau, il n'y a pas création de nouveau vecteur index et la mise-à-jour du vecteur contexte de t se fait par addition du vecteur index de c_1 ; ce qui conduit au nouveau vecteur contexte de t : $[2, 1, 0, -3]$. La distance entre ces vecteurs contextes peut être évaluée au moyen de différentes mesures de distance. Sahlgren et Karlgren (2005) utilisent la mesure cosinus (Sahlgren & Karlgren, 2005).

Une version pondérée du Random Indexing a été proposée par (Gorman & Curran, 2006) et les auteurs l'utilisent pour mesurer la similarité sémantique entre phrases. Le vecteur contexte d'un mot y est calculé comme la somme pondérée des vecteurs index des contextes qui lui sont associés. Les auteurs comparent plusieurs fonctions de pondération dans une tâche d'extraction de synonymie : fréquence du mot dans le contexte, fréquence relative, *tf-idf, tf-idf*† (version log-pondérée du tf-idf), *DICE*, etc. Ils concluent à une nette amélioration des performances de RI en présence de grands corpus de données. Pour des ensembles de données réduits, RI est suffisamment robuste et la pondération n'a, au mieux, qu'un effet mineur. Ils constatent également une grande variabilité dans l'effet des fonctions poids utilisées et les bonnes performances de la fonction *tf-idf*†.

4 Calcul de la similarité entre phrases

Le calcul de la similarité entre phrases a été mis en œuvre en effectuant les étapes suivantes.

- Un étiqueteur syntaxique (en l'occurrence TreeTagger [3]) traite l'ensemble des articles de Wikipédia et convertit chacun de leurs termes en lemmes (*"travaille"* → *"travailler"*).
- Ensuite, le coefficient de pondération du *tf-icf* (Term Frequency-Inverse Corpus Frequency) (Reed *et al.*, 2006) de chaque terme (lemme) est calculé pour chaque article :

$$\textit{tf-icf}_{ij} = \log(1 + f_{ij}).\log(\frac{N+1}{n_i+1}) \quad (1)$$

  où f_{ij} est le nombre d'occurrences du terme d'indice i dans le document d'indice j, N le nombre total de documents d'un sous-corpus choisi suffisamment large et diversifié et n_i le nombre de documents où apparaît le terme d'indice i. Le coefficient *tf-icf* fournit une approximation du véritable *tf-idf* construit sur le corpus entier et il permet de traiter à moindre coût, des corpus dynamiques ou de très grande taille. Dans les expérimentations que nous présentons, nous avons considéré une version complète et statique de Wikipedia.
- L'ensemble des concepts est identifié avec celui des articles, chaque article définissant un concept et un concept n'existant que s'il existe un article qui le définit. Les valeurs du *tf-icf* d'un terme par rapport à l'ensemble des articles sont les composantes d'un vecteur appelé *vecteur sémantique de terme* dans la base des concepts.
- La valeur sémantique d'une phrase est calculée à partir des vecteurs sémantiques des termes qui la composent.

4.1 Calcul des vecteurs sémantiques

Un vecteur de terme est la représentation des liens entre ce terme et chacun des concepts, où l'ensemble des concepts est identifié à l'ensemble des articles de Wikipédia. Selon nos calculs, après avoir appliqué les étapes de prétraitement du corpus Wikipédia : filtrage du texte proprement dit, suppression des articles trop courts ou ayant un nombre trop faible

3. http ://www.cis.uni-muenchen.de/schmid/tools/TreeTagger

de liens, suivant les étapes suivies dans (Bawakid, 2011, p. 129), il reste 1 015 879 articles avec le Wikipédia français du 20/11/2013 et 3 766 589 articles avec le Wikipédia anglais du 02/12/2013.

Pour résoudre le problème de la réduction de dimension des vecteurs de terme, nous avons utilisé la méthode d'indexation aléatoire du Random Indexing décrite dans la section 3, en suivant les étapes ci-après.

- **Définition des vecteurs index de concept :** à chaque concept Wikipédia est attribué un vecteur index unique $\vec{c_j}$ dans un espace de dimension d fixée (cf. section 3). Étant donné le nombre de concepts des Wikipédia anglais et français, une dimension d de quelques milliers suffit pour assurer l'existence de vecteurs index presque orthogonaux.
- **Définition des vecteurs sémantiques de terme :** les vecteurs des termes présents dans le corpus Wikipédia sont calculés selon la formule (2).

$$\overrightarrow{terme_i} = \sum_{j=1}^{N} \textit{tf-icf}_{ij}.\overrightarrow{c_j} \tag{2}$$

  où N est le nombre de concepts Wikipédia, $\vec{c_j}$ est le vecteur index du concept j et *tf-icf*$_{ij}$, le *Term Frequency-Inverse Corpus Frequency* du terme d'indice i dans le document (concept) d'indice j calculé suivant la formule (1).

