Un analyseur hybride pour la détection et la correction des erreurs cachées sémantiques en langue arabe

Chiraz BEN OTHMANE ZRIBI, Hanène MEJRI , Mohamed BEN AHMED
Laboratoire de recherche RIADI, Université La Manouba
ENSI, La Manouba, Tunisie
Chiraz.benothmane@riadi.rnu.tn, Hanene.mejri@riadi.rnu.tn,
Mohamed.benahmed@riadi.rnu.tn

Résumé. Cet article s'intéresse au problème de la détection et de la correction des erreurs cachées sémantiques dans les textes arabes. Ce sont des erreurs orthographiques produisant des mots lexicalement valides mais invalides sémantiquement. Nous commençons par décrire le type d'erreur sémantique auquel nous nous intéressons. Nous exposons par la suite l'approche adoptée qui se base sur la combinaison de plusieurs méthodes, tout en décrivant chacune de ces méthodes. Puis, nous évoquons le contexte du travail qui nous a mené au choix de l'architecture multi-agent pour l'implémentation de notre système. Nous présentons et commentons vers la fin les résultats de l'évaluation dudit système.

Abstract. In this paper, we address the problem of detecting and correcting hidden semantic spelling errors in Arabic texts. Hidden semantic spelling errors are morphologically valid words causing invalid semantic irregularities. After the description of this type of errors, we propose and argue the combined method that we adopted in this work to realize a hybrid spell checker for detecting and correcting hidden spelling errors. Afterward, we present the context of this work and show the multi-agent architecture of our system. Finally, we expose and comment the obtained results.

Mots-clés : erreur cachée, erreur sémantique, détection, correction, système multi-agent, langue arabe.

Keywords: hidden error, semantic error, detection, correction, multi-agent system, Arabic language.

1 Introduction

Les erreurs cachées sont des erreurs orthographiques produisant des mots valides lexicalement et causant des dérèglements de haut niveau : syntaxique, sémantique, voire même pragmatique. Les erreurs cachées surviennent lorsqu'une ou plusieurs modifications sur un mot le transforme en un autre mot de la langue. Dans ce cas, l'erreur, est dans la plupart du temps, une graphie semblable au mot que l'utilisateur avait l'intention d'écrire.

*Le jardinier utilise le **gâteau** (râteau) pour bêcher la terre*

Dans cet exemple, le mot « *gâteau »* est introduit dans un contexte qui ne lui est pas approprié. Cette faute de frappe peut être corrigée en rétablissant le mot correct *« râteau ».*

Dans (Verberne, 2002) on lit que les statistiques réalisées pour la langue anglaise par (Eastman, Oakman, 1991) affirment que les erreurs cachées représentent 25% parmi toutes les erreurs orthographiques commises et contenues dans leur corpus de référence. (Mitton, 1987) cité par le même auteur, leur attribue une valeur plus grande à savoir : 40% parmi toutes les erreurs orthographiques étudiées. Ces deux valeurs assez importantes ont rendu l'étude de ce genre d'erreurs une nécessité en soi. Plusieurs recherches ont été entreprises dans le but de remédier à ce problème. Nous pouvons citer par exemple les recherches de Golding qui a étudié ce genre d'erreurs pour la langue anglaise. Il a ainsi proposé différentes méthodes comme la méthode de *Bayes* (Golding, 1995), la méthode des trigrammes des parties du discours (Golding, Schabes, 1996) et la méthode à base de réseaux neuronaux dite *Winnow* (Golding, Roth, 1999). Le chinois a été aussi traité avec les deux chercheurs (Xiaolong, Jianhua, 2001). Le suédois a également fait l'objet d'une recherche avec (Bigert, Knutsson, 2002).

En ce qui concerne la langue arabe, aucun autre travail n'a concerné le traitement des erreurs cachées malgré l'importance de l'entreprise d'une telle recherche. La langue arabe présente, en effet, des spécificités dont nous citons principalement : l'agglutination, l'ambiguïté grammaticale et la proximité lexicale. Toutes ces caractéristiques rendent le risque de commettre une erreur cachée plus important que pour les autres langues notamment latines.

