Simbow : une mesure de similarité sémantique entre textes

RÉSUMÉ

Cet article décrit une mesure de similarité sémantique non-supervisée qui repose sur l'introduction d'une matrice de relations entre mots, dans un paradigme de mesure cosinus entre sacs de mots. La métrique obtenue, apparentée à *soft*-cosinus, tient compte des relations entre mots qui peuvent être d'ordre lexical ou sémantique selon la matrice considérée. La mise en œuvre de cette métrique sur la tâche qui consiste à mesurer des similarités sémantiques entre questions posées sur un forum, a remporté la campagne d'évaluation SemEval2017. Si l'approche soumise à la campagne est une combinaison supervisée de différentes mesures non-supervisées, nous présentons dans cet article en détail les métriques non-supervisées, qui présentent l'avantage de produire de bons résultats sans nécessiter de ressources spécifiques autres que des données non annotées du domaine considéré.

ABSTRACT

Simbow : a semantic similarity metric between texts

This article describes an unsupervised similarity metric which relies on the introduction of a word relation matrix, in a cosine bag of words paradigm. The obtained *soft*-cosine metric takes into acount word relations that can be lexical or semantic depending on the matrix under consideration. The implementation of this metric in a task consisting of computing semantic similarities between questions from a forum, won the SemEval2017 evaluation campain. If the approach submitted to the campain is a supervised combination of several unsupervised metrics, we present in detail in this article the unsupervised metrics, which already yield good results without needing any specific ressources other than unannotated in domain data.

MOTS-CLÉS : similarité sémantique entre textes, relations sémantiques, questions de forums.

KEYWORDS: semantic textual similarity, semantic relations, forum questions.

1 Introduction

Les forums permettent à leurs participants de poser des questions et d'interagir avec les autres pour obtenir des réponses pertinentes. La popularité des forums montre la capacité de ce type d'interactions à produire des réponses pertinentes aux questions. Leur popularité est telle que le premier réflexe d'un internaute, lorsqu'il se pose une question, est de faire appel à son moteur de recherche préféré, pour vérifier si une question similaire n'a pas déjà été posée et résolue. Pour répondre de façon pertinente à cette requête, il faudrait pouvoir mesurer une similarité entre la nouvelle question posée et les questions déjà postées sur le forum, qui tienne compte de la sémantique et pas seulement du nombre de mots en communs entre les questions, sur quoi s'appuient les moteurs de recherche usuels.

Dans le cadre de la campagne d'évaluation des technologies sémantiques SemEval 2017, la tâche *Community Question Answering*, est consacrée à l'analyse des questions et leurs fils de discussions dans les forums (Nakov *et al.*, 2017). Une sous-tâche traite précisément de la similarité entre questions : il s'agit pour une question, dite originale, de trier par similarité sémantique décroissante les 10

questions, dites relatives, remontées par un moteur de recherche sur ce forum. Cette tâche peut être vue comme une pure tâche de similarité sémantique entre textes, sur des données textuelles bruitées générées par des utilisateurs, à la différence d'une autre tâche de similarité textuelle (Agirre *et al.*, 2016), qui traite de textes courts bien formés. Dans la tâche *Community Question Answering*, le corpus d'étude est issu du forum en anglais Qatar Living, dans lequel les expatriés occidentaux discutent de tous les aspects de la vie quotidienne au Qatar (où pratiquer le basket, comment faire pour embaucher une nounou,...). Cette tâche est apparue dans SemEval en 2016 (Nakov *et al.*, 2016), et s'est poursuivie en 2017. Les approches proposées en 2016 étaient principalement des fusions tardives de mesures de similarités supervisées ou non. De nombreuses mesures non-supervisées étaient basées sur le comptage d'éléments en commun entre les questions, ces éléments pouvant être des n-grams de lettres ou de mots, ou des composants de plus haut niveau, comme des entités nommées, des roles sémantiques ou des portions d'arbres syntaxiques (e.g. (Franco-Salvador *et al.*, 2016)). Les plongements de mots ont également été beaucoup utilisés (e.g. (Mihaylov & Nakov, 2016)), souvent simplement moyennés au niveau de la question, et utilisé dans une mesure en cosinus ou en entrée d'un classifieur neuronal.

