# Une mesure de Similarité Sémantique basée sur la Recherche d'Information

Davide Buscaldi[1]

Laboratoire d'Informatique de Paris Nord,
CNRS, (UMR 7030)
Université Paris 13, Sorbonne Paris Cité,
F-93430, Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr

**Résumé** : Dans cet article, nous décrivons une mesure de similarité sémantique basée sur la recherche d'information qui a été utilisée dans le "shared task" SemEval 2013. Habituellement, la similarité sémantique est utilisée pour améliorer la performance des systèmes de recherche d'information. Nous avons essayé de faire le contraire : deux textes à comparer sont traités comme des requêtes et on compare les listes des résultats de la recherche afin d'établir la similarité sémantique des textes. Les résultats préliminaires obtenus dans le "shared task" SemEval 2013 montrent que cette mesure a fonctionné mieux que les mesures de similarité sémantique basées sur WordNet et des mesures classiques comme le cosinus ou la distance d'édition. **Mots-clés** : Similarité Sémantique. Recherche d'Information.

## 1 Introduction

La détermination du niveau de similarité sémantique entre textes est une tâche très importante pour différentes applications dans le contexte du Traitement Automatique de la Langue (TAL). Une des premières applications de similitude de texte est, peut-être, le modèle d'espace vectoriel utilisé dans la Recherche d'Information (RI), où il faut classer des documents dans une collection par ordre de pertinence par rapport à une requête en entrée (Salton (1971)). La pertinence est calculée en fonction d'une mesure de similarité entre la requête et le document, dont le contenu est représenté par un vecteur de mots. Les mesures de similarité de texte ont été également utilisées pour la désambiguïsation lexicale (Lesk (1986)) et le résumé automatique de texte (Lin & Hovy (2003)), entre autres.

Très récemment, une évaluation comparative pour la tâche de similarité sémantique a été proposée en tant que "pilot task" au SemEval 2012 (Semantic Textual Similarity, STS) par Agirre *et al.* (2012) et confirmée pour SemEval 2013, élevée ainsi au rang de "shared task" du *SEM2013 [1], dans le but d'établir un cadre d'évaluation qui favorise le développement de nouveaux systèmes et mesures de similarité sémantique. L'objectif final des campagnes d'évaluation SemEval est de promouvoir la recherche dans tous les aspects de la sémantique computationnelle et d'améliorer les performances dans toutes les tâches où il est nécessaire d'analyser sémantiquement un texte de façon automatique. Il s'agit de déterminer, sur une échelle de $0$ (aucune similarité) à $5$ (même signification), la similarité sémantique entre couples de phrases. Cette particularité distingue la tâche STS de la tâche de détection de paraphrases, dont la réponse à la question si deux phrases sont similaires est une réponse binaire.

Dans cet article, nous présentons une mesure de similarité sémantique qui utilise la RI pour déterminer la similarité entre textes ; cette mesure est basée sur l'hypothèse que, si on utilise deux textes en tant que requêtes d'entrée dans un même système de RI, alors leur similarité dépend de la similarité des listes des documents récupérés par le système. Cette mesure a été vérifiée expérimentalement dans le cadre de la participation de l'équipe RCLN du LIPN au STS du SemEval 2013 (Buscaldi *et al.* (2013)), où elle s'est révélée la mesure la plus efficace, parmi celles mis en œuvre par le système LIPN. La suite du papier est articulée comme suit : dans la Section 2, nous présentons des travaux connexes. Dans la Section 3, nous décrivons notre mesure de similarité basée sur la RI. Dans la Section 4, nous rendons compte de l'expérimentation et de l'évaluation menées. Enfin, dans la Section 5, nous concluons et nous proposons quelques perspectives.

## 2 Travaux connexes

Traditionnellement, les mesures de similarité sémantique ont été définies sur des mots ou des concepts, mais non sur des fragments de texte. La dérivation d'une mesure de similarité sémantique compositionnelle à partir d'une mesure de similarité sémantique lexicale (c'est à dire, déterminer la similarité entre deux phrases à partir des mots qui les composent) n'est pas triviale. Par exemple, les phrases « un chien mord un homme » et « un homme mord un chien » ont la même similarité sémantique mot-à-mot mais sont très différentes globalement. L'accent mis sur les mesures de

