Word2Vec vs LSA pour la détection des erreurs orthographiques produisant un dérèglement sémantique en arabe

Chiraz Ben Othmane Zribi[1]
(1) Laboratoire RIADI-GDL, ENSI, Université La Manouba, 2010, La Manouba
Chiraz.zribi@ensi-uma.tn

RESUME

Les mots en arabe sont très proches lexicalement les uns des autres. La probabilité de tomber sur un mot correct en commettant une erreur typographique est plus importante que pour le français ou pour l'anglais. Nous nous intéressons dans cet article à détecter les erreurs orthographiques plus précisément, celles générant des mots lexicalement corrects mais causant un dérèglement sémantique au niveau de la phrase. Nous décrivons et comparons deux méthodes se basant sur la représentation vectorielle du sens des mots. La première méthode utilise l'analyse sémantique latente (LSA). La seconde s'appuie sur le modèle Word2Vec et plus particulièrement l'architecture Skip-Gram. Les expérimentations ont montré que Skip-Gram surpasse LSA.

ABSTRACT

Word2Vec vs LSA for detecting semantic errors in Arabic language.
Arabic words are lexically close to each other. The probability of having a correct word by making a typographical error is greater than for French or English. We are interested in this article to detect spelling errors more precisely, those generating lexically correct words but causing a semantic disturbance in the sentence. We describe and compare two word embedding based methods. The first one uses Latent Semantic Analysis (LSA). The second, is based on the Word2Vec model and more precisely the Skip-Gram architecture. Experiments have showed that LSA is more efficient than Skip-Gram for both precision and recall.

MOTS-CLES : Erreurs orthographiques, dérèglement sémantique, représentation vectorielle, LSA, Word2Vec, Skip-Gram, langue arabe.

KEYWORDS : Spelling Errors, semantic disturbance, word embedding, LSA, Word2Vec, Skip-Gram, Arabic language.

1 Introduction

Les erreurs orthographiques qui produisent des mots lexicalement corrects causant un dérèglement sémantique au sein du contexte où elles se trouvent peuvent être dues à des problèmes de performance (i.e. faute de frappe) ou à des problèmes d'ignorance (i.e. confusion avec un autre mot).

Quand l'erreur est due à une faute de frappe par exemple, le mot erroné est généralement proche lexicalement du mot correct, comme dans les exemples amusants suivants :

Exemple en français : erreur de substitution d'une lettre par une autre :

La maman prépare un bon ***râteau*** (gâteau)

Exemple en arabe : interversion de deux lettres adjacentes

ترك له والده **ثورة** (ثروة)
/trk lh wAldh **vwrp** (vrwp)/
Son père lui a laissé une **révolution** (fortune)

Ces erreurs dites « sensibles au contexte », comptent environ 40% parmi toutes les erreurs orthographiques étudiées, selon (Verberne, 2002). Cette valeur assez importante a rendu l'étude de ce genre d'erreurs une nécessité en soi. En effet, plusieurs recherches ont été entreprises dans le but de remédier à ce problème notamment pour les langues indo-européennes telles que le français et l'anglais. Toutefois, en arabe, rares sont les travaux qui se sont attelés à les traiter en dépit de l'importance de cette tâche. En effet, les mots arabes sont lexicalement proches les uns des autres. Le risque de tomber sur un mot correct en commettant une erreur typographique (ajout d'une lettre, suppression d'une lettre, substitution d'une lettre par une autre et interversion de deux lettres adjacentes) est relativement important comme l'ont montré (Ben Othmane Zribi et al., 2005). Selon ces auteurs le nombre moyen de formes voisines qui diffèrent d'une seule opération d'édition est de 26,5 pouvant atteindre un maximum de 185, valeur importante comparée à celle calculée pour la langue française égale à 3,5 et celle relative à l'anglais égale à 3. Aussi, les auteurs nous renseignent sur la probabilité d'obtenir un mot correct lorsqu'une erreur est commise sur un mot. Cette probabilité pour un mot arabe (5,79%) est 10 fois plus grande que pour un mot anglais (0,59%) et 14 fois plus grande que pour un mot français (0,39%).