- **Similarité entre phrases :** pour calculer la similarité entre deux phrases, chacune d'elles doit d'abord être représentée comme un vecteur sémantique. On suppose que Wikipédia a une couverture des concepts et des mots suffisamment large pour contenir la plupart des termes sémantiquement significatifs utilisés dans les phrases en question. Le vecteur sémantique d'une phrase se calcule en faisant la somme des vecteurs sémantiques des termes qui la composent, suivant la formule (3).

$$\vec{S} = \sum_{i=1}^{n} \overrightarrow{terme_i}. \tag{3}$$

  Toutefois, cette mesure ne prend pas en considération le poids interne des mots dans le texte ou dans l'ensemble de textes d'où la phrase est extraite. L'hypothèse est que, si un mot est très fréquent dans les documents concernés, il convient de minimiser son importance au niveau de la phrase. Pour cela et conformément aux travaux de Neto et al., nous utilisons la pondération par le *tf-isf* (term frequency $\times$ inverse sentence frequency) (Neto *et al.*, 2000, 2002). Le *tf* est ici le nombre d'occurrences du terme dans la phrase et l'*isf* est calculé d'après la proportion de phrases dans l'ensemble des documents qui contiennent le terme :

$$\textit{tf-isf}_{is} = tf_{is}.\log(\frac{|S|}{SF_i}) \tag{4}$$

  où $|S|$ est le nombre de phrases et SF_i le nombre de phrases qui contiennent le terme d'indice i. Ainsi, l'importance d'un terme qui apparaît dans un grand nombre de phrases de l'ensemble des documents s'en trouve réduite.
  La sémantique d'une phrase est finalement représentée par une combinaison linéaire des vecteurs des termes qui la composent, pondérés par leurs *tf*s respectifs :

$$\vec{S_i} = \sum_{j=1}^{n} \textit{tf}_{ij}.\overrightarrow{terme_j}. \tag{5}$$

  La similarité entre deux phrases S_i et S_j dans un document (ou multi-document) est ensuite définie comme le cosinus de leurs vecteurs sémantiques respectifs [4] :

$$Sim_{WikiRI}(S_i, S_j) = \cos(\vec{S_i}, \vec{S_j}). \tag{6}$$

4.2 Nouveau calcul des vecteurs de termes

Nos premières expérimentations ayant donné des résultats décevants, nous avons analysé finement les mesures de similarités obtenues entre certains termes et groupements de termes pour mieux comprendre les insuffisances de la méthode. Des dysfonctionnements s'observent lorsque sont associés des termes qui diffèrent de par leur fréquence. Après avoir décrit le phénomène, nous proposons une modification dans le calcul des coordonnées des vecteurs de termes.

4. D'autres mesures ont été testées sans qu'une amélioration significative des résultats n'ait été constatée.

Les mots grammaticaux (*stop-words*) sont très fréquents dans les articles de Wikipédia, comme dans tous les textes écrits en langue française ou anglaise. Malgré leur importance pour la bonne compréhension d'un texte par ses lecteurs, ces termes ne sont pas pris en compte dans le calcul des vecteurs sémantiques.

Certains termes, que nous désignerons par *termes généraux*, ne sont pas des mots grammaticaux mais sont néanmoins très fréquents dans les articles de Wikipédia. La table 1 en donne quelques exemples pour la langue française, avec leur nombre d'occurrences dans Wikipédia, le pourcentage des articles dans lesquels ils apparaissent et la valeur de leur coefficient *icf*.

Terme	cf	Couverture	icf	Terme	cf	Couverture	icf
naître	298 963	29,60%	0,52	Lune	6 667	0,66%	2,18
pouvoir	293 035	29,01%	0,53	NASA	3 528	0,35%	2,45
grand	263 987	24,14%	0,58	peste	4 917	0,49%	2,31
nouveau	235 462	23,31%	0,63	sida	1 524	0,15%	2,82

TABLE 1 – Exemples de l'importance des termes généraux dans le Wikipédia français.