Les hypothèses de travail que nous avons prises pour cette tâche sont les suivantes : nous avons considéré que les données de forum étaient trop bruitées pour obtenir des sorties fiables de nos outils d'analyse linguistiques, et nous avons souhaité nous focaliser sur la similarité sémantique entre textes. C'est pourquoi nous n'avons utilisé aucune analyse de méta-données (issues des dates, profils utilisateurs etc...), afin d'obtenir des résultats qui pourraient se généraliser à d'autres tâches de similarité sémantique de textes. Ainsi, nous avons développé des mesures de similarité sémantiques entre textes, sans ressources externes, presque sans pré-traitements linguistiques, reposant uniquement sur la disponibilité d'un gros corpus non-annoté représentatif des données.

Contrairement à l'article à paraître lors de la conférence SemEval (Charlet & Damnati, 2017), où nous décrivons en détail l'optimisation de la combinaison supervisée de nos métriques non-supervisées, nous nous concentrons ici sur l'analyse des métriques non supervisées. La suite de cet article est organisée de la façon suivante : dans la section 2, la mesure de similarité non-supervisée est présentée, tandis que des mesures alternatives sont présentées dans la section 3. Les résultats sont présentés et discutés dans la section 4.

2 Mesure de similarité *soft*-cosine

2.1 Principe

Dans une approche classique en sacs de mots, les textes sont représentés par un vecteur de poids (e.g. TF-IDF) de taille N où N est la taille du lexique considéré. La mesure de similarité en cosinus entre ces 2 vecteurs est directement liée au nombre de mots en commun entre les 2 textes :

$$cos(X, Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{ où } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad (1)$$

Si il n'y a aucun mot en commun entre X et Y (i.e. aucun indice i pour lequel à la fois x_i et y_i sont non nuls), la mesure vaut 0. Or les textes peuvent être sémantiquement similaires, même si ils n'ont aucun mot en commun, mais si les mots qu'ils contiennent sont eux-mêmes sémantiquement liés.

C'est pourquoi nous proposons d'introduire les relations entre mots dans la mesure de similarité, en introduisant dans la formule du cosinus une matrice de relations :

$$cos(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \text{ où } X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad (2)$$

où M est une matrice dont l'élément $m_{i,j}$ traduit une relation entre le mot i et le mot j. Avec une telle métrique, la similarité entre 2 textes est non nulle dès que les textes ont des mots liés entre eux, même si ils n'ont aucun mot en commun. Introduire la matrice M également au dénominateur est nécessaire pour conserver le rôle de normalisation du dénominateur et une similarité réflexive égale à 1. On notera que si les mots n'ont de relations qu'avec eux-mêmes ($m_{i,i} = 1$ et $m_{i,j} = 0\ \forall i, j, i \neq j$), M devient la matrice identité, et on retrouve la mesure en cosinus classique.

Nous avons d'abord exploré cette mesure de similarité dans le contexte de la segmentation thématique de journaux télévisés (Bouchekif *et al.*, 2016) en utilisant les relations sémantiques entre mots pour améliorer le calcul de la cohésion sémantique entre passages consécutifs. Cette mesure avait été également proposée par d'autres chercheurs (e.g. (Sidorov *et al.*, 2014)), et dénommée *soft*-cosinus, où la matrice de relations était fondée par exemple sur une distance d'édition entre n-grams. Dans notre travail, nous explorons différents types de relations entre mots pour le calcul de M.