Nous nous focalisons dans cet article sur la détection des erreurs orthographiques en arabe qui produisent des mots lexicalement et syntaxiquement corrects mais qui causent des incohérences sémantiques. Nous utilisons et comparons à cet effet deux méthodes vectorielles qui permettent d'inférer le sens des mots à partir de leur distribution les uns par rapport aux autres, à savoir LSA (Landauer, 1998) et Skip-Gram (Mikolov et al., 2013). LSA est une méthode très utilisée pour représenter le sens, elle a été déjà utilisée pour détecter ce type d'erreurs avec quelques variantes au niveau de l'implémentation. La méthode Word2Vec, apparue ces dernières années, permet quant à elle de créer des vecteurs de sens utilisant des réseaux de neurones. Ceci a pour principal avantage de faciliter l'utilisation de grandes quantités de données d'apprentissage. Beaucoup de travaux ont été conduits utilisant Word2Vec mais, à notre connaissance, cette méthode n'a pas été auparavant testée pour détecter le type d'erreurs que nous visons. Afin de restreindre les champs de nos investigations, nous émettons l'hypothèse de l'existence d'une seule erreur sémantique par phrase et par mot. Cette erreur consisterait en une seule faute typographique générant un mot lexicalement et syntaxiquement correct, relevant de l'une des opérations d'édition citées précédemment. Des statistiques ont en effet montré que l'une (seulement) de ces opérations est à l'origine d'une erreur orthographique dans 90% des cas (Ben Hamadou, 1993). Nous avons également considéré l'Arabe moderne standard non voyellé car les écrits arabes sont généralement dépourvus de voyelles, c'est le cas des textes fréquemment rencontrés dans les journaux, les revues, les romans, etc. L'arabe voyellé concerne seulement quelques ouvrages poétiques ou littéraires didactiques ou encore le coran.

Le plan de l'article est comme suit : La section 2 est consacrée à la présentation de l'état de l'art. Nous décrivons dans la section 3 les deux méthodes que nous proposons. Nous présentons par la suite les expérimentations et les résultats obtenus dans la section 4. Enfin, nous concluons et donnons quelques perspectives dans la dernière section.

2 Etat de l'art

Dans la littérature, le problème des erreurs sémantiques a été considéré selon deux visions. Certains chercheurs ont considéré ce problème comme une tâche de résolution d'ambiguïté lexicale. Ils utilisent des ensembles de mots préétablis nommés "ensembles de confusion", contenant des mots semblables par le son (i.e. {stationary, stationery}), par l'écriture (i.e. {dessert, desert}) et par l'usage (i.e. {between, among}). Selon cette approche, un mot est simplement soupçonné lorsqu'un membre de son ensemble de confusion est mieux adapté à son contexte. Ce mot est corrigé en sélectionnant l'alternative la plus probable par rapport au contexte. (Golding, 1995) est à l'origine des méthodes basées sur le jeu de confusion. Il a proposé avec ses collègues plusieurs méthodes d'apprentissage automatique (pour le même ensemble de confusion), présentées ici dans l'ordre chronologique : la méthode hybride bayésienne basée sur les probabilités ainsi que les collocations (Golding, 1995), la méthode Tribayes combinant une méthode trigramme avec une méthode hybride bayésienne (Golding & Schabes, 1996) et l'algorithme de Winnow utilisant les mots voisins et adjacents avec un vote à majorité pondérée (Golding , Roth, 1999). Ces méthodes ont donné respectivement un taux de précision de 83%, 89% et 93,5% en moyenne. Le meilleur résultat a été obtenu plus tard par (Carlson et al., 2001). Ils ont proposé une méthode basée sur l'architecture d'apprentissage SNOW (un classificateur multi-classes) et ont testé jusqu'à 265 ensembles de confusion avec une précision de 99%. D'autres chercheurs les ont rejoints, comme par exemple le cas de (Mangu & Brill, 1997). Ils ont proposé de nouvelles méthodes et testé leur système sur les mêmes ensembles de confusion. Plus récemment, certains travaux ont développé des systèmes de correction basés sur les modèles n-gram à l'échelle du Web. Dans ces systèmes, le choix du mot dépend de la fréquence à laquelle chaque candidat (un membre de l'ensemble de confusion) a été vu dans le contexte donné dans des données d'apprentissage du Web, comme le Google N-gram Corpus. Nous pouvons citer par exemple (Bergsma et al., 2010) qui ont amélioré la précision (95,7% en moyenne) pour 5 ensembles de confusion dont la performance moyenne rapportée dans (Golding, Roth, 1999) est inférieure à 90%.

