Une mesure de similarité entre phrases basée sur des noyaux sémantiques

Samir Amir*, Adrian Tanasescu*, Djamel A. Zighed*

*Institut des Sciences de l'Homme, 14 avenue Berthelot,
69007 Lyon, France, prenom.nom@ish-lyon.cnrs.fr,

Résumé. Nous proposons une nouvelle approche pour le calcul de similarité sémantique entre phrases en utilisant les noyaux sémantiques qui les composent. Ces noyaux, sous la forme de triplets (sujet, verbe et objet) sont supposés porteurs de l'information des phrases dont ils sont extraits. Sur la base de la comparaison sémantique de noyaux, on extrait un ensemble d'indicateurs descriptifs. Nous utilisons ensuite un apprentissage automatique, sur un benchmark contenant des phrases dont la similarité sémantique a été évaluée par des experts humains, afin de déterminer l'importance de chaque indicateur et de construire ainsi un modèle capable de fournir une mesure de similarité sémantique entre phrases. Les expérimentations et les études comparatives, effectuées avec d'autres approches permettant l'estimation des similarités sémantiques entre phrases, montrent les bonnes performances de notre approche. En se basant sur cette dernière, un outil de navigation sémantique est en cours de développement.

1 Introduction

Avec l'émergence de l'internet et les différents moyens de son utilisation, la capacité de déterminer la similarité sémantique entre textes s'est avérée utile pour un large nombre d'applications. Le calcul de similarité entre textes en utilisant les méthodes classiques issues du domaine de la recherche d'information et du traitement automatique de la langue présente certaines limitations, notamment pour les textes courts (phrases). Cela est du au fait que le nombre de mots communs ou leur cooccurrence n'est pas toujours une information pertinente sur le plan statistique. Dans ce contexte, plusieurs mesures de similarité entre phrases ont été proposées. Cependant, due à la complexité du langage humain, des améliorations sont toujours requises.

Deux phrases peuvent avoir une sémantique équivalente (ou proche) avec une structure syntaxique et lexicale complètement hétérogène en terme de nombre et de position de mots, ainsi qu'en termes de relations grammaticales. Par exemple, selon le benchmarks décrits dans (O'shea et al. (2014)), les deux phrases "*I am so hungry I could eat a whole horse plus dessert*." et "*I could have eaten another meal, I'm still starving*." Ont une forte similarité évaluée dans ce benchmark à 0.77 alors qu'on peut clairement voir que leurs structures présentent des hétérogénéités sur plusieurs niveaux.

Afin de faire face au défi décrit précédemment, nous proposons une approche basée sur des noyaux sémantiques. Dans la prochaine section, nous décrivons cette approche plus en détails

dans. L'évaluation expérimentale est montrée dans la Section 3. La Section 4 conclut le travail tout en rappelant les avancées réalisées dans cet article par rapport à l'état de l'art et explique les travaux futurs.

2 Mesurer la similarité entre phrases à partir des noyaux sémantiques

Nous proposons d'utiliser les noyaux sémantiques des phrases pour le calcul de la similarité sémantique entre ces dernières. Ces noyaux sont des triplets composés de sujets, verbes et objets. Il est évident que la sémantique de la phrase réside dans tous ses composants et peut-être pas seulement dans son noyau *(sujet, verbe, objet)*. Cependant, selon plusieurs études, les noyaux sémantiques sont porteurs de la plus grande partie de la sémantique de la phrase (Heidinger (1984)). Cette hypothèse nous permet de nous concentrer sur la partie essentielle des phrases lors de l'estimation de leur similarité sémantique. L'approche que nous décrivons dans la suite de cet article commence par l'extraction des noyaux sémantique. Ensuite, nous calculons la similarité sémantique entre les différents éléments de ces noyaux. Ainsi, un ensemble d'indicateurs agrégeant les similarités entre éléments des noyaux est extrait. Enfin, pour chaque indicateur, nous déterminons le coefficient relatif en appliquant un apprentissage automatique sur une vérité terrain.

2.1 Similarité sémantique entre les éléments des noyaux

Dans cette étape, l'analyseur *Stanford Parser* est utilisé pour l'extraction des relations grammaticales. Ce dernier est capable de détecter plus de 50 types de relation. Pour chaque relation, nous avons défini un ensemble de règles permettant de parcourir l'arbre grammatical et d'en construire l'ensemble de noyaux contenus dans la phrase en question. Ensuite, des similarités sémantiques sont calculées entre les noyaux appartenant aux deux phrases. Plus précisément, entre les différents éléments ayant le même règle grammatical. A savoir, entre les sujets, les verbes et les objets appartenant aux différents noyaux. Pour ce faire, nous utilisons la distance de Hirst and St-Onge (HSO)(Hirst et St Onge (1998)). HSO calcule la similarité entre deux termes en fonction de leurs positions dans l'ontologie. Notre choix pour cette distance était motivé par l'étude comparative effectuée dans (Budanitsky et Hirst (2006)).