2.2 Relations sémantiques

Les plongements de mots, comme par exemple l'approche `word2vec` (Mikolov *et al.*, 2013) ont connu un immense succès ces dernières années. Ils permettent, entre autres, d'obtenir des relations sémantiques entre mots, simplement à partir d'une mesure de similarité (par exemple en cosinus) entre les vecteurs représentatifs des mots. Dans ce travail, 2 plongements sont calculés, avec l'outil `word2vec`, dans la configuration *cbow* : l'un à partir d'une version du Wikipedia en anglais, l'autre à partir du corpus non-annoté de 100 millions de mots disponible pour la campagne d'évaluation (questions et fils de discussions du forum Qatar Living). La dimension des vecteurs est fixée à 300 (les variantes de dimensions testées n'ont pas eu d'influence significative) et la fréquence minimale des mots à 50. A partir de ces plongements, M peut être calculée de différentes façons. Après avoir exploré différentes variantes, les meilleurs résultats ont été obtenus en considérant, pour traduire la relation entre deux mots w_i (représenté par v_i) et w_j (représenté par v_j) :

$$m_{i,j} = max(0, cosine(v_i, v_j))^2 \quad (3)$$

Le seuillage à 0 est motivé par l'observation que les cosinus négatifs entre mots sont difficiles à interpréter et ne semblent pas pertinents. Le passage au carré accentue la dynamique des relations sémantiques, en insistant sur les plus fortes et en diminuant les plus faibles.

2.3 Relations à partir de distances d'édition

Une similarité d'édition entre mots peut être calculée, à partir d'une distance d'édition comme celle de Levenshtein : elle permet de pallier certaines petites erreurs de typographie, fréquentes dans les textes produits par les réseaux sociaux. Nous définissons la matrice de similarité d'édition comme étant $m_{i,i} = 1$ et pour $i \neq j$:

$$m_{i,j} = \alpha * \left(1 - \frac{Levenshtein(w_i, w_j)}{max(||w_i||, ||w_j||)}\right)^{\beta} \quad (4)$$

où $||w||$ est le nombre de lettres du mot w, α est un facteur de pondération par rapport aux éléments diagonaux, et β un facteur pour amplifier la dynamique des scores. Les expériences sur le corpus $train$ et dev nous ont mené à fixer $\alpha = 1.8$ et $\beta = 5$.

3 Autres mesures

3.1 Mesure par appariement de mots

La spécificité de l'approche en soft-cosinus est de tenir compte des relations de tous les mots d'un texte par rapport à tous les mots d'un autre texte, ce qui contraste avec des approches d'alignement d'un texte par rapport à un autre, où chaque mot d'un texte est associé à un seul mot, au maximum, d'un autre texte. Ainsi, une mesure contrastive a été développée, dans une approche d'appariement entre les mots de chaque texte. Dans cette approche, chaque mot x_i de poids non nul du vecteur X est apparié à son plus proche voisin d'indice $vois(i)$ de poids non nul du vecteur Y :

$$vois(i) = argmax_{j,y_j \neq 0} m_{i,j}$$

En restant dans un paradigme inspiré du cosinus, une fois cet appariement réalisé, on propose la mesure de similarité de Y par rapport à X :

$$simvois(X,Y) = \frac{\sum_i x_i m_{i,vois(i)} y_{vois(i)}}{\sqrt{\sum_i x_i^2}\sqrt{\sum_j y_j^2}}$$