D'autres chercheurs ne se sont pas restreints à des ensembles de confusion prédéfinis. Ils ont utilisé le contexte pour détecter les erreurs sémantiques en appliquant des méthodes basées sur des informations sémantiques ou probabilistes. Les résultats obtenus sont souvent moins bons, car la tâche est plus difficile. Nous pouvons citer (Verberne, 2002) qui a appliqué une méthode trigramme et l'a testée sur 5 500 mots du British National Corpus (un sous-ensemble des données d'apprentissage) avec 606 erreurs. Les taux de rappel et de précision de détection sont respectivement de l'ordre de 72% et de 98%. Lorsque cette méthode a été testée sur des données de test hors entraînement les résultats pour la détection ont été largement inférieurs avec un taux de rappel de 51% et un taux de précision de 5%. (Hirst, Budanitsky, 2005) ont utilisé des mesures de distance sémantique dans WordNet pour détecter et corriger les malapropismes. Une erreur est signalée lorsqu'une variante d'orthographe (tout mot dont la distance d'édition est 1 du mot d'origine) entraîne un nouveau mot sémantiquement plus proche au contexte. Cette méthode a atteint une précision d'environ 23% lorsqu'elle a été testée sur environ 300 000 mots du corpus Wall Street Journal de 1987-89, avec environ 1 400 malapropismes introduits aléatoirement à une fréquence d'environ un mot sur 200. Plus récemment, (Zesch, 2012) a combiné une méthode statistique utilisant le modèle n-gram avec une méthode à base de connaissances utilisant WordNet, inspirée de celle de (Hirst, Budanitsky, 2005) pour détecter les malapropismes en anglais et en allemand. La combinaison des deux méthodes s'est révélée avantageuse au niveau des taux de précision qui sont de l'ordre de 90%. Les taux de rappel sont par contre faibles, ils sont d'environ 50% en moyenne. Aussi, pour détecter les erreurs sémantiques, (Gutierrez et al., 2014) a proposé une méthode basée sur le raisonnement logique utilisant une ontologie du domaine. Les taux de F-mesure obtenus varient entre 58% et 90%.

Pour l'arabe, un seul travail, à notre connaissance, s'est intéressé à la détection des erreurs sémantiques. (Ben Othmane Zribi, Ben Ahmed, 2013) ont proposé un Système Multi-Agents (SMA) combinant quatre méthodes contextuelles dont LSA et n'utilisant pas d'ensemble de confusion. Un système de vote permet de décider de la présence d'une erreur au sein d'une phrase. Les taux de précision et de rappel rapportés pour environ 1400 erreurs sémantiques générées artificiellement sont respectivement de l'ordre de 90% et de 83%.

3 Une méthode à base de représentation vectorielle des mots pour détecter les erreurs sémantiques

La majorité des chercheurs, en s'intéressant au problème des erreurs orthographiques sensibles au contexte, ont utilisé des ensembles de confusion. Les résultats obtenus sont d'une manière générale très satisfaisants car la problématique est relativement simple. Nous avons choisi dans ce travail de ne pas utiliser d'ensembles de confusion et de détecter toute erreur générant une incohérence sémantique au sein de son contexte. Ce choix est doublement motivé. D'une part, nous avons voulu traiter le problème des erreurs sémantiques dans sa globalité et ne pas nous limiter à un ensemble restreint d'erreurs prédéfinies. D'autre part, nous pensons qu'utiliser des ensembles de confusion pour l'arabe ne serait pas très judicieux. En effet, due à la proximité lexicale des mots, les ensembles de confusion seraient nombreux et de taille importante.
Afin de détecter ces erreurs sémantiques, nous faisons appel à deux méthodes se basant sur la représentation vectorielle des mots, à savoir LSA et Skip-Gram. Le modèle vectoriel n'est pas récent, il a en effet été introduit par (Salton et al., 1975) en recherche documentaire. Sa réhabilitation dans les recherches en TALN est par contre relativement récente notamment avec l'apparition des plongements lexicaux (word embeddings). Cette technique correspond à la représentation des mots par des vecteurs de nombres réels qui capturent leurs sens, leurs liens sémantiques et les différents contextes de leur utilisation. La représentation vectorielle des mots est principalement utilisée pour comparer les mots entre eux. Elle a ceci de particulier que les mots apparaissant dans des contextes similaires, et donc liés sémantiquement, possèdent des vecteurs correspondants qui sont relativement peu distants dans l'espace vectoriel où ils sont définis.
Chacune des deux méthodes proposées fournit sa propre représentation vectorielle des mots en fonction de leurs contextes. Nous calculons pour chaque mot à vérifier un coefficient de « validité sémantique » obtenu en comparant le vecteur de ce mot à tous les autres vecteurs-mot de la phrase. Un mot est soupçonné d'erreur si son coefficient de validité sémantique est inférieur à un seuil (déterminé empiriquement et préalablement établi), autrement dit, s'il est jugé suffisamment distant de ses voisins.