Nous nous sommes donc intéressés à ce problème en construisant un système permettant à la fois de détecter et de corriger ce type d'erreurs pouvant survenir dans des textes arabes. Dans un premier temps ce système a concerné uniquement les anomalies syntaxiques (Ben Othmane et al., 2005). Nous l'avons amendé par la suite pour qu'il puisse traiter l'ensemble des anomalies (syntaxiques et sémantiques).

Dû à la complexité de ce travail, nous avons été amenés à émettre certaines hypothèses pour restreindre les champs de nos investigations. Nous avons considéré alors l'arabe non voyellé et ce pour une raison capitale. C'est que malgré l'importance des voyelles[1] dans la compréhension du discours arabe, elles n'apparaissent que très rarement dans les textes. Ainsi, à part quelques ouvrages poétiques ou littéraires didactiques, les écrits arabes sont généralement dépourvus de voyelles, et c'est le cas des textes fréquemment rencontrés dans les journaux, les revues, les romans, etc. Aussi, nous émettons l'hypothèse de l'existence d'une seule erreur par phrase et par mot. Cette erreur consisterait en une seule faute typographique du type : ajout d'un caractère, omission d'un caractère, substitution d'un caractère par un autre ou interversion de deux caractères adjacents. 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).

Dans ce qui suit, nous décrivons dans la première section le type d'erreurs sémantiques auquel nous nous sommes intéressés et formant ce qu'on appelle des erreurs cachées sémantiques. Dans la deuxième section, nous présentons l'approche proposée pour la conception de notre système de détection–correction erreurs cachées sémantiques. Dans la troisième section de l'article, nous abordons le contexte de notre travail, ainsi, que l'architecture d'implémentation adoptée pour la réalisation de notre système. La quatrième et dernière section est consacrée, quant à elle, à la description des résultats de l'évaluation du système mis en place.

---

[1] Signes diacritiques ajoutées aux lettres arabes pour permettre leur lecture

2 Les erreurs cachées sémantiques

Nous entendons par « erreur cachée sémantique » tout mot ressemblant typographiquement à un caractère près au mot correct qu'il remplace mais invalide sémantiquement dans le contexte où il se trouve. Les dérèglements sémantiques causées par ce type d'erreurs peuvent être réparties en deux catégories: les *incompatibilités sémantiques* et les *incomplétudes sémantiques*. Quand l'erreur cause des contresens ou encore rend la phrase dépourvue de sens, nous parlons dans ce cas d'*incompatibilité sémantique*. Quand à l'incomplétude sémantique, elle concerne principalement l'oubli de mots, de syntagmes ou d'outils de coordination nécessaires à l'interprétation de la phrase.

Nous nous intéressons ici qu'aux anomalies mettant en cause le sens. Les erreurs d'incomplétude sont plus difficiles à déceler.

يعرضون عليه أموالا كبيرة (كثيرة)

Ils lui proposent de grandes (beaucoup) d'argent

Dans cette phrase erronée, l'adjectif "كبيرة" (grandes) est utilisé au lieu de l'adjectif "كثيرة" (beaucoup) et il se trouve dans un contexte inapproprié par la substitution de la lettre ب par la lettre ث.

3 Détection des anomalies sémantiques

Pour que la machine puisse traiter la sémantique des mots, elle doit disposer, par analogie à l'être humain, des connaissances à propos du sens des mots et des différents contextes dans lesquels ils apparaissent. Ces connaissances peuvent être obtenues à partir de plusieurs ressources informatiques telles que les dictionnaires sémantiques, les thésaurus, les réseaux sémantiques, les ontologies ou les corpus textuels.