Le dénominateur assure que $simvois(X,X) = 1$ car le plus proche voisin d'un mot est toujours lui-même, et $m_{i,i} = 1$. Cette mesure est non-symétrique car la relation de plus proche voisin ne l'est pas. On la symétrise alors en calculant :

$$simvois_{sym}(X,Y) = 0.5 * (simvois(X,Y) + simvois(Y,X))$$

3.2 Mesure entre représentations moyennes des textes

A partir des plongements de mots, une façon très simple de représenter un texte consiste à calculer la moyenne, pondérée ou non, des plongements de mots présents dans le texte. Ensuite, la mesure de similarité entre textes est calculée par exemple par une mesure de similarité en cosinus entre les vecteurs représentatifs de chaque texte. Cette mesure sans pondération des mots a été très fréquemment utilisée par les participants des campagnes d'évaluation de similarité sémantique entre textes de SemEval. L'introduction d'une pondération permet des améliorations significatives de la représentation (e.g. (Arora *et al.*, 2017)), et l'approche de similarité cosinus entre vecteurs moyens pondérés constitue une base de départ robuste pour les représentations de texte. Nous choisissons ici de pondérer les mots par leur poids TF-IDF de leur représentation en sacs de mots, où x_i est le poids du terme i dans le texte X. Si $v_{i,k}$ est la k-ième composante du mot i dans l'espace des plongements de mots, la k-ième composante du vecteur représentatif $\tilde{X}$ du texte X est obtenue par :

$$\tilde{X}_k = \frac{1}{\sum_i x_i} \sum_i x_i v_{i,k}$$

La mesure $wavg - w2v(X,Y)$ consiste alors en une similarité en cosinus entre $\tilde{X}$ et $\tilde{Y}$.

4 Evaluation

4.1 Protocole expérimental

Nous présentons ici des évaluations de la tâche de similarité question-question (Task3-subtaskB de SemEval 2017). Etant donnée une nouvelle question, dite originale, il s'agit de trier par similarité sémantique décroissante les 10 questions, dites relatives, remontées par un moteur de recherche sur le forum Qatar Living. Une description détaillée du corpus et de la métrique d'évaluation se trouve dans (Nakov *et al.*, 2017). L'évaluation est faite avec la mesure MAP (Mean-Average-Precision), mesure standard en recherche d'information. Elle vaut 100% si, pour chaque question originale, toutes les questions relatives jugées pertinentes dans la vérité-terrain ont un score de similarité sémantique supérieur aux questions jugées non-pertinentes. On présente les résultats sur 3 corpus : dev (50 questions originales $\times$ 10 questions relatives), $test2016$ (70 questions originales $\times$ 10 questions relatives) et $test2017$ (88 questions originales $\times$ 10 questions relatives).

Il faut noter que le programme d'évaluation MAP utilisé dans la campagne est sensible au nombre de questions originales qui n'ont aucune question relative pertinente : le programme d'évaluation leur attribue une MAP de 0. Ainsi, une évaluation Oracle (où toutes les questions relatives labellisées pertinentes dans la vérité-terrain ont un score de similarité de 1, et toutes les questions relatives labellisées non-pertinentes dans la vérité-terrain ont un score de similarité de 0) ne donne pas un MAP de 100%, mais un MAP qui correspond au taux de question originale qui ont au moins une question relative labellisée pertinente. Cette borne supérieure du MAP vaut 86.0% pour le dev, 88.6% pour le $test2016$ et seulement 67.0% pour le $test2017$. De plus, pour le $test2017$, le nombre moyen de questions relatives labellisées pertinentes par question originale (pour les questions originales ayant au moins une relative pertinente) n'est que de 2.7, tandis qu'il est de 3.7 pour $test2016$. C'est pourquoi le $test2017$ est un corpus plus difficile que le $test2016$.

4.2 Pré-traitements des données

Des pré-traitements très simples sont appliqués sur les textes : suppression de la casse, de la ponctuation, des mots-outils, remplacement des urls et des images par des termes génériques "_url_" et "_img_". En ce qui concerne les poids TF-IDF de la représentation sacs de mots, nous proposons une approche particulière, où les IDF sont estimés sur le gros corpus non-annoté du domaine du forum *Qatar Living*. Les "documents" considérés sont à la fois les titres et les paragraphes du forum (commentaires inclus).