Afin déterminer le coefficient de validité sémantique d'un mot $\Omega(m_i)$ nous calculons la moyenne des distances angulaires entre le vecteur mot Vm_i et tous les $n-1$ vecteurs-mot dans la phrase (n étant le nombre de mots de la phrase) tout en privilégiant les mots contextuels les plus proches par rapport aux mots contextuels les plus éloignés.
Soit et $Cg = \{m_1, \ldots, m_{i-1}\}$ et $Cd = \{m_{i+1}, \ldots, m_n\}$ respectivement le contexte gauche et droit du mot à analyser m_i :

$$\overline{D_i} = \frac{\sum_{j=1}^{i-1} \frac{1}{i-j} D(Vm_i, Vmg_j) + \sum_{j=i+1}^{n} \frac{1}{j-i} D(Vm_i, Vmd_j)}{n-1}$$

$$\text{avec} \quad D(Vm_i, Vm_j) = \arccos \frac{Vm_i \bullet Vm_j}{\|Vm_i\| \times \|Vm_j\|} .$$

Cette distance est interprétée comme suit : “Deux mots *x* and *y* sont sémantiquement proches si *D(Vx, Vy)* ≤ 45 °. Lorsque *D (Vx, Vy)* > 45 °, la proximité sémantique est faible et pour 90° *x* et *y* n’ont aucune relation” (Schutze, 1998).

3.1 Principe et application de la méthode LSA

L’analyse sémantique latente (*LSA*) est l’une des méthodes les plus utilisées pour représenter le sens des mots. Elle permet d’identifier la similarité sémantique entre deux mots, deux segments textuels ou la combinaison des deux même si ces mots ou segments textuels ne sont pas co-occurrents. LSA prend en entrée un corpus textuel d’entrainement, construit une matrice de co-occurrences dont les lignes correspondent aux unités lexicales et les colonnes aux unités textuelles. Une normalisation est d’abord appliquée afin de réduire les poids des mots qui sont fréquents mais non informationnels. Ensuite, une réduction des dimensions de l’espace vectoriel des mots est réalisée à l’aide d’une analyse factorielle appelée décomposition en valeurs singulières.