À l'inverse, un grand nombre de termes ont un nombre d'occurrences beaucoup plus faible. Il s'agit souvent de termes spécifiques à un domaine déterminé et qui sont essentiels pour une modélisation pertinente de la sémantique d'une phrase.

Ainsi, lorsque l'on évalue la similarité entre groupements de termes où sont associés un terme très fréquent avec un terme spécifique, on constate que l'influence du terme le plus fréquent écrase celui du terme spécifique. Par exemple, les lemmes *robot* et *infection* ont respectivement des *cf* relativement faibles, respectivement égaux à 5930 et 3593. À ce titre, ils peuvent être considérés comme des mots spécifiques. Par ailleurs, leur score de similarité (calculé comme le cosinus de leurs vecteurs de terme) est très faible (peu différent de 0,007). Or, les groupements de termes *petit robot/petite infection* obtiennent, avec le calcul de similarité défini précédemment, un score peu différent de 0,89, une valeur intuitivement beaucoup trop élevée, due à la prééminence du vecteur de termes *petit* sur les deux autres vecteurs de termes.

Autrement dit, bien que l'*icf* ait considérablement réduit le poids des termes généraux, la réduction qu'il opère n'est pas suffisante.

4.2.1 Modification des coordonnées des vecteurs de terme

L'objectif est donc de rééquilibrer le poids des termes très fréquents (mots généraux) par rapport à celui des termes plus rares, souvent spécifiques à un domaine donné, par rapport aux valeurs obtenues par le calcul classique du *tf-icf*. Pour ce, on introduit un paramètre $\alpha \geq 1$, destiné à renforcer le poids du *icf*, selon la formule (7).

$$\textit{tf-icf}_{\alpha} = tf.icf^{\alpha}, \tag{7}$$

Le paramètre α est estimé par apprentissage sur les ensembles de données SemEval-2012 TASK 6[5], choisies comme données d'entraînement pour le système.

Plus précisément, pour chacun des cinq ensembles de données SemEval-2012, nous avons calculé les similarités pour chaque paire de phrases, puis les scores obtenus par le système ont été comparés avec les similarités du "gold standard" qui sont fournies par SemEval-2012 pour obtenir les scores d'évaluations. Après avoir examiné les résultats obtenus avec différentes valeurs du paramètre α comprises entre 1 et 7, nous avons constaté que la valeur $\alpha = 3$ correspondait au meilleur résultat d'évaluation pour chacun des cinq corpus de Semeval-2012 testés.

Avec la valeur $\alpha = 3$, le calcul de la similarité des groupes de termes *petit robot* et *petite infection*, qui combinent des mots très généraux avec des mots moins fréquents, donne un résultat intuitivement acceptable, avec une valeur égale à 0,091.

4.2.2 Modification des vecteurs sémantiques de phrase

Les résultats sont améliorés par l'introduction du paramètre α. Cependant, cette modification du calcul des coordonnées des vecteurs sémantiques des termes agit sur la partie *icf* du *tf-icf* : elle ne fait donc que modifier la norme des vecteurs de

5. http ://www.cs.york.ac.uk/semeval-2012/task6/

termes. En particulier, elle ne résoud pas le caractère creux des vecteurs sémantiques des termes peu fréquents. En d'autres termes, ces derniers contiennent toujours principalement des coordonnées nulles. Conformément aux auteurs (Higgins & Burstein, 2007), les vecteurs des mots rares peuvent être enrichis en utilisant le vecteur centroïde du texte défini suivant la formule suivante.

$$\overrightarrow{centroid} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{terme_i}, \tag{8}$$

où n est le nombre de termes distincts dans le texte à calculer.

L'introduction dans le calcul du vecteur sémantique d'une phrase de son vecteur centroïde augmente l'apparition des coordonnées des vecteurs des termes rares et amoindrit le biais introduit par la fréquence des termes généraux. Le vecteur sémantique d'une phrase est finalement calculé en remplaçant la formule (3) par la formule (9).

$$\overrightarrow{S_i} = \sum_{j=1}^{n} tf_{ij}.(\overrightarrow{terme_j} - \overrightarrow{centroid}), \tag{9}$$

oàą $\overrightarrow{term_j}$ est le vecteur du terme d'indice j et n le nombre de termes distincts dans la phrase d'indice i.

5 Expérimentations et résultats

Les expérimentations ont été effectuées sur deux langues, l'anglais et le français.