4.3 Évaluation des mesures de similarité non-supervisées

Le tableau 1 présente les performances MAP des différentes mesures de similarités non-supervisées. Nous ne présentons que les résultats des similarités mesurées entre les textes $sujet + corps$ de la question originale et des questions relatives, configuration optimale pour la tâche. A titre de comparaison, la performance du moteur de recherche qui produit les 10 questions relatives, ainsi que la performance du système qui a gagné la compétition SemEval2016 (Franco-Salvador *et al.*, 2016) sont également présentées.

Nous prenons comme point de départ des évaluations le système dit *baseline_token_cos*, tel que défini

similarités	dev	$test2016$	$test2017$
moteur de recherche	71.35	74.75	41.85
gagnant SemEval2016	-	77.33	-
mesures non supervisées			
baseline_token_cos	62.22	68.54	40.88
baseline_pp_cos	67.49	71.05	42.80
baseline_pp_cos_tfidf	69.41	75.53	44.37
cos_{Mrel} relations WP	72.25	77.11	45.38
cos_{Mrel} relations QL	**75.24**	**77.96**	45.27
cos_{Mlev} Levenshtein	70.02	76.34	46.10
$simvois_{sym}$ relations QL	74.40	77.47	45.67
wavg-word2vec sur QL	73.31	75.77	**46.99**
combinaison supervisée des mesures			
soumission SemEval2017	**77.30**	**79.77**	**47.22**

TABLE 1 – Performance MAP pour les mesures de similarité non-supervisées

dans les évaluations de similarité sémantique de phrases de SemEval2015-Task2 (Agirre *et al.*, 2015) : c'est une simple similarité en cosinus entre sacs de $tokens$, sans pondération (un $token$ étant défini comme une séquence de caractères non espaces entre 2 caractères d'espaces, avec des poids valant 1 ou 0). Les performances obtenues avec *baseline_pp_cos*, qui est aussi un cosinus entre vecteurs binaires, mais sur les termes obtenus après pré-traitements, rappelle l'importance de pré-traitements appropriés. Ensuite, *baseline_pp_cos_tfidf* confirme l'influence d'une pondération pertinente des termes.

Les résultats suivants montrent une amélioration significative des performances lorsqu'on introduit une matrice M dans le calcul de la similarité en cosinus. Lorsqu'il s'agit de relations sémantiques, une différence significative est observée sur dev entre les relations estimées sur un corpus générique (WP : Wikipedia) et sur un corpus spécialisé du domaine (QL : Qatar Living). Cette différence est plus faible pour $test2016$ et même négative pour $test2017$. A l'inverse, l'utilisation de la matrice M basée sur une distance d'édition Levenshtein améliore le plus sur $test2017$ tandis que ces gains sont marginaux pour dev et $test2016$. L'approche $simvois_{sym}$ qui ne considère que le plus proche voisin d'un mot dans l'autre phrase donne également des performances intéressantes, et très proches de l'approche soft-cosine qui tient compte des relations entre tous les mots. Finalement, la mesure $wavg - w2v$ est moins performante que le *soft*-cosinus avec relations sémantiques sur dev et $test2016$ mais donne de bons résultats sur $test2017$. Les particularités de ce corpus, où les appariements pertinents sont largement minoritaires parmi les propositions du moteur de recherche, font que les résultats sont beaucoup moins contrastés. On notera que la mesure cos_{Mrel} avec relations sémantiques sur le corpus spécialisé aurait gagné la compétition de 2016. La combinaison des trois mesures, via une régression logistique optimisée sur le $train$ selon un processus décrit en (Charlet & Damnati, 2017), produit les meilleurs résultats quel que soit le corpus considéré, comme le montre la dernière ligne du tableau.