1. `http://clic2.cimec.unitn.it/starsem2013/`

similarité mot-à-mot est probablement dû à la disponibilité de ressources sémantiques telles que WordNet ou des ontologies qui codent des relations concept-à concept. Pedersen *et al.* (2004) a implémenté et mis à disposition un *package*, WordNet :Similarity [2], qui code différentes mesures de similarité mot-à-mot, souvent utilisées dans plusieurs applications de TAL, par exemple la détection de l'implication textuelle (Harabagiu & Hickl (2006)) ou la résolution de coréférences (Ponzetto & Strube (2006)). La plupart des systèmes qui participent à SemEval STS ont utilisé des mesures classiques, modèle vectoriel, des extensions du modèle de n-grammes (Buscaldi *et al.* (2012)), la distance d'édition, ou des combinaisons de mesures de similarité sémantique lexicale (Banea *et al.* (2012); Bär *et al.* (2012); Šarić *et al.* (2012)). Les meilleurs résultats ont été obtenus en intégrant les différentes mesures avec des méthodes de régression linéaire (Schölkopf *et al.* (1999)).

## 3 Mesure de Similarité Basée sur la Recherche d'Information

Étant donnés deux textes $p$ and $q$, un système de RI $S$ et une collection de documents $D$ indexé par $S$, cette mesure est basée sur l'hypothèse que $p$ et $q$ sont sémantiquement similaires si les documents récupérés par $S$ pour les deux textes utilisés en tant que requêtes d'entrée, sont classés de façon similaire.

Soient $L_p = \{d_{p_1}, \ldots, d_{p_K}\}$ et $L_q = \{d_{q_1}, \ldots, d_{q_K}\}$, $d_{x_i} \in D$ les ensembles des premiers $K$ documents récupérés par $S$ pour $p$ et $q$, respectivement. Soient $s_p(d)$ et $s_q(d)$ les scores assignés par $S$ au document $d$ pour les requêtes $p$ et $q$, respectivement. Le score de similarité est calculé de la façon suivante :

$$sim_{IR}(p,q) = 1 - \frac{\sum\limits_{d \in L_p \cap L_q} \frac{\sqrt{(s_p(d) - s_q(d))^2}}{\max(s_p(d), s_q(d))}}{|L_p \cap L_q|} \quad (1)$$

si $|L_p \cap L_q| \neq \emptyset$, 0 en cas contraire.

La valeur optimale de $K$ a été déterminée avec des expériences menées sur le jeux de données de test du SemEval-2012, composée par $3108$ couples de phrases en anglais, avec des jugements de similarité gradués entre $0$ et $5$. Le score de similarité calculé avec la formule (1) est toujours compris entre $0$ et $1$, donc pour pouvoir calculer la corrélation avec le gold standard (jugements de similarité), le score calculé pour chaque couple de phrases

2. `http://wn-similarity.sourceforge.net/`

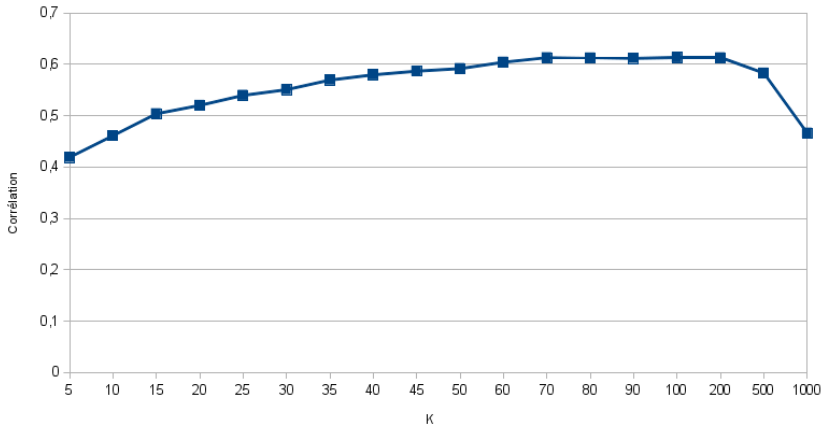


FIGURE 1 – Coefficient de corrélation de Pearson pour différentes valeurs de $K$, calculés sur le jeu de données SemEval 2012.

a été multiplié par $5$. Les résultats en Figure 1 montrent que la valeur optimale se situe entre $70$ et $100$. Nous avons utilisé une valeur $K = 80$

## 4 Expérimentation

La mesure de similarité a été testée sur les jeux de données SemEval-2012 et SemEval-2013, et comparée avec des mesures de similarité habituellement utilisées dans la tâche STS, décrites dans la Section 4.1. La tâche STS consiste à comparer deux phrases et déterminer leur similarité sémantique sur une échelle de $0$ à $5$. Chaque couple de phrases a été jugé par $4$ experts, le jugement de similarité du gold standard est la moyenne de tous les $4$ jugements. La Figure 2 présente des exemples de phrases à comparer, avec le jugement de corrélation donné par les experts.