Dans le cadre de ce travail, nous optons pour une solution basée sur l'apprentissage du sens des mots à partir des corpus textuels. Cette orientation repose sur un principe de la linguistique distributionnelle qui dit que : "*le sens d'un mot peut être défini statistiquement, à partir de l'ensemble des contextes (i.e., paragraphes, phrases, textes) dans lesquels ce mot apparaît*" (Landauer et al., 1998). Par exemple, le mot *avion* apparaît souvent conjointement avec des mots comme *décoller*, *aile*, *aéroport*, et rarement conjointement avec des mots comme *lion* ou *forêt*.

Pour détecter les erreurs cachées sémantiques, nous proposons une approche qui se base sur l'étude de la validité sémantique de chaque mot du texte à analyser dans son contexte et ceci par la combinaison de plusieurs méthodes permettant de représenter chaque mot en fonction du contexte proche et lointain dans lequel il apparaît et de comparer cette représentation aux représentations antérieures obtenues lors de l'apprentissage.

Nous faisons ainsi appel à quatre méthodes, de nature statistique ou mixte (linguistique et statistique), responsables chacune de vérifier la validité sémantique d'une phrase donnée. L'idée derrière cette combinaison est d'obtenir un analyseur d'erreurs cachées sémantiques capable de tirer profit des avantages de toutes les méthodes d'analyses sémantiques proposées. Ceci implique la construction de plusieurs systèmes de traitement d'erreurs cachées qui seront mis en confrontation quant à la sélection d'une erreur cachée sémantique dans une phrase. Cette confrontation est réalisée suite à l'application d'une procédure de vote qui prendra en considération tous les résultats issus de l'application des méthodes d'analyses sémantiques proposées et procèdera à un vote pour l'identification de l'erreur la plus probable garantissant ainsi une meilleure qualité d'analyse.

Pendant la phase d'apprentissage, sont récoltées à partir d'un corpus dit d'entraînement traité au préalable[2] toutes les connaissances nécessaires aux différentes méthodes proposées et formant leurs entrées. Ce corpus[3] comporte *30* textes de type économique, et compte environ *30 000* mots, *1827* phrases et *4029* lemmes. Les connaissances extraites se présentent sous forme de données linguistiques et statistiques et varient selon les besoins de chaque méthode d'analyse utilisée.

3.1 Méthode Coocurrence-Collocation

Cette méthode vérifie la validité contextuelle d'un mot en se basant sur sa probabilité contextuelle déduite du calcul des trois mesures suivantes :

- **Probabilité de cooccurrence :** Cette probabilité est calculée pour chaque mot m_i de la phrase à analyser pour une fenêtre de 10 mots[4]. Elle est exprimée par la formule de probabilité conditionnelle de Bayes suivante :

$$P(m_i|C) = P(m_i|c_k,.....c_{-1},c_1,.....c_k) = \frac{P(c_k,.....c_{-1},c_1,.....c_k|m_i)P(m_i)}{P(c_k,.....c_{-1},c_1,.....c_k)}$$

  Où m_i représente le mot à analyser, c_i les mots voisins du contexte proche et $P(m_i)$ la probabilité d'apparition du mot m_i dans le corpus d'apprentissage.

- **Coefficient de collocation :** Une collocation est une expression ayant une structure morphosyntaxique précise et une fréquence d'apparition importante dans le corpus d'apprentissage, exemple : شوارع المدينة (les rues de la ville). Pour calculer ce coefficient nous procédons d'abord à l'identification des collocations existantes dans une phrase en se basant sur une liste de collocations obtenue lors de la phase d'apprentissage. Pour se faire, nous avons utilisé et adopté une partie du système réalisé par (Mlayeh, 2004). Lorsque une collocation est identifiée dans une phrase, un coefficient collocationnel est attribué à chaque mot de cette expression. Ce coefficient n'est autre que la mesure de Kulczinsky, qui est un critère d'association permettant d'identifier le degré de corrélation de deux lemmes l_i et l_j, calculée à l'aide de la formule suivante :

$$KUC = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$$

  Où : a : le nombre d'occurrences du couple (l_i, l_j)
  b : le nombre d'occurrences des couples où l_i apparaît non suivi de l_j
  c : le nombre d'occurrences des couples où l_j est non précède dé l_i

  La valeur de ce coefficient varie entre 0 et 1 et il est égal à 0,5 quand l_i est toujours observé avec l_j. Une expression est considérée comme collocation si son coefficient de KUC est supérieur à 0,5.