D'après (Kanerva *et al.*, 2000) et étant donné la taille des corpus obtenus après les opérations de prétraitement, les vecteurs index ont été représentés dans des espaces de dimension $d = 5\ 000$ pour le Wikipédia français et $d = 10\ 000$ pour le Wikipédia anglais. Suivant les indications des mêmes auteurs, le nombre de composantes non nulles est fixé à $s = 20$ dans le premier cas et à $s = 26$ dans le second.

Les résultats rendus par le système WIKIRI ont été évalués en utilisant le coefficient de corrélation de Pearson entre les scores de système et les scores des annotateurs humains, comme il est habituel pour ce type de tâche.

5.1 Évaluation pour l'anglais

L'évaluation a été réalisée sur les données de la tâche 10 de **SemEval-2014** (Agirre *et al.*, 2014) qui contient 6 types de corpus à évaluer pour l'anglais :

1. **Discussion de forum** (deft-forum) : 450 paires de phrases.
2. **Discussion de l'actualité** (deft-news) : 300 paires de phrases.
3. **Titres de l'actualité** (headlines) : 750 paires de phrases.
4. **Descriptions d'images** (image) : 750 paires de phrases.
5. **Définitions extraites de OntoNotes et de WordNet** (OnWN) : 750 paires de phrases
6. **Titres et commentaires de nouvelles sur tweeter** (tweet-news) : 750 paires de phrases.

La table 2 présente une analyse comparative des corpus de Semeval où figurent leur nombre de mots (non grammaticaux) par phrase, leurs pourcentages d'adverbes, d'adjectifs, de noms communs, de noms propres, de verbes, ainsi que le pourcentage moyen de mots (non grammaticaux) communs entre les phrases des paires testées. Le faible pourcentage de noms propres dans certains corpus correspond au fait que le choix y a été fait de supprimer les majuscules. Par ailleurs, on peut également noter le très important pourcentage de mots qu'ont en commun les phrases testées.

SemEval fournit les "gold standard" des 6 corpus et un outil pour évaluer les systèmes. En 2014, 15 équipes ont participé à cette évaluation et les résultats de 38 systèmes ont été comparés. En utilisant la valeur de $\alpha = 3$ déterminée avec les corpus de SemEval-2012, notre système a obtenu les scores suivants : 47,005% avec deft-forum, 63,820% avec deft-news, 56,584% avec headlines, 75,884% avec image et 73,995% avec OnWN. La Figure 2 compare les résultats du système (en rose) avec ceux des systèmes qui ont participé à SemEval2014. WikiRI se place au-dessus de la moyenne des systèmes pour tous les corpus, à l'exception de celui concernant les titres de l'actualité.

Or, les meilleurs systèmes utilisent des corpus qui sont soit plus grands soit plus élaborés que Wikipédia, tels que Stanford WebBase Project (Kashyap *et al.*, 2014) ou des corpus de paraphrases (Sultan *et al.*, 2014). WikiRI obtient donc des

	Nb_Mots/Ph	ADV	ADJ	NC	NP	V	Communs/Ph
deft-news	11,8	1,9%	11,2%	33,7%	0%	14,8%	32,6%
headlines	6,3	0,7%	7,5%	25,3%	21,1%	11,6%	22,4%
images	5,8	0,4%	10,4%	30,8%	0,7%	9,5%	25,1%
OnWN	5,25	2%	6,2%	24,9%	0,2%	14,8%	25,2%
deft-forum	6,6	6%	5,6 %	16,8%	5,2%	19%	33%
tweet-news	7,4	2,2%	5,4%	18,7%	20,8%	11,1%	19%

TABLE 2 – Analyse comparative des différents corpus de tests de Semeval.

résultats tout à fait encourageants puisqu'il obtient des résultats au niveau de l'état de l'art en utilisant Wikipédia pour seule ressource.

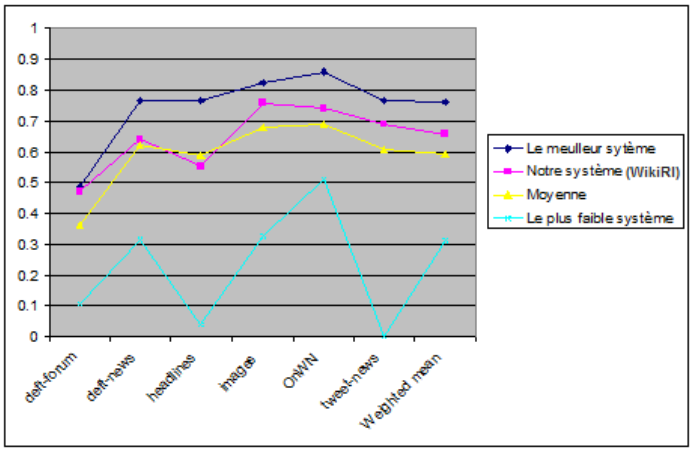

FIGURE 2 – Comparaison des résultats de WIKIRI avec ceux des systèmes proposés à SemEval-2014.