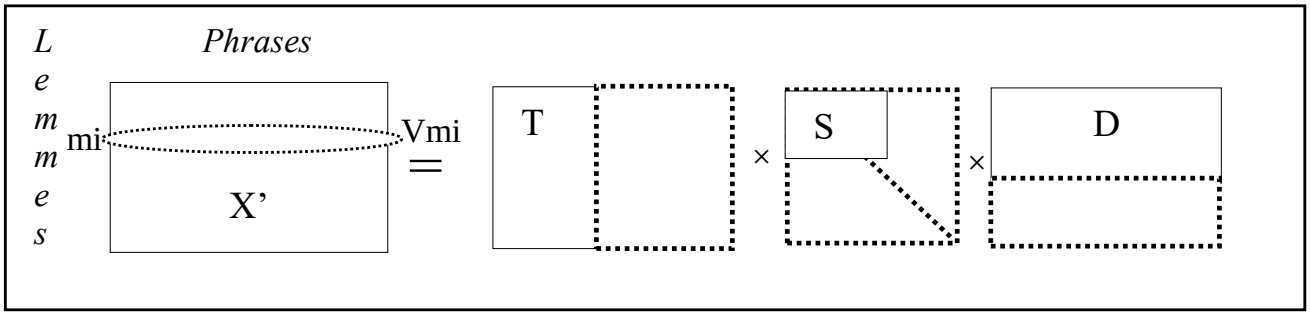

FIGURE 1 : Matrice de co-occurrence réduite

Dans ce travail, les vecteurs mots de la phrase à vérifier sont obtenus à partir de la matrice de cooccurrence réduite dont les lignes correspondent aux lemmes et les colonnes aux phrases. La taille de l’espace vectoriel est fixée à 300, valeur déterminée empiriquement. Le choix de la phrase comme contexte nous a paru raisonnable vu que celle-ci représente une unité sémantique dont le contenu se dégage du rapport établi entre les mots qu’elle contient.

3.2 Principe et application de l’architecture skip-gram du modèle Word2Vec

(Mikolov et al., 2013) considèrent que les méthodes déterministes basées sur le calcul des fréquences, telles que LSA, sont limitées pour représenter le sens des mots et ont introduit le modèle « Word2vec » à la communauté du TAL. Ce modèle est prédictif car il affecte des probabilités aux mots et a montré son efficacité par rapport à l’état de l’art pour des tâches de calcul de similarité et d’analogie entre les mots. Il est capable d’effectuer des tâches, comme le fameux exemple, vec(Roi)-vec(Homme)+vec(Femme) = vec(Reine), qui est un résultat assez intéressant. Il se base sur l’utilisation d’un réseau de neurones entrainé par des exemples de mots et de leurs contextes à partir d’un corpus d’apprentissage. Une fois entrainé, la transformation linéaire apprise au niveau de la couche cachée constitue la représentation vectorielle du mot cible. Le modèle Word2vec a été proposé en deux versions CBOW et Skip-Gram. CBOW permet de prédire un mot à partir d’un contexte tandis que Skip-gram prédit un contexte pour un mot. Pour notre problème, nous pensons que l’architecture Skip-Gram est plus adéquate, vu que nous vérifions pour chaque mot dans la phrase sa validité sémantique au sein de son contexte.

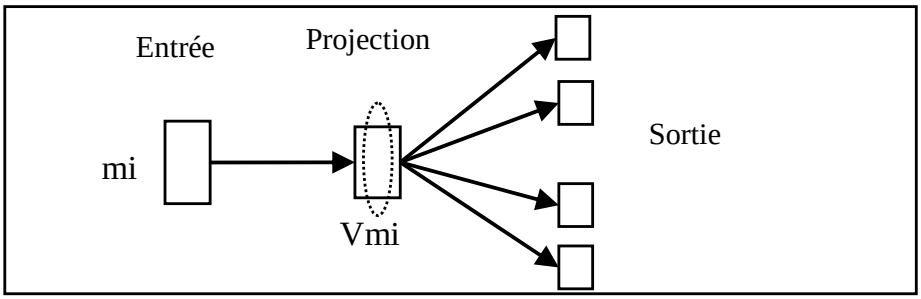

FIGURE 2 : Architecture Skip-Gram

Pareillement que pour LSA, nous avons fixé la taille des vecteurs mots à 300, ce qui correspond aux nombres de nœuds de la couche cachée. En outre, puisque LSA telle que nous l'avions définie tient compte du contexte de la phrase, nous avons choisi de faire en sorte que la taille de la fenêtre soit dynamique et toujours égale à la taille de la phrase en cours de traitement.