- **Probabilité de répétition:** "*les mots ou plus précisément les lemmes des mots d'un texte ont tendance à se répéter dans le texte lui-même*". Cette hypothèse est déduite des comptages réalisés par (Ben Othmane, Ben Ahmed, 2003) sur un corpus textuel en langue arabe appartenant à un domaine particulier qui montrent qu'une forme

---

[2] Analysé morpho-syntaxiquement, découpé en phrases et en syntagmes nominaux et verbaux.

[3] Ces textes proviennent à l'origine du corpus de l'arabe contemporain collecté par Al-Sulaiti L. http://www.comp.leeds.ac.uk/eric/latifa/arabic-corpora.htm. Ils ont été choisis par ce qu'ils sont relatifs à un même domaine.

[4] La taille de la fenêtre est paramétrable et peut être facilement ajustée.

textuelle apparaît en moyenne 5,6 fois dans un même texte alors qu'un lemme apparaît en moyenne 6,3 fois et ce dans le même texte. Subséquemment, si le lemme d'un mot se répète très peu dans le texte, le mot en question peut correspondre à une erreur cachée. Cette probabilité concerne donc le taux d'apparition de chaque lemme des mots de la phrase, objet de vérification, dans le corpus de test. Ce taux est calculé par la formule suivante :

$$P(l_i) = \frac{\text{nombre d' occurences de } l_i}{\text{nombre total de lemmes}}$$

La combinaison de ces trois mesures en vue de l'obtention de la probabilité contextuelle $P(m_i)$ de chaque mot de la phrase se fait selon la formule linéaire suivante :

$$P(m_i) = \alpha * P(m_i|C) + \beta * KUC(m_i) + \delta * P(l_i)$$

Où $P(m_i|C)$ est la probabilité de cooccurrence du mot m_i, $KUC(m_i)$ est le coefficient collocationelle attribué à un mot m_i, $P(l_i)$ est la probabilité de répétition pour un lemme l_i du mot m_i. α, β, et δ sont des poids attribués aux différentes probabilités afin de mettre en évidence la contribution de chaque probabilité. Il est à noter que ces valeurs ne sont pas connues à l'avance et sont déterminées lors des expérimentations[5]. Toutefois, nous estimons que la valeur de α doit être plus importante que celles de β, et δ vu que le contexte voisin est plus déterminant pour le sens du mot à analyser que son contexte lointain.

Une fois les probabilités relatives à tous les mots de la phrase en question sont calculées, elles seront comparées à une valeur *seuil* déterminé lors des expérimentations. Le ou les vocables ayant une probabilité inférieure à ce *seuil* forment une liste d'erreurs cachées éventuelles.

3.2 Méthode Vecteur-Contexte

Cette méthode consiste à représenter chaque mot de la phrase par un vecteur en fonction du contexte dans lequel il apparaît. De ce fait, un vecteur mot Vm_i n'est autre qu'une représentation vectorielle de la probabilité de cooccurrence de ce mot avec chaque mot de la phrase. Considérons par exemple, la phrase suivante :

شرب الرجل كلبا(كأسا)

L'homme a bu un chien (un verre)

La matrice ci-dessus illustre la probabilité de cooccurrence de chaque mot m_i de la phrase avec les mots voisins de ce même contexte. Les colonnes de la matrice représentent les mots m_i et les lignes représentent les composantes du vecteur Vm_i. Ainsi, une cellule contient la probabilité de cooccurrence du mot m_i avec le mot m_j, calculée selon la formule suivante:

$$P(m_i \mid m_j) = \frac{\text{nombre de fois où } m_i \text{ et } m_j \text{ cooccurent}}{\text{nombre d' occurrence de } m_j}$$

$V_{\text{كلبا}}$

	كلبا	الرجل	شرب
شرب	0,3	0,6	
الرجل	0,1		0,6
كلبا		0,1	0,3

Tableau 1 : Matrice de cooccurrence des mots d'une phrase

[5] Pour nos expérimentations nous avons choisi: α=2, β=1 et δ =0,5.