D'autres ressources linguistiques telle que DBpedia sont également sollicitées dans le cas où l'information recherchée n'est pas disponible dans WordNet. Enfin, la distance de Jaro-Winkler (Winkler (1999)) est utilisée dans le cas où un des termes n'est pas couvert pas les ressources linguistiques disponibles.

Le résultat du processus décrit ci-dessus est un ensemble de similarités entre sujets, verbes et objets appartement a deux noyaux. Si les phrases contiennent plus d'un noyau, processus le sera exécuté autant de fois que nécessaire pour comparer deux à deux chacun des noyaux composant les phrases.

Considérons que nous avons deux phrases P_1 and P_2 à comparer. P_1 est composée de deux noyaux $K_1 = (s_1, v_1, o_1)$ et $K_2 = (s_2, v_2, o_2)$. Alors que P_2 est composée d'un seul noyau $K_3 = (s_3, v_3, o_3)$. Le résultat du processus décrit précédemment sera un ensemble de simila-

rité :

$\sigma(s_1, s_3), \sigma(s_2, s_3)$ - similarités entre sujets
$\sigma(v_1, v_3), \sigma(v_2, v_3)$ - similarités entre verbes
$\sigma(o_1, o_3), \sigma(o_2, o_3)$ - similarités entre objets

2.2 Construction d'indicateurs basés sur la similarité partielle entre noyaux

Comme nous l'avons précisé précédemment, le calcul de la similarité sémantique entre les éléments des noyaux des deux propositions, nous fournit des informations partielles quant à la similarité globale entre ces propositions. Dans ce qui suit, nous proposons une manière d'agréger ces similarités partielles afin d'en déduire une similarité globale entre propositions.

Si on considère les propositions P_a et P_b, nous calculons un ensemble de mesures de similarités entre les sujets, verbes et objets qui composent les noyaux de P_a et P_b. Pour chaque noyau i dans P_a et chaque noyau j dans P_b : $X_{ab} = \{\sigma(s_i, s_j)\}$, $Y_{ab} = \{\sigma(v_i, v_j)\}$ et $Z_{ab} = \{\sigma(o_i, o_j)\}$.

En utilisant X_{ab}, Y_{ab} et Z_{ab}, nous calculons des agrégats nous permettant de capturer des informations partielles relatives à la similarité globale entre P_a et P_b. Ces agrégats sont :

- $\overline{X_{ab}}$, $\overline{Y_{ab}}$ et $\overline{Z_{ab}}$: les moyennes des similarités dans les ensembles X_{ab}, Y_{ab} et Z_{ab}
- $max(X_{ab})$, $max(Y_{ab})$ et $max(Z_{ab})$: les similarités maximales dans X_{ab}, Y_{ab} et Z_{ab}.
- N_{ab} le nombre de comparaisons de noyaux effectuées entre P_a et P_b.

A partir de ces indicateurs, nous souhaitons déterminer, par un apprentissage automatique, quels sont ceux qui apportent une vraie information quant à la similarité globale entre deux phrases et quelle est la faslimanislimanion dont les éléments retenus peuvent se combiner dans une mesure de similarité globale. De ce fait, et c'est une des nouveautés de notre approche, nous proposons d'estimer les contributions des similarités partielles décrites ci-dessus à la similarité globale par un apprentissage automatique. Dans ce sens, nous avons choisi les modèles de regression linéaires pour sa simplicité d'interpretation et sa facilité de déploiment.

2.3 Estimation des coefficients des indicateurs

L'objectif ici est d'estimer une mesure de similarité sémantique entre phrases qui se rapproche le plus possible de la similarité sémantique telle qu'elle est perçues par les humains. Cette mesure est construite sur la base des indicateurs agrégés construits précédemment. Afin de déterminer, parmi ces indicateurs, ceux qui contribuent à l'obtention d'une mesure de similarité globale entre phrases, nous avons utilisé une méthode d'apprentissage basée sur la regression linéaires pas à pas. Ces techniques permettent d'utiliser, successivement, des sous-ensembles d'indicateurs afin de déterminer la meilleure combinaison d'indicateurs à retenir, par l'optimisation de l'erreur d'estimation des modèles testés.