4 Expérimentations et comparaison des résultats

A cause de la non disponibilité de corpus contenant des erreurs naturelles correspondants à des mots appartenant au lexique, les travaux s'intéressant à la détection de ces erreurs ont été évalués dans leur majorité sur de erreurs générées de manière artificielle et introduites dans des corpus. Ne disposant pas d'un tel corpus pour l'arabe et dans le souci de comparer notre travail avec ce qui a été déjà proposé pour traiter ce type d'erreurs dans cette langue, nous avons utilisé les mêmes corpus d'apprentissage et de test que (Ben Othmane Zribi, Ben Ahmed, 2013). Rappelons que ce travail consiste en l'application d'un SMA combinant quatre méthodes contextuelles dont LSA. Signalons toutefois que la méthode LSA que nous utilisons considère la même taille de l'espace vectoriel (k = 300) mais diffère de cette dernière par le choix des unités lexicales car elle utilise le contexte de la « macro-phrase »[1] alors que nous considérons le contexte de la phrase simple. Le corpus d'apprentissage est composé d'un ensemble d'articles dans le domaine de l'économie[2] extraits du journal égyptien Al-Ahram (2009-2010) contenant environ 1 million de mots. Le corpus de test est extrait du même journal, mais hors apprentissage. Il contient environ 300 000 mots et 1 398 erreurs orthographiques produisant des dérèglements sémantiques générées artificiellement et introduites à la fréquence d'une erreur chaque 200 mots. Ces erreurs ont été générées semi-automatiquement et diffèrent des mots à remplacer d'une seule opération d'édition (insertion, suppression ou substitution d'un seul caractère, ou encore transposition de deux caractères adjacents). Nous avons utilisé le détecteur-correcteur d'erreurs orthographiques de (Ben Othmane & Zribi, 1999) qui fournit, en lui soumettant un mot correct, tous les mots lexicalement proches d'une seule opération d'édition. Les erreurs sont choisies manuellement parmi ces mots proches en faisant en sorte qu'elles ne soient pas des mots outils, qu'elles soient correctes syntaxiquement et qu'elles engendrent une incohérence sémantique au sein de la phrase. Voici un exemple d'erreur sémantique insérée dans notre corpus de test :

...ويتطلب المشروع قرضا من **الحنك** (البنك) الدولي ...

/…wytTlb Alm$rwE qrDA mn **AlHnk** (Albnk) Aldwly…/

… et le projet nécessitera un crédit **du palais** (bancaire) international …

[1] Macro-phrase est une phrase obtenue à l'issue d'un découpage à base de ponctuations et non de délimiteurs lexicaux. Elle peut correspondre à tout un paragraphe en français.

[2] Ce type de corpus est écrit en arabe standard moderne, plus facile à traiter que l'arabe classique qui est relativement plus ancien.

Comme l'illustre le tableau ci-dessous, notre méthode LSA avec un contexte restreint au niveau des unités lexicales, donne de meilleurs résultats au niveau de la précision que LSA avec un contexte plus élargi avec certes une petite perte au niveau du rappel. Elle reste cependant supérieure au niveau de la F-mesure. La méthode SMA est supérieure à notre méthode LSA aussi bien au niveau de la précision que de la F-mesure et ceci s'explique par la combinaison des quatre méthodes utilisées qui créent selon l'auteur une certaine synergie. En ce qui concerne Skip-Gram, nous remarquons d'abord sa supériorité à LSA aussi bien pour la précision que pour le rappel. Ensuite, comparativement à la méthode SMA, Skip-Gram est légèrement inférieure au niveau de la précision. Ceci pourrait s'expliquer par le fait que la méthode SMA utilise un processus de vote entre ses méthodes pour décider d'une erreur, ce qui la rend moins sensible au bruit. Cependant, Skip-Gram donne de meilleurs résultats globalement au niveau de la F-mesure avec une hausse de 5 points au niveau du rappel. Néanmoins, cette supériorité reste à vérifier en faisant par exemple varier la taille du corpus d'apprentissage. Nous pouvons citer (Altszyler, 2016) qui ont montré que LSA pouvait dépasser Skip-Gram quand le corpus est de petite taille.

Méthode	Précision (%)	Rappel (%)	F-mesure (%)
LSA (Ben Othmane Zribi et al., 2013)	79,62	84,44	81,96
SMA (Ben Othmane Zribi et al., 2013)	**90,55**	82,73	86,46
LSA	85,33	83,78	84,54
Skip-Gram	89,48	**86,15**	**87,78**

TABLE 1 : Evaluation et comparaison de la détection des erreurs

4.1 Un exemple d'échec

L'exemple suivant illustre un exemple d'erreur détectée à tort (faux positif) par les deux méthodes LSA et Skip-Gram :

لضمان لحاق الاقتصاد **بركب** الانتعاش وبناء القوة الاقتصادية الحقيقية

/… lDmAn lHAq AlAqtSAd **brkb** AlAntEA$ wbnA' Alqwp AlAqtSAdyp AlHqyqyp…/

… pour assurer que l'économie rattrape le **convoi** de la relance et la construction de la vraie force économique …

L'expression "لحاق... بركب"/lHAq …brkb /(rattraper le convoi) est une collocation non contiguë. Les mots de la phrase n'ont pas de lien sémantique avec le terme "ركب "/ rkb /(convoi), c'est pourquoi sa validité sémantique a été jugée faible.