5.2 Évaluation pour le français

Si SemEval2014 contient des données pour l'anglais et pour l'espagnol, il n'existe pas de corpus annoté en français actuellement pour la tâche qui nous intéresse. Créer un tel corpus est un travail long et difficile : tester toutes les paires d'un ensemble de n phrases devient rapidement impraticable de par la croissance quadratique du nombre de paires en fonction de n. Nous avons extrait du Web deux corpus de textes français dans deux domaines différents définis respectivement par les mots-clefs "Épidémies" et "Conquête spatiale". Dans chaque corpus, nous avons sélectionné un ensemble de soixante-dix phrases, dont la longueur varie de 10 à 65 mots. Dix d'entre elles ont été choisies comme phrases de référence : elles contiennent diverses informations importantes concernant les domaines testés. Chacune de ces dix phrases a été associée à six autres phrases choisies de sorte que les différents niveaux de similarité entre phrases (sur une échelle de 0.0 à 4.0) soient représentés. La table 3 contient les mêmes indications que celles données pour le corpus Semeval : nombre de mots non grammaticaux par phrase, pourcentages d'adverbes, d'adjectifs de noms propres et de verbes, moyenne du nombre de mots non grammaticaux communs entre les phrases des paires testées. Ces données montrent que les phrases sont notablement plus longues que celles des corpus de Semeval, excepté celles du corpus *deft-news* ; par ailleurs, l'application visée étant le résumé multi-textes, le pourcentage de mots communs entre phrases est également beaucoup plus faible, notre échantillon se voulant représentatif de la tâche à laquelle devrait se confronter le système.

Sept volontaires humains, âgés de 18 à 60 ans, ont été impliqués dans la tâche d'annotation dont trois experts et quatre candides. Ils ont évalué la similitude des paires de phrases sur une échelle de 0,0 à 4,0 (les décimales étaient autorisées), selon les consignes indiquées dans la Table 4 et suivant la procédure d'annotation décrite dans (Li *et al.*, 2006).

La Table 5 donne l'une des phrases de référence (en gras) avec les phrases qui lui ont été associées. Les données du tableau correspondent à la moyenne des scores de similarité attribués par les sept annotateurs à chacune des six paires de phrases.

	Nb_Mots/Ph	ADV	ADJ	NC	**NP**	V	Communs/Ph
Epidémies	12,6	2,5%	10,9%	22,7%	**3,7%**	10%	9,7%
Conquête spatiale	16,1	2,2%	10,7%	21,4%	**8,1%**	11,4%	6,8%

TABLE 3 – Comparaison des corpus de tests *épidémies* et *conquête spatiale*.

4.0 : *Les phrases sont complètement équivalentes ;*
3.0 : *Les phrases sont globalement équivalentes, mais elles diffèrent par quelques détails ;*
2.0 : *Les phrases ne sont pas équivalentes, mais elles partagent certaines parties de l'information ;*
1.0 : *Les phrases ne sont pas équivalentes, mais elles traitent du même sujet ;*
0.0 : *Les phrases ne sont pas liées.*

TABLE 4 – Les instructions d'annotation pour le choix du score de similarité entre phrases

(1) ***Mars est l'astre le plus étudié du système solaire, puisque 40 missions lui ont été consacrées, qui ont confirmé la suprématie américaine - des épopées Mariner et Viking aux petits robots Spirit et Opportunity (2003 et 2004).***
(2) *Le 28 novembre 1964, la sonde Mariner 4 est lancée vers Mars, 20 jours après l'échec de Mariner 3.*
(3) *Les robots Spirit et Opportunity, lancés respectivement le 10 juin 2003 et le 8 juillet 2003 par la NASA, représentent certainement la mission la plus avancée jamais réussie sur Mars.*
(4) *Le bilan de l'exploration de Mars est d'ailleurs plutôt mitigé : deux tiers des missions ont échoué et seulement cinq des quinze tentatives d'atterrissage ont réussi (Viking 1 et 2, Mars Pathfinder et les deux MER).*
(5) *Le 6 août 2012, le rover Curiosity a atterri sur Mars avec 80 kg de matériel à son bord.*
(6) *Arrivé sur Mars en janvier 2004 comme son jumeau Spirit, et prévu comme lui pour fonctionner au moins trois mois, Opportunity (alias MER-B) roule encore et plusieurs de ses instruments répondent présents.*
(7) *Mars est mille fois plus lointaine que la Lune et son champ d'attraction plus de deux fois plus intense : la technologie n'existe pas pour envoyer un équipage vers Mars et le ramener sur Terre.*