L'apprentissage du meilleur modèle d'estimation de la similarité globale à partir des agrégats présentés dans la section précédente a été effectué sur un sous-ensemble du benchmark présenté dans O'Shea et al. (2008). Le choix de ce sous-ensemble de 30 couples de phrases est

uniquement lié au besoin d'évaluation de notre approche que nous développons dans la section suivante, car nous n'avons disposé des résultats des autres approches que sur ce sous-ensemble.

Le meilleur modèle d'estimation de la similarité globale entre deux phrases que nous avons obtenu ne retient que trois des sept indicateurs initialement fournis. Il se présente de la manière suivante :

$$\sigma(P_a, P_b) = -0.13213 + 0.60341 * \overline{X_{ab}} + 0.38057 * \overline{Z_{ab}} + 0.04893 * N_{ab}$$

Premièrement, nous observons que le modèle de calcul de la similarité globale entre deux phrases ne retient pas d'indicateurs relatifs à la similarité des verbes. Cela ne signifie pas que la similarité entre les verbes ne contribue pas à la similarité globale entre les phrases mais seulement qu'il n'y a pas de relation linéaire directe et significative entre les deux. Une explication pouvant justifier empiriquement cette absence de lien direct serait, par exemple, le fait qu'une similarité maximale entre des verbes de deux phrases différentes peut ne rien apporter à la similarité globale si les sujets et objets ne présente aucune similarité.

Statistiquement, le modèle calculant la similarité globale par l'équation présentée ci-dessus n'explique que partiellement la variance de la similarité humaine qu'il est censé estimer. Il faut noter que la similarité humaine qui est utilisée ici comme "golden standard" est elle même issue d'une moyenne avec un écart-type non négligeable (O'Shea et al. (2008)). Le R^2 ajusté, qui lui estime la qualité de la regression, s'établit à 0.66 alors que l'erreur résiduelle standard (RSE) est de 0.16 pour une échelle attendue entre [0,1]. Le test du modèle sur l'intégralité du benchmark décrit dans (O'Shea et al. (2008)) a révélé une correlation de 0.776 avec la similarité humaine moyenne et une déviation moyenne de 0.126.

Dans la section suivante nous comparons les résultats obtenus par le modèle que nous avons obtenu avec d'autres approches de calcul de similarité entre phrases décrites dans la littérature.

3 Evaluation de l'approche

Tel que précisé précédemment, nous avons utilisé le benchmark décrit dans (O'Shea et al. (2008)) pour déterminer notre modèle d'estimation de la similarité sémantique. Dans un premier temps, nous avons comparé nos résultats avec ceux des approches réputées dans l'état de l'art sur ce même benchmark. Puis, afin de prouver que notre modèle est robuste, nous avons effectué des expérimentations avec un autre benchmark non utilisé pour l'apprentissage (O'shea et al. (2014)).

3.1 Evaluation sur des benchmarks existants

Notre approche, baptisée ***SK*** *(Semantic Kernels)*, a été comparée avec quelques approches existantes dans l'état de l'art. L'évaluation a été faite sur un sous-ensemble de 30 paires de phrases sur les 65 présentes dans le benchmark car nous n'avons disposé des résultats des autres approches que sur ce sous-ensemble. La Figure 1 résume le résultat de cette comparaison en termes de corrélation et déviation moyenne.

Nous pouvons observer que notre approche est très bien située en comparaison avec les autres approches. Le coefficient de corrélation de notre modèle *SK* est à l'ordre de 0.83, juste derrière et très proche de *STS* (Islam et Inkpen (2008)) et *Omiotis* (Tsatsaronis et al. (2010)).

En termes de déviation moyenne, *SK* prend la deuxième place où il est très proche de *STS* et largement supérieur à *LSA, LSS* (Croft et al. (2013))(Landauer et al. (1998)) et *STASIS* (Li et al. (2006)).

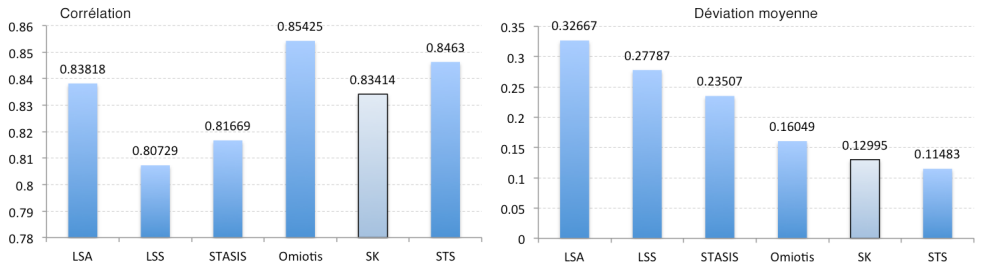

FIG. 1 – *Correlations et deviations moyennes des approches comparées avec la similarité humaine.*