La collection de documents utilisée pour les expériences est en anglais et composée par la collection AQUAINT-2 [3] et la collection NTCIR-8 [4]. Le moteur de recherche utilisé est Lucene [5], plus précisément dans sa version 4.2, et la mesure de similarité utilisée pour calculer la similarité entre les requêtes et les documents est la mesure BM25 (Jones *et al.* (2000)). La

3. `http://www.nist.gov/tac/data/data_desc.html#AQUAINT-2`
4. `http://metadata.berkeley.edu/NTCIR-GeoTime/ntcir-8-databases.php`
5. `http://lucene.apache.org/core`

| Jugement | Phrase 1 | Phrase 2 |
|---|---|---|
| 4.8 | Tehran: Discussions with Venezuelan President Hugo Chavez | city of tehran : conversations with president of venezuela hugo chavez |
| 3.8 | Anti-Putin protesters form human chain in Moscow | Anti-Putin protesters form human chain |
| 2.6 | Egyptians vote to pick president for first time | Egyptians wait for key presidential vote results |
| 1 | Dozens killed in Nigerian riots | Dozens killed in Kenyan clashes |
| 0 | 20-member parliamentary, trade team leaves for India | Commerce secretary to take leave of absence |

FIGURE 2 – Examples de phrases de la tâche STS.

valeur $K$ a été mis à 20 pour les expériences du SemEval-2013, pour des raisons de rapidité (il y a un temps limite pour produire les résultats).

## 4.1 Mesures de similarité pour comparaison

### 4.1.1 Cosinus

Soient $\mathbf{p} = (w_{p_1}, \ldots, w_{p_n})$ et $\mathbf{q} = (w_{q_1}, \ldots, w_{q_n})$ les vecteurs de poids $tf.idf$ associés aux phrases $p$ et $q$, alors la distance cosinus est :

$$sim_{cos}(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{p_i} \times w_{q_i}}{\sqrt{\sum_{i=1}^{n} {w_{p_i}}^2} \times \sqrt{\sum_{i=1}^{n} {w_{q_i}}^2}} \tag{2}$$

Où la valeur idf (inverse document frequency) a été calculée sur Google Web 1T.

### 4.1.2 Distance d'édition

La distance d'édition a été calculé de la façon suivante :

$$sim_{ED}(p, q) = 1 - \frac{Lev(p, q)}{\max(|p|, |q|)} \tag{3}$$

où $Lev(p, q)$ est la distance de Levenshtein entre $p$ et $q$, au niveau des caractères.

### 4.1.3 Mesure de similarité basée sur WordNet

Cette mesure de distance a été introduite pour Buscaldi *et al.* (2012). Soient $C_p$ et $C_q$ les ensembles de concepts contenus dans les phrases $p$ et $q$, avec $|C_p| \geq |C_q|$, alors la similarité entre $p$ et $q$ est me résultat de :

$$sim_{WN}(p,q) = \frac{\sum\limits_{c_1 \in C_p} \max\limits_{c_2 \in C_q} s(c_1, c_2)}{|C_p|} \tag{4}$$

où $s(c_1, c_2)$ est une mesure de similarité conceptuelle calculée sur la hiérarchie de WordNet, selon la formulation de Wu & Palmer (1994).

## 4.2 Mesure de similarité basée sur N-grammes

Cette mesure a été proposée par Buscaldi *et al.* (2009) pour la tâche de recherche de passages dans le contexte des systèmes question-réponse. La similarité entre $p$ et $q$ est calculée de la façon suivante :

$$sim_{ngrams}(p,q) = \frac{\sum\limits_{\forall x \in Q} h(x,P) \frac{1}{d(x, x_{max})}}{\sum_{i=1}^{n} w_i} \tag{5}$$

où $w_i$ est le poids $idf$ du mot $t_i$, $P$ est l'ensemble des n-grammes composés par les mots de $p$ qui sont aussi dans $q$ , $Q$ est l'ensemble de tous les n-grammes possibles dans $q$, et $n$ le nombre de mots dans la phrase la plus longue. $\frac{1}{d(x,x_{max})}$ est un facteur de distance qui réduit le poids des n-grammes en fonction de la distance du plus grand n-gramme. Le poids pour un n-gramme est calculé comme la somme des poids des mots qui le composent :

$$h(x, P_j) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in P_j \\ 0 & \text{autrement} \end{cases} \tag{6}$$

où $w_k$ est le poids du *k-ième* mot et *j* la taille du n-gramme $x$ ;