Pour représenter le degré de corrélation de chaque mot m_i avec tous les autres mots m_j de la phrase, nous proposons de calculer la norme de chaque vecteur Vm_i exprimée comme suit :

$$\|Vm_i\| = \sqrt{\sum_{j=1}^{n} c_j^2}$$

Où c_j est la probabilité de cooccurrence du mot m_i avec le mot m_j de la phrase. Dans l'exemple précédent, les normes des vecteurs des mots شرب, الرجل, كلبا sont respectivement égales à 0,67 ; 0,6 et 0,31. Le mot ayant la norme la moins élevée est كلبا, est soupçonné d'une erreur cachée. D'une manière générale, nous évaluons la norme de chaque vecteur mot Vm_i à une valeur *seuil*. Le ou les mots ayant une norme inférieure au *seuil* sont ajoutés à la liste des mots suspectés.

3.3 Méthode Vecteur-Vocabulaire

Le vocabulaire (termes représentatifs) d'un texte ou d'un domaine en question est un élément caractéristique de ce dernier et un bon indicateur de la cohérence de ce texte. Nous pouvons, par conséquent et en adoptant le principe de représentation vectorielle précédemment cité, étudier la validité sémantique d'une phrase en représentant chaque mot lui appartenant par un vecteur en fonction de sa probabilité de cooccurrence avec le vocabulaire. Pour évaluer la proximité entre deux vecteurs, nous utilisons la métrique de distance angulaire exprimée comme suit :

$$Dist(Vm_i, Vm_j) = \arccos(Sim(Vm_i, Vm_j))$$

$$Sim(Vm_i, Vm_j) = \cos(Vm_i, Vm_j) = \frac{Vm_i Vm_j}{\|Vm_i\| * \|Vm_j\|} = \frac{\sum_{t=1}^{n} Vm_{it} Vm_{jt}}{\sqrt{\sum_{t=1}^{n} Vm_{it}^2 \sum_{t=1}^{n} Vm_{jt}^2}}$$

Le calcul de la distance angulaire se fait pour chaque vecteur mot m_i par rapport à tous les autres vecteurs mot m_j de la phrase. Le vecteur le plus éloigné du contexte correspond au mot qui apparaît le moins avec les mots du vocabulaire en corrélation avec le contexte courant. Pour sélectionner ce vecteur, la somme des distances angulaires de chaque vecteur mot m_i est calculée puis comparée à une valeur *seuil*. Le ou les mots qui correspondent à la somme des distances la plus élevée et supérieur au seuil sont soupçonnés d'erreurs cachées.

3.4 Méthode LSA

"LSA (Latent semantic Analysis : Analyse sémantique latente) est une méthode permettant l'acquisition des connaissances à partir de l'analyse entièrement automatique de grands corpus textuels" (Landauer et al., 1998). Plus précisément, cette méthode 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.

Le principe de la méthode LSA consiste à représenter les mots dits unités lexicales et les segments textuels (phrases, paragraphes, textes) dits unités textuelles par des vecteurs dans un espace vectoriel de dimensions réduites par rapport à l'espace d'origine et le mieux représentatif de ce dernier. L'espace d'origine est représenté par une matrice de cooccurrence initiale *X(m, n)* représentative du corpus d'apprentissage où les *m* lignes correspondent aux unités lexicales, et les *n* colonnes aux unités textuelles. Une cellule contient le nombre d'occurrences d'une unité lexicale dans une unité textuelle. Cette matrice est décomposée en produits de trois matrices *T(m,t)*, *S(t,t)* et *D(t,n)* grâce à une forme d'analyse factorielle appelée décomposition en valeurs singulières. La matrice *T* est une matrice orthogonale de *m*×*t* dimensions, *D* est une matrice orthogonale de *t*×*n* dimensions et *S* est une matrice

diagonale de $t \times t$ dimensions dite aussi matrice de valeurs singulières. Les valeurs de cette dernière représentent les dimensions de l'espace d'origine.