Paires des phrases	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)	(1)-(6)	(1)-(7)
Score de similarité	0,49	2,06	1,86	1,19	1,57	1,1

TABLE 5 – Les scores de similarité d'une phrase de référence avec ses six phrases associées.

Les participants ont travaillé indépendamment et sans contrainte de temps sur une application Web [6] conçue pour leur faciliter la tâche d'annotation. Pour chaque phrase de référence choisie au hasard, ses phrases associées ont été aléatoirement et successivement présentées à l'annotateur. Ce dernier disposait d'un historique des scores de similarité qu'il avait déjà attribués et il était libre de les modifier à tout moment. Pour estimer l'accord inter-annotateurs, nous avons comparé les scores de chaque annotateur à la moyenne des scores calculée sur le reste du groupe. Les coefficients de corrélation ainsi obtenus sont présentés dans la table 6 [7]. Compris entre 0,8 et 0,941, ils indiquent que les évaluateurs humains sont largement d'accord sur les définitions utilisées dans l'échelle, même s'ils ont trouvé la tâche d'annotation particulièrement difficile.

Pour chacun des deux corpus, le système a été testé avec différentes valeurs du paramètres α. Les résultats ont été évalués à l'aide du coefficient de corrélation de Pearson, comme dans la tâche correspondante de SemEval. Ils sont donnés dans la première partie du tableau (lignes WikiRI) de la table 7. La deuxième partie du tableau contient les résultats obtenus avec un système précédemment implémenté (Vu *et al.*, 2014) inspiré de la méthode ESA (Gabrilovich & Markovitch, 2007), une variante du modèle GVSM. Chacun des corpus étant lié à un domane spécifique, un choix des concepts les plus pertinents basé sur l'étude des liens Wikipédia précédait la construction de la matrice termes×concepts. D'après Gottron *et all*, une réduction de dimension est d'autant plus efficace que l'on travaille dans un domaine spécifique (Gottron *et al.*, 2011).

6. http ://vuhaihieu-001-site1.smarterasp.net
7. Le choix de laisser les annotateurs utiliser des valeurs décimales ne permettait pas d'utiliser un kappa pour estimer l'accord.

Annotateurs	1	2	3	4	5	6	7
Corrélation (c. spatiale)	0,872	0,869	0,844	**0,941**	0,886	0,815	0,855
Standard Déviation (c. spatiale)	0,586	0,640	0,714	0,364	0,624	0,671	0,568
Corrélation (épidémies)	0,862	0,904	0,903	0,931	0,846	0,846	**0,800**
Standard Déviation (épidémies)	0,544	0,514	0,622	0,367	0,651	0,580	0,617

TABLE 6 – Les coefficients de corrélation entre les scores de chaque annotateur et la moyenne des scores des six autres.

WikiRI α	1	2	2,25	2,5	3	4	4,5	4,75	5
Epidémies	0,64788	0,79418	**0,79955**	0,79593	0,7754	0,72649	0,70075	0,68663	0,67148
Conquête spatiale	0,64811	0,74963	0,76075	0,77049	0,79191	0,83745	0,84846	**0,84943**	0,84696
ESA α	1	1,25	1,5	2	3,75	4	4,25	4,5	5
Epidémies	0,52541	**0,53926**	0,5388	0,51306	0,38146	0,36229	0,34119	0,31904	0,27683
Conquête Spatiale	0,55626	0,5611	0,56051	0,56563	0,61389	**0,61692**	0,61659	0,61241	0,59465

TABLE 7 – Les résultats du système pour les deux corpus en langue française suivant différentes valeurs du paramètre α.