## 4.3 Évaluation

Le coefficient de corrélation de Pearson a été calculé pour chaque mesure, sur les jeux de données de test du SemEval-2012 (Table 1) et du SemEval-2013 (Table 2). Le jeu de données de test SemEval-2012 est composé par différents sous-ensembles de phrases extraites de différents

corpus : *MSRpar* est composé par des couples de phrases extraites du corpus de paraphrases de Microsoft[6] ; *MSRvid* est un corpus composé par des annotations de vidéos ; *OnWN* est composé par des définitions de WordNet ; *Europarl* est composé par des transcriptions de traductions automatiques des sessions du parlement européen ; *News*, *Headlines* sont des corpus composés par des titres d'actualités. Le jeu de données SemEval-2013 contient aussi des définitions de frames de FrameNet (*FNWN*) et des phrases issues de l'évaluation automatique de systèmes de traduction automatique (*SMT*). La Table 4 présente la taille, en nombre moyen de mots pour chaque phrase, dans chaque jeu de données.

| | **MSRpar** | **MSRvid** | **OnWN** | **Europarl** | **News** | **ALL** |
|---|---|---|---|---|---|---|
| $sim_{ngrams}$ | 0.419 | 0.543 | 0.453 | 0.505 | 0.408 | 0.412 |
| $sim_{WN}$ | 0.380 | 0.784 | **0.507** | 0.556 | 0.426 | 0.609 |
| $sim_{ED}$ | 0.251 | 0.290 | **0.507** | **0.625** | 0.426 | 0.300 |
| $sim_{cos}$ | **0.468** | 0.688 | 0.458 | 0.556 | 0.349 | 0.513 |
| $sim_{IR}$ | 0.167 | **0.785** | 0.359 | 0.584 | **0.523** | **0.613** |

TABLE 1 – Coefficient de corrélation de Pearson, calculé sur le test set SemEval 2012.

| | **FNWN** | **Headlines** | **OnWN** | **SMT** | **ALL** |
|---|---|---|---|---|---|
| $sim_{ngrams}$ | 0.285 | 0.532 | 0.459 | 0.280 | 0.336 |
| $sim_{WN}$ | **0.395** | **0.606** | 0.552 | 0.282 | 0.477 |
| $sim_{ED}$ | 0.220 | 0.536 | 0.089 | **0.355** | 0.230 |
| $sim_{cos}$ | 0.306 | 0.573 | 0.541 | 0.244 | 0.382 |
| $sim_{IR}$ | 0.067 | 0.598 | **0.628** | 0.241 | **0.541** |

TABLE 2 – Coefficient de corrélation de Pearson, calculé sur le test set SemEval 2013.

La Table 3 présente les résultats du test d'ablation mené sur le système utilisé pour la participation au SemEval-2013, où chaque mesure a été utilisée en tant que caractéristique pour entraîner un modèle de régression linéaire.

Dans tous les tests on peut observer que la mesure basée sur la RI, en moyenne, s'est montrée meilleure, surtout pour les corpus basés sur les

6. `http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/`

| **Mesure enlevée** | **Corrélation** | **Perte** |
|---|---|---|
| *aucune* | 0.597 | 0 |
| $sim_{ngrams}$ | 0.596 | 0.10% |
| $sim_{WN}$ | 0.563 | 3.39% |
| $sim_{ED}$ | 0.584 | 1.31% |
| $sim_{cos}$ | 0.596 | 0.10% |
| $sim_{IR}$ | 0.510 | **8.78%** |

TABLE 3 – Test d'ablation sur différentes mesures utilisées dans le test set SemEval-2013.

| **corpus** | **# moyen de mots** |
|---|---|
| MSRpar | 17.71 |
| MSRvid | 6.63 |
| OnWN 2012 | 7.5 |
| Europarl | 10.7 |
| News | 11.72 |
| FNWN | 19.92 |
| Headlines | 7.21 |
| OnWN 2013 | 7.17 |
| SMT | 26.39 |

TABLE 4 – Nombre moyen de mots pour chaque phrase, dans chaque corpus.

actualités (c'est à dire, dans le même domaine du corpus indexé par le moteur de recherche), même si pour certains sous-ensembles de données elle a obtenu des résultats non satisfaisants, en particulier sur les données MSRpar et FNWN. On a supposé que la taille moyenne des phrases joue un rôle dans la qualité des résultats ; apparemment, c'est le cas pour MSRpar et FNWN qui ont une taille en moyenne double que la plupart des autres corpus (Table 4), mais ce n'est pas si évident dans le cas du corpus SMT, où la perte de performances ne semble pas en corrélation avec la taille des phrases.