5 Conclusion

Nous avons présenté dans cet article deux méthodes vectorielles afin de détecter les erreurs orthographiques causant un dérèglement sémantique au niveau de la phrase, à savoir LSA et Skip-Gram. La méthode Skip-Gram a donné des résultats encourageants par rapport à l'état de l'art. Elle a également montré sa supériorité par rapport à LSA aussi bien au niveau de la précision qu'au niveau du rappel. Nous comptons vérifier cette supériorité, dans le futur proche en faisant varier les tailles des corpus d'apprentissage et aussi le type des textes et voir dans quelle mesure le domaine des textes peut-il influer sur les résultats. Aussi, la correction automatique de ces erreurs en utilisant le contexte représente une perspective proche pour ce travail.

Références

ALTSZYLER E. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. CoRR Journal abs/1610.01520.

BEN HAMADOU A. (1993). Vérification et correction automatique par analyse affixale des textes écrits, le cas de l'arabe non voyellé. Thèse d'état, Faculté des sciences de Tunis, 1993.

BEN OTHMANE ZRIBI C., BEN FRAJ F., BEN AHMED M. (2005). Un Système Multi-agent pour la Détection et la Correction des Erreurs Cachées en Langue Arabe. Actes de la 5ème conférence sur le Traitement Automatique des Langues Naturelles, Dourdan, France, 143-153.

BEN OTHMANE ZRIBI C., BEN AHMED M. (2013). Detection of semantic errors in Arabic texts. Artificial intelligence journal (195), 249-264.

BERGSMA S., PITLER E., LIN D. (2010). Creating Robust Supervised Classifiers via Web-Scale N-gram Data. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Sweden, 865-874.

CARLSON A.J., ROSEN J., ROTH D. (2001). Scaling up Context-sensitive Text Correction. Proceedings of 13th Conference on Innovative Applications of Artificial Intelligence IAAI'01, Washington, USA, 45-50.

GOLDING A.R. (1995). A Bayesian hybrid method for context-sensitive spelling correction. Proceedings of the 3rd Workshop on Very Large Corpora, Massachusetts, USA, 39–53.

GOLDING A. R., SCHABES Y. (1996). Combining trigram based and feature based methods for context sensitive spelling correction, in: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, 71-78.

GOLDING A.R., ROTH D. (1999). A winnow-based approach to context-sensitive spelling correction, Machine Learning journal, 34(1-3), 107-130.

GUTIERREZ F., DOU D., FICKAS S., GRIFFITHS G. (2014). Online Reasoning for Ontology-Based Error Detection in Text. OTM international conference on ontologies, databases and application of semantics, 562-579.

HIRST G., BUDANITSKY A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering (11), 87-111.

LANDAUER T. K., FOLTZ P.W. , LAHAM D.(1998). An introduction to Latent Semantic Analysis, Discourse Processes (25), 259-284.

MANGU L., BRILL E. (1997). Automatic Rule Acquisition for Spelling Correction, in: Proceedings of the 14th International Conference on Machine Learning, Nashville, 734-747.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G., DEAN J. (2103). Distributed representations of phrase and their compositionality. Advances in neural information processing systems, 3111-3119.

SALTON G., WONG A., YANG C. S. (1975). A vector space model for automatic indexing. Communications of the ACM (18), New York, NY, USA, 613-620.

SCHUTZE H. (1998). Automatic Word sense discrimination, Journal of Computational Linguistics (24), 97-123.

VERBERNE S. (2002). Context sensitive spell checking based on word trigram probabilities. Master thesis Taal, Spraak & Informatica, Nijmegen University.

ZESCH T. (2012). Detecting Malapropisms Using Measures of Contextual Fitness. Special Issue of the TAL Journal on “Managing Noise in the Signal: Error Handling in Natural Language Processing (53), 11-31.