Une première constatation est que les résultats obtenus par le système WikiRI, qui utilise l'ensemble de Wikipédia, sont très largement supérieurs à ceux obtenus par le système ESA pour des espaces de concepts limités à ceux des domaines considérés. Par ailleurs, l'introduction du paramètre α est plus efficace pour le système WikiRI que pour le système inspiré de la méthode ESA. Ces résultats sont conformes aux conclusions de Gordon et al. (Gorman & Curran, 2006) concernant l'influence des pondérations sur le système RI.

La seconde constatation est que, si la valeur optimale du paramètre α reste stable entre les différents corpus en langue anglaise de SemEval, il n'en est pas de même entre les deux corpus de domaine en langue française, puisque le meilleur résultat est obtenu avec $\alpha = 2,25$ pour le corpus *épidémies* et $\alpha = 4,75$ pour le corpus *conquêtes spatiales*. Néanmoins, l'introduction du paramètre s'avère très efficace : les résultats obtenus pour $\alpha = 1$, qui correspondent à l'utilisation du *tf-icf* classique, sont largement inférieurs à ceux obtenus pour les valeurs optimales (0,648 contre 0,800 et 0,648 contre 0,849). Par ailleurs, on constate la même variabilité de la valeur optimale de α pour le système inspiré de la méthode ESA que pour le système WikiRI.

Il est actuellement difficile de savoir si cette instabilité constatée du α optimal est imputable à la langue ou à la nature même des corpus que nous avons volontairement choisis très différents. D'après les données de la table 3, la principale différence concerne les noms propres (NP), presque trois fois plus fréquents dans le corpus *conquête spatiale* que dans le corpus *épidémies*. Dans ce second corpus en effet, les termes spécifiques au domaine sont souvent des noms communs : *peste, choléra, vaccin, bacille, virus,* etc. alors qu'ils concernent plus fréquemment des hommes ou des engins spatiaux dans le premier : *Gagarine, Curiosity, Spoutnik, Armstrong,* etc. Cependant, des expérimentations supplémentaires seront nécessaires pour pouvoir mieux comprendre la relation qui peut exister entre le choix du meilleur α et la nature du corpus.

6 Conclusion et perspectives

Nous avons présenté une méthode de modélisation de la sémantique d'un mot ou d'un texte basée sur l'utilisation de Wikipédia, qui utilise la technique d'indexation aléatoire RI pour réduire la dimension des espaces vectoriels de représentation. Par ailleurs, des modifications ont été introduites dans le calcul des vecteurs représentant les termes et les phrases pour réduire le bruit que peut engendrer la multiplicité des concepts dans une ressource linguistique aussi foisonnante. La technique d'indexation aléatoire a montré son efficacité dans la réduction de la complexité des calculs, mais elle semble très sensible au choix des pondérations utilisées. Les résultats obtenus sur les données de SemEval2014 pour l'anglais sont au niveau de l'état de l'Art, ce qui prouve l'efficacité de l'approche. Testée également sur la langue française, la méthode donne des résultats très encourageants, même si des expérimentations supplémentaires sont nécessaires pour mieux comprendre l'influence du paramètre α que nous avons introduit. Elle offre l'avantage d'être utilisable pour d'autres langues, à la condition d'y disposer de ressources Wikipédia suffisamment développées. Si le choix a été fait d'utiliser la totalité de Wikipédia, la question reste ouverte de savoir quel est le nombre minimal de documents qui pourrait assurer une qualité suffisante à la détermination des vecteurs de termes.

Nos travaux actuels cherchent à utiliser les similarités entre phrases pour implémenter une méthodologie de résumés multi-textes. Pour l'anglais, le système sera testé sur les données DUC. Pour le français, il utilisera les données du corpus rpm2 (de Loupy *et al.*, 2010).

Références

AGIRRE E., BANEA C., CARDIE C., CER D., DIAB M., GONZALEZ-AGIRRE A., GUO W., MIHALCEA R., RIGAU G. & WIEBE J. (2014). Semeval-2014 task 10 : Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 81–91, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

BALASUBRAMANIAN N., ALLAN J. & CROFT W. B. (2007). A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 813–814 : ACM.

BAWAKID A. (2011). *Automatic Documents Summarization Using Ontology based Methodologies.* PhD thesis, University of Birmingham.

BURIOL L. S., CASTILLO C., D. D., S. L. & S. M. (2006). Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence.*, p. 45–51.

CAPOCCI A., SERVEDIO V., COLAIORI F. & BURIOL L. (2006). Preferential attachment in the growth of social networks : the case of wikipedia. *Arxiv preprint physics*.