Les résultats de la deuxième évaluation, avec un autre benchmark ont été comparées avec *LSA* dont les valeurs sont calculés via le portail web [1] mis à disposition par ses auteurs. La corrélation avec la similarité humaine est à l'ordre de 0.74 pour *LSA* et 0.90 pour notre approche *SK*. Notons que *LSA* était la seule approche pour laquelle nous avons pu fournir des valeurs de comparaison étant donné que les autres approches étaient indisponibles. Ces résultats montrent que notre approche, basée sur des noyaux sémantiques, atteint facilement les résultats des approches réputées dans la littérature.

4 Conclusion

Dans cet article, nous avons présenté une nouvelle approche dédiée à la mesure de similarité entre phrases. La particularité de cette approche est qu'elle estime la similarité sémantique entre phrases en se basant sur leurs noyaux sémantiques. Sur la base des benchmarks existants, nous avons utilisé des méthodes d'apprentissage automatique permettant d'obtenir à partir dun ensemble d'indicateurs, construits à partir de la similarité des noyaux, le meilleur modèle d'estimation de la similarité sémantique. L'évaluation expérimentale montre que les performances de notre approche sont comparables avec celles issues des méthodes décrites dans l'état de l'art.

Bien que les résultats obtenus soient performants, nous pensons que notre approche reste perfectible sur plusieurs niveaux. Etant donné que notre méthode utilise *Stanford Parser*, pour détecter les noyaux des phrases, il serait intéressant de prendre en compte d'autres informations fournies par ce dernier telles que le contexte de la phrase (interrogative, exclamative, etc). Nous pensons également que le modèle sémantique construit pourrait être amélioré par l'ajout de certaines dimensions sémantiques telles que l'aspect temporel. Enfin, il est évident que le benchmark utilisé dans l'apprentissage ne couvre pas tous les cas qu'on peut trouver dans la vie réelle. Pour cette raison, nous travaillons simultanément sur le développement d'un nouveau benchmark, plus élargi.

1. http ://lsa.colorado.edu/

Références

Budanitsky, A. et G. Hirst (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist. 32*(1), 13–47.

Croft, D., S. Coupland, J. Shell, et S. Brown (2013). A fast and efficient semantic short text similarity metric. In *Computational Intelligence (UKCI), 2013 13th UK Workshop on*, pp. 221–227.

Heidinger, V. (1984). *Analyzing Syntax and Semantics : Workbook*. Gallaudet University Press.

Hirst, G. et D. St Onge (1998). *Lexical Chains as representation of context for the detection and correction malapropisms*. The MIT Press.

Islam, A. et D. Inkpen (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data 2*(2), 10 :1–10 :25.

Landauer, T. K., P. W. Foltz, et D. Laham (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes* (25), 259–284.

Li, Y., D. McLean, Z. A. Bandar, J. D. O'Shea, et K. Crockett (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng. 18*(8), 1138–1150.

O'shea, J., Z. Bandar, et K. Crockett (2014). A new benchmark dataset with production methodology for short text semantic similarity algorithms. *ACM Trans. Speech Lang. Process. 10*(4), 19 :1–19 :63.

O'Shea, J., Z. Bandar, K. A. Crockett, et D. McLean (2008). A comparative study of two short text semantic similarity measures. In *Agent and Multi-Agent Systems : Technologies and Applications, Second KES International Symposium, KES-AMSTA 2008, Incheon, Korea, March 26-28, 2008. Proceedings*, pp. 172–181.

Tsatsaronis, G., I. Varlamis, et M. Vazirgiannis (2010). Text relatedness based on a word thesaurus. *J. Artif. Int. Res. 37*(1), 1–40.

Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Summary

We propose a new approach for semantic similarity between sentences by using semantic kernels that compose the sentences. Kernels, composed of triples (subject, verb, and object), are supposed to summarize the general meaning of each sentence they belong to. Based on the semantic similarities between kernel elements, we build descriptive features summarizing information about semantic similarity between phrases the kernels originate from. Then, using a supervised machine learning technique we estimate the coefficients of the descriptive features. The learning process is done on a benchmark containing phrases whose semantic similarities were evaluated human experts. Comparative studies with other semantic similarity measures in the litterature show good performances of our approach. Based on the latter, an application is being developed for highlighting semantic parts related to the elements described in abstracts of scientific articles.