## 5 Conclusion et perspectives

Dans ce papier, nous avons présenté une mesure de similarité sémantique entre deux textes à partir d'une collection de documents indexé avec un système de RI. L'évaluation de cette mesure sur les jeux de données de la

tâche SemEval STS 2012 et 2013 montre qu'elle obtient, en moyenne, des meilleures valeurs de corrélation par rapport à d'autres mesures de similarité entre textes, basées soit sur des caractéristiques de surface du texte, soit sur la similarité conceptuelle entre concepts. Cependant, cette mesure montre des limites, surtout si la taille des textes à comparer augmente, et si le corpus de documents indexés n'est pas dans le même domaine que les textes à comparer. Nous souhaitons, en perspective, vérifier si la mesure est indépendante du modèle de RI utilisé par le moteur de recherche, et implémenter la mesure sur le web, pour résoudre le problème lié au domaine du corpus indexé. Dans ce cas, un problème additionnel à résoudre est constitué par le poids à donner à chaque document récupéré.

## Remerciements

Le travail présenté a été en partie soutenu par le LabEx EFL (Empirical Foundations of Linguistics) et par le projet OSEO Quaero.

## Références

AGIRRE E., CER D., DIAB M. & GONZALEZ A. (2012). A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantcis (*SEM 2012)*, Montreal, Quebec, Canada.

BANEA C., HASSAN S., MOHLER M. & MIHALCEA R. (2012). UNT : A Supervised Synergistic Approach to Semantic Text Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, p. 635–642, Montreal, Canada.

BÄR D., BIEMANN C., GUREVYCH I. & ZESCH T. (2012). UKP : Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, p. 435–440, Montreal, Canada.

BUSCALDI D., ROSSO P., GÓMEZ J. M. & SANCHIS E. (2009). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems (JIIS)*, **34**(2), 113–134.

BUSCALDI D., ROUX J. L., FLORES J. G. G. & POPESCU A. (2013). LIPN-CORE : Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, GA, USA. to appear.

BUSCALDI D., TOURNIER R., AUSSENAC-GILLES N. & MOTHE J. (2012). Irit : Textual similarity combining conceptual similarity with an n-gram comparison method. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Quebec, Canada.

DUDOGNON D., HUBERT G. & RALALASON B. (2010). Proxigénéa : Une mesure de similarité conceptuelle. In *Proceedings of the Colloque Veille Stratégique Scientifique et Technologique (VSST 2010)*.

HARABAGIU S. & HICKL A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, p. 905–912, Stroudsburg, PA, USA : Association for Computational Linguistics.

JONES K. S., WALKER S. & ROBERTSON S. E. (2000). A probabilistic model of information retrieval : development and comparative experiments part 2. *Inf. Process. Manage.*, **36**(6), 809–840.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, p. 24–26, New York, NY, USA : ACM.

LIN C.-Y. & HOVY E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 71–78, Stroudsburg, PA, USA : Association for Computational Linguistics.

PEDERSEN T., PATWARDHAN S. & MICHELIZZI J. (2004). WordNet : :Similarity : measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, p. 38–41, Stroudsburg, PA, USA : Association for Computational Linguistics.

PONZETTO S. P. & STRUBE M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, p. 192–199, Stroudsburg, PA, USA : Association for Computational Linguistics.

SALTON G. (1971). *The SMART Retrieval System&#8212 ;Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

SCHÖLKOPF B., BARTLETT P., SMOLA A. & WILLIAMSON R. (1999). Shrinking the tube : a new support vector regression algorithm. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, p. 330–336, Cambridge, MA, USA : MIT Press.

ŠARIĆ F., GLAVAŠ G., KARAN M., ŠNAJDER J. & BAŠIĆ B. D. (2012). TakeLab : Systems for Measuring Semantic Text Similarity. In *Proceedings of*

*the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, p. 441–448, Montreal, Canada.

WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, p. 133–138, Stroudsburg, PA, USA : Association for Computational Linguistics.