CARBONELL J. G., YANG Y., FREDERKING R. E., BROWN R., GENG Y. & D. L. (1997). Translingual information retrieval : a comparative evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97 Distinguished Paper Award)*.

CHAN P., HIJIKATA Y. & NISHIDA S. (2013). Computing semantic relatedness using word frequency and layout information of wikipedia. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, p. 282–287 : ACM.

CHAVEZ A., DÁVILA H., GUTIÉRREZ Y., FERNÁNDEZ-ORQUÍN A., MONTOYO A. & MUÑOZ R. (2014). Umcc_dlsi_semsim : Multilingual system for measuring semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 716–721, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

DE LOUPY C., GUÉGAN M., AYACHE C., SENG S. & MORENO J.-M. T. (2010). A french human reference corpus for multi-document summarization and sentence compression. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, **22**(1), 457–479.

GABRILOVICH E. & MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, p. 1606–1611.

GILES J. (2005). Internet encyclopedias go head to head. *Nature*, **438**, 900–901.

GORMAN J. & CURRAN J. R. (2006). Random indexing using statistical weight functions. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, p. 457–464, Stroudsburg, PA, USA : Association for Computational Linguistics.

GOTTRON T., ANDERKA M. & STEIN B. (2011). Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, p. 1961–1964 : ACM.

HADJ TAIEB M. A., BEN AOUICHA M. & BEN HAMADOU A. (2013). Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, **50**, 260–278.

HIGGINS D. & BURSTEIN J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, p. 1–12.

HIRAO T., OKUMURA M. & ISOZAKI H. (2005). Kernel-based approach for automatic evaluation of natural language generation technologies : Application to automatic summarization. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 145–152 : Association for Computational Linguistics.

JAAP K. & MARIJN K. (2009). Is wikipedia link structure different ? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, p. 232–241.

KANERVA P. (1988). *Sparse distributed memory*. MIT Press.

KANERVA P., KRISTOFERSSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, volume 1036 : Erlbaum.

KASHYAP A., HAN L., YUS R., SLEEMAN J., SATYAPANICH T., GANDHI S. & FININ T. (2014). Meerkat mafia : Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 416–423, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

KO Y., PARK J. & SEO J. (2002). Automatic text categorization using the importance of sentences. In *COLING*.

KUMAR C. A. (2009). Analysis of unsupervised dimensionality reduction techniques. *Comput. Sci. Inf. Syst.*, **6**(2), 217–227.

LI Y., MCLEAN D., BANDAR Z. A., O'SHEA J. D. & CROCKETT K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, **18**(8), 1138–1150.

LIN C.-Y. (2004). Rouge : a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*, p. 25–26.

NAKAYAMA K., HARA T. & NISHIO S. (2008). Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology.

NETO J. L., FREITAS A. A. & KAESTNER C. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, p. 205–215. Springer.

NETO J. L., SANTOS A. D., KAESTNER C. A. & FREITAS A. A. (2000). Generating text summaries through the relative importance of topics. In *Advances in Artificial Intelligence*, p. 300–309. Springer.

REED J. W., JIAO Y., POTOK T. E., KLUMP B. A., ELMORE M. T. & HURSON A. R. (2006). TF-ICF : A new term weighting scheme for clustering dynamic data streams. In M. A. WANI, T. LI, L. A. KURGAN, J. YE & Y. LIU, Eds., *The Fifth International Conference on Machine Learning and Applications, ICMLA 2006, Orlando, Florida, USA, 14-16 December 2006*, p. 258–263 : IEEE Computer Society.

SAHLGREN M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5.

SAHLGREN M. & KARLGREN J. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Journal of Natural Language Engineering*, **11**(3). Special Issue on Parallel Texts.

STRUBE M. & PONZETTO S. P. (2006). Wikirelate ! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, p. 1419–1424.

SULTAN M. A., BETHARD S. & SUMNER T. (2014). Dls@cu : Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 241–246, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

VOSS J. (2005). Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

VU H. H., VILLANEAU J., SAÏD F. & MARTEAU P. (2014). Sentence similarity by combining explicit semantic analysis and overlapping n-grams. In P. SOJKA, A. HORÁK, I. KOPECEK & K. PALA, Eds., *Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, volume 8655 of *Lecture Notes in Computer Science*, p. 201–208 : Springer.

WILLIAM B. & LINDENSTRAUSS J. (1984). Extensions of lipschitz mappings into a hilbert space. In *Conference in Modern Analysis and Probability.*