4.4 Expériences contrastives

Nous présentons dans cette section quelques expériences contrastives sur le corpus dev afin de mesurer plus précisément l'influence des différents éléments de la chaine, depuis les pré-traitements jusqu'à l'introduction d'une matrice de relation sémantique. Le tableau 2 présente l'influence des

pré-traitements sur les performances. Dans l'approche simple qui consiste à évaluer un cosinus entre les sacs de mots, nous pouvons voir que la lemmatisation dégrade les performances lorsque les sacs de mots ne sont pas pondérés. En revanche la mise en œuvre d'une pondération appropriée (estimée sur le corpus QL lemmatisé) apporte un gain substantiel, qui est accru par le passage à la lemmatisation. Lorsque l'on introduit une matrice de relations, apprise sur le corpus QL également lemmatisé, on observe une amélioration importante des résultats par rapport au cosinus, mais non significative par rapport à l'approche non lemmatisée.

similarité	**lemme**	**TF-IDF**	**MAP on** dev
cosinus	-	-	67.49
cosinus	✓	-	65.46
cosinus	-	✓	69.41
cosinus	✓	✓	71.67
soft-cosinus	-	✓	75.24
soft-cosinus	✓	✓	75.85

TABLE 2 – Influence des pré-traitements sur les performances

5 Conclusion

Dans ce travail, nous avons exploré une extension de la mesure de similarité en cosinus entre sacs de mots, par l'introduction d'une matrice M de relations entre mots dans un *soft*-cosinus. Le calcul de M est non-supervisé et peut être issu de plongements de mots. Ce *soft*-cosinus donne de bonnes performances sur la tâche de similarité entre questions de la campagne SemEval2017-Task3. Une simple combinaison par régression logistique des différentes mesures de similarité non-supervisées s'est classée première de la compétition officielle. Nous souhaitons approfondir l'étude des similarités basées sur le *soft*-cosinus dans 2 directions : d'une part inclure de nouvelles relations dans M, par exemple à partir de rôles sémantiques, et d'autre part, étudier comment cette matrice M, efficacement initialisée de façon non-supervisée grâce aux plongements de mots, pourrait être ensuite apprise pour des tâches spécifiques.

Références

AGIRRE E., BANEA C., CARDIE C., CER D., DIAB M., GONZALEZ-AGIRRE A., GUO W., LOPEZ-GAZPIO I., MARITXALAR M., MIHALCEA R., RIGAU G., URIA L. & WIEBE J. (2015). Semeval-2015 task 2 : Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 252–263, Denver, Colorado : Association for Computational Linguistics.

AGIRRE E., BANEA C., CER D. M., DIAB M. T., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 497–511.

ARORA S., LIANG Y. & MA T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*, Toulon, France.

BOUCHEKIF A., DAMNATI G., CHARLET D., CAMELIN N. & ESTÈVE Y. (2016). Title assignment for automatic topic segments in TV broadcast news. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, p. 6100–6104.

CHARLET D. & DAMNATI G. (2017). Simbow at semeval-2017 task 3 : Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada : Association for Computational Linguistics.

FRANCO-SALVADOR M., KAR S., SOLORIO T. & ROSSO P. (2016). Uh-prhlt at semeval-2016 task 3 : Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 814–821, San Diego, California : Association for Computational Linguistics.

MIHAYLOV T. & NAKOV P. (2016). Semanticz at semeval-2016 task 3 : Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. *Proceedings of SemEval*, p. 879–886.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

NAKOV P., HOOGEVEEN D., MÀRQUEZ L., MOSCHITTI A., MUBARAK H., BALDWIN T. & VERSPOOR K. (2017). SemEval-2017 task 3 : Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada : Association for Computational Linguistics.

NAKOV P., MÀRQUEZ L., MOSCHITTI A., MAGDY W., MUBARAK H., FREIHAT A. A., GLASS J. & RANDEREE B. (2016). Semeval-2016 task 3 : Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 525–545.

SIDOROV G., GELBUKH A. F., GÓMEZ-ADORNO H. & PINTO D. (2014). Soft similarity and soft cosine measure : Similarity of features in vector space model. *Computación y Sistemas*, **18**(3).