Dans notre cas, la matrice *X* a été construite durant la phase d'apprentissage. Les lignes correspondent aux lemmes dudit corpus, et ils sont au nombre de **4029**, les colonnes représentent les phrases dont le nombre est **1827**. La réduction des dimensions consiste à choisir parmi les *n* dimensions les *k* dimensions les plus pertinentes et les plus représentatives de l'espace d'origine à partir de la matrice diagonale *S* triée selon l'ordre de ses valeurs singulières. Ainsi, nous obtenons trois matrices *T (m,k)*, *S(k,k)* et *D(k,n)* de dimensions réduites (k=300 valeur choisie après plusieurs tests). Le produit scalaire de ces matrices génère la matrice *X'(m,n)* représentative de l'espace résultat.

La variante de la méthode *LSA* que nous proposons étudie la validité sémantique des mots d'une phrase donnée en comparant leurs vecteurs sémantiques extraits de la matrice de cooccurrence transformée et obtenue lors de la phase d'apprentissage. Pour mesurer la proximité sémantique entre les vecteurs issus de la matrice obtenue, nous utilisons, comme le cas de la méthode *Vecteur-Vocabulaire,* la métrique de distance angulaire. Ainsi, chaque vecteur sémantique Vm_i du mot m_i est comparé à tous les vecteurs Vm_j des mots m_j du contexte en fonction de la distance angulaire. La somme de ces distances est ensuite calculée pour chaque mot m_i et comparée à une valeur *seuil*. Si cette valeur est supérieure au *seuil*, le mot correspondant est soupçonné d'une erreur cachée.

3.5 Procédure de vote

Étant donné que notre système global de détection d'erreurs cachées se base sur l'hypothèse stipulant une erreur au plus par phrase et que les prétendues erreurs sont toujours classées par ordre de probabilité décroissante, nous avons choisi un vote de type *uninominal par classement* (les candidats sont triés et un seul parmi eux sera élu). Nous présentons dans ce qui suit le principe de la méthode que nous avons adoptée par notre procédure de vote.

1. Compter le nombre d'occurrences des différentes erreurs proposées par toutes les méthodes d'analyses sémantiques présentes dans chaque liste et se trouvant au premier rang.
2. Sélectionner les erreurs qui ont recueilli le plus grand nombre d'occurrences. Si une seule erreur obtient la majorité absolue du nombre d'occurrences, elle est élue comme étant l'erreur la plus probable dans la phrase. Sinon, on calcule une nouvelle valeur d'occurrences des erreurs retenues au rang suivant.
3. Ce processus se répète autant de fois jusqu'à ce qu'une seule erreur ayant la majorité absolue d'occurrences soit retenue.

Toutefois, la méthode de vote proposée peut conduire parfois à une situation de blocage où le nombre d'occurrence de deux ou plusieurs erreurs sélectionnées en premier rang reste toujours invariant. Dans ce cas, nous nous référons au *degré de confiance* attribué à chaque méthode afin de sélectionner, parmi la liste des erreurs retenues, celle détectée par la méthode du plus grand degré de confiance.

4 Correction des erreurs cachées sémantiques

Pour corriger les erreurs cachées, nous procédons à la génération de toutes les formes proches de la forme erronée, à un caractère d'édition près pour former ainsi une liste contenant les

candidats à la correction. Nous avons utilisé et adapté à cet effet un correcteur orthographique développé par (Ben Othmane, 1998).

Comme nous nous attendons à avoir un grand nombre de propositions, dû à la proximité lexicale de la langue arabe, nous avons pensé réduire cette liste. L'idée étant de substituer la forme erronée par chacune des formes proposées et former ainsi un ensemble de phrases candidates. Ces dernières seront soumises à notre détecteur d'erreur sémantique. Celles qui produisent des dérèglements dans la phrase seront éliminées et c'est le même sort que subissent leurs propositions respectives. La liste des propositions restantes est par la suite triée par ordre de pertinence et présentée à l'utilisateur.

5 Contexte de travail

Ce travail vient compléter nos recherches précédentes (Ben Othmane et al., 2005) qui ont concerné le problème d'erreurs cachées (syntaxiques et sémantiques) pouvant se produire dans un texte en langue arabe. Le système qui a été proposé pour le traitement de ces erreurs est à base d'agents. Ce système (SMA) se compose principalement d'un agent pour la correction et de deux groupes d'agents pour la détection : un groupe d'agents syntaxiques permettant de traiter les anomalies syntaxiques pouvant se produire dans une phrase donnée et un groupe d'agents sémantiques permettant de traiter les incohérences sémantiques. Seul l'agent correction et le groupe d'agents syntaxiques ont été bien étudiés et implémentés, nous venons donc compléter par notre travail la partie sémantique. La figure 1 illustre l'architecture globale du système de traitement des erreurs cachées.

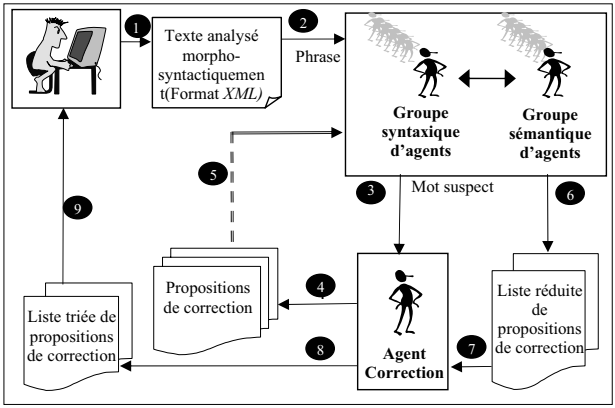

Figure 1 : Architecture du système global de détection et correction des erreurs cachées

Nous avons ainsi implémenté notre vérificateur sémantique sous forme d'un groupe d'agents sémantiques, où chaque méthode proposée est appliquée par un agent spécifique. En plus, un agent *Superviseur* du groupe est chargé de l'activation des différents sous agents sémantiques responsables d'analyser la phrase en cours et de détecter les incohérences sémantiques qu'elle peut renfermer. Les agents sémantiques travaillent en parallèle et communiquent leurs résultats à l'agent *Superviseur* qui joue en plus, dans ce cas, le rôle de décideur en sélectionnant l'erreur la plus probable parmi l'ensemble des listes d'erreurs détectées par les différents agents en appliquant la procédure de vote.

6 Expérimentations et résultats

Pour l'évaluation de notre système, nous avons choisi un texte de test de même type et appartenant au même domaine que le corpus d'apprentissage utilisé. Il compte 1 564 mots, 100 phrases dont 50 contiennent une erreur cachée.

La figure suivante illustre les performances de chaque agent, ainsi, que du système global de détection des erreurs cachées sémantiques en terme de précision.

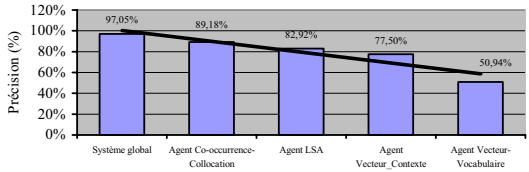

Figure 2 : Performances du système de détection des erreurs cachées sémantiques

Le taux de précision le plus élevé pour l'ensemble des agents sémantiques est celui de l'agent *Cooccurrence-Collocation* avec une valeur de **89,18%**. Cette performance s'explique par la complémentarité des phénomènes de cooccurrence, de collocation et de répétition. Contre toute attente, le taux fourni par l'agent *LSA* (**82,92%**) s'avère plus faible ; ceci est dû sans doute à la modestie de nos données d'apprentissage qui cause un taux élevé de sur-détection d'erreurs. Toutefois, la méthode *LSA* reste toujours prometteuse par rapport aux méthodes basées uniquement sur les cooccurrences des mots. En effet, le taux de précision de l'agent *Vecteur-Contexte*, est relativement faible (**77,5%)** et celui de l'agent Vecteur_Vocabulaire n'est pas bon (**50,94%**). L'amélioration des résultats de ces derniers nécessiterait à notre avis un grand corpus d'apprentissage, une stratégie d'extraction du vocabulaire du domaine plus fiable et une sélection fine et bien étudiée des textes formant le corpus d'apprentissage. Pour ce qui est du résultat de l'évaluation du système global, nous pouvons dire que le taux de précision qui est égal à **97,05%** est très satisfaisant**.** La performance du système de vote et son apport quant à la sélection de l'erreur la plus probable dans la phrase se confirment donc.

Quant à la phase de correction, elle a été testée à deux niveaux ; d'abord après l'obtention de toutes les propositions de correction, ensuite après la minimisation de la liste de ces propositions. Les résultats obtenus sont illustrés dans le tableau ci-après.

	Couverture	Précision	Ambiguïté	Proposition	Position
Initialement	100%	100%	100%	46,67	13,82
Minimisation	100%	80%	80%	5,98	3,43

Tableau 2 : Performance du système de correction des erreurs cachées sémantiques

Nous remarquons que notre méthode de minimisation de la liste des propositions a permis de réduire, considérablement (98%), le nombre moyen des propositions (46,67 à 5,98 propositions en moyenne). Cette diminution, bien qu'elle ait réduit l'ambiguïté de notre correcteur de 20%, ne s'est pas passée sans dégât. Elle s'est faite au dépend de la précision (diminution de 20%).

7 Conclusion

Notre système de détection d'erreurs cachées sémantiques a donné des résultats satisfaisants (taux de précision de **97,05%**) en dépit des contraintes et des restrictions liées à la taille ainsi qu'à la non diversité de nos données d'apprentissage. Nous signalons, aussi, l'apport de la démarche suivie pour la correction de la forme erronée qui a permis de minimiser la liste des propositions de correction de **98%** et d'avancer la forme correcte aux premiers rangs. Cependant, nous estimons que les résultats obtenus peuvent être encore améliorés d'abord par l'utilisation d'un bon corpus d'apprentissage de nature plus varié et de taille plus importante. D'autres perspectives proches sont également en vue, nous pensons effectivement intégrer les deux groupes d'agents syntaxiques et sémantiques ensemble afin de former le système global de traitement des erreurs cachées en langue arabe.

Références

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

BEN OTHMANE Z. C. (1998). De la synthèse lexicographique à la détection et la correction des graphie fautives arabes. Thèse de doctorat, Université de Paris XI, Orsay.

BEN OTHMANE Z. C., BEN AHMED M. (2003). Le contexte au service des graphies fautives arabes. TALN'03, Batz-sur-Mer.

BEN OTHMANE Z. C., BEN FRAJ F., BEN AHMED M. (2005). Un système multi-agent pour le traitement des erreurs cachées en langue arabe. Actes de la 12ème Conférence sur le Traitement Automatique des langues naturelles TALN'05, Dourdan, vol. 1, p. 143-153.

BIGERT J., KNUTSSON O. (2002). Robust Error Detection : A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge. In Proceedings of Robust Methods in Analysis of Natural Language Data (ROMAND'02), Frascati, Italie.

GOLDING A. (1995). A Bayesian hybrid method for context-sensitive spelling correction. In Proceedings of the third Workshop On Very Large Corpora, Cambridge, Massachuses, USA, (1995), 39-53.

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

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

MLAYEH I. (2004). Extraction de collocations à partir de corpus textuels en langue arabe. Mémoire de mastère, Ecole nationale des sciences informatiques, Université de la Manouba.

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

XIAOLONG W., JIANHUA L. (2001). Combine trigram and automatic weight distribution in Chinese spelling error correction. Journal of computer Science and Technology, Volume 17 Issue 6, Province, China.