Meta
Journal des traducteurs
Translators' Journal

META

Outil d'extraction de la sémantique d'un corpus textuel

Eugène Sandford et Sylvain Fraïssé

Volume 42, numéro 2, juin 1997

Lexicologie et terminologie II (1) et Traduction et post-colonialisme en Inde
Translation and Postcolonialism: India (2)

URI : https://id.erudit.org/iderudit/004027ar
DOI : https://doi.org/10.7202/004027ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Sandford, E. & Fraïssé, S. (1997). Outil d'extraction de la sémantique d'un corpus textuel. *Meta*, *42*(2), 356–363. https://doi.org/10.7202/004027ar

Résumé de l'article

La méthode proposée a pour but l'extraction du sens des textes analysés. Ceux-ci sont préalablement soumis aux analyses morphologique et syntaxique développées sur le système SYGMART afin d'obtenir tous les renseignements concernant les fonctions et places des mots dans la phrase. Cette analyse pré-sémantique donne certaines ambiguïtés qui ne peuvent être levées que par une étude du sens. Ce travail s'inscrit dans le cadre d'une analyse globale de la langue écrite où l'on associe à une structure syntaxique une composante sémantique.

OUTIL D'EXTRACTION DE LA SÉMANTIQUE D'UN CORPUS TEXTUEL[1]

EUGÈNE SANDFORD ET SYLVAIN FRAÏSSÉ*
LIRMM, Montpellier, France

Résumé

La méthode proposée a pour but l'extraction du sens des textes analysés. Ceux-ci sont préalablement soumis aux analyses morphologique et syntaxique développées sur le système SYGMART afin d'obtenir tous les renseignements concernant les fonctions et places des mots dans la phrase. Cette analyse pré-sémantique donne certaines ambiguïtés qui ne peuvent être levées que par une étude du sens. Ce travail s'inscrit dans le cadre d'une analyse globale de la langue écrite où l'on associe à une structure syntaxique une composante sémantique.

Abstract

This article proposes a method for extracting meaning of texts. Morphological and syntactic analyses of texts using the SYGMART system provide relevant information concerning function and place of words in a sentence. This presemantic analysis can give rise to certain ambiguities which can be clarified only through meaning analysis. This task is part of a global analysis of written language, an analysis in which a syntactic structure is associated with a semantic component.

1. INTRODUCTION

Dans le cadre du traitement automatique de la langue naturelle, après des recherches pour obtenir des analyses morphosyntaxiques, il faut s'intéresser maintenant à la désambiguïsation sémantique couplée avec cette analyse morphosyntaxique. Il existe quelques méthodes pour représenter la signification des mots dans l'ordinateur.

Historiquement, dans un domaine adjacent, la recherche d'informations dans les corpus des textes, qui existe maintenant en tant que domaine d'applications depuis une quinzaine d'années, Salton et McGill (1983) avec un traitement statistique dépourvu de syntaxe qui s'appuie sur des mots clés (Salton 1988), sont arrivés à des résultats de comparaison sémantique qui sont, d'après Niwa et Nitta (1995), meilleurs que ceux donnés par la recherche de vecteurs distance entre les mots dans des dictionnaires. Cela dit, la conception de Salton repose sur une sémantique rudimentaire (Rastier 1987), et nous pouvons espérer un meilleur formalisme.

Les travaux des sémanticiens logiciens, comme Sowa (1984), ne correspondent pas à notre approche sur la désambiguïsation. En effet, le passage du texte au formalisme qu'ils proposent est loin d'être évident, même s'il y a eu des tentatives comme la grammaire de Montagu et la logique intentionnelle typée.

Des travaux en cours aux États-Unis semblent prometteurs pour la désambiguïsation sémantique. En effet, en travaillant sur WordNet, une énorme taxonomie des mots anglais, Resnik (1995), par exemple, cherche une distance entre les mots à partir de cette hiérarchie. Le seul ennui est qu'en français, nous ne disposons pas de ce matériel, et il semble que nous n'en disposerons pas avant des années.

Dans cet article, un mot est représenté par un vecteur de l'ensemble des sens qui le composent. Ces sens sont donnés par une traduction du *Roget's Thesaurus of English Words*

and Phrases (1987), consensuellement reconnu dans le monde littéraire, des concepts élémentaires généraux dont le nombre est fixe.

Ici, nous définirons comme Yarowsky (1992) ce que nous appellerons **sens** ou **concepts élémentaires**. Les sens d'un mot ou encore les concepts élémentaires d'un mot sont donnés par la liste des catégories listées pour ce mot dans le thésaurus de Roget.

À partir de cette projection des sens des mots sur un espace de concepts élémentaires, on peut construire des vecteurs pour les phrases et les textes en faisant la somme des vecteurs associés aux mots, avec un coefficient calculé suivant la place que le mot ou le groupe de mots occupe dans la phrase. Il est donc nécessaire de connaître la fonction syntaxique des mots dans la phrase. Une analyse morphosyntaxique permet de récupérer les mots non vides de sens du document en utilisant un lexique, et ceux-ci sont projetés sur l'espace de concepts élémentaires qui est la base de connaissances de la désamguïsation sémantique. Celle-ci sert à lever les ambiguïtés du texte donnés par les analyses morphologique et syntaxique, ainsi qu'à connaître les rapprochements sémantiques entre les mots.

Dans le cadre d'une application de cette désambiguïsation sémantique, nous l'avons couplée à l'analyse syntaxique afin de générer une application pour la navigation dans un corpus de textes dirigée par thèmes. Cette application est en cours de réalisation.

Voyons, avant d'aller plus loin, comment nous décrivons cette représentation du signifié des mots.

2. L'APPROCHE

2.1. La méthode

2.1.1. Application Vecteur Sémantique

Un texte est composé de paragraphes, de phrases, et de mots.

Le but est de construire une représentation sémantique d'un texte, des paragraphes, des phrases et des mots.

Cette représentation est construite au-dessus de l'analyse syntaxique pour lever les ambiguïtés que ne manquent pas d'apporter les analyses syntaxique et morphologique.

Au niveau sémantique, ces ambiguïtés se traduisent par du bruit ou du silence.

Les textes à traiter sont analysés syntaxiquement avec entre autres moyens un lexique de mots (dictionnaire). Ces lexies sont alors projetées sur l'espace vectoriel de concepts pour avoir leurs significations.

Pour couvrir une représentation des sens d'un texte, une application vectorielle *Vs* (pour vecteur sémantique) de l'ensemble des textes analysés vers un espace vectoriel de concepts est nécessaire.

L'ensemble Source de l'application Vecteur Sémantique (*Vs*) est l'ensemble L des lexies signifiantes des textes traités, pour l'instant. Plus loin, dans la sous-section 2.4., l'application *Vs* sera étendue à l'ensemble P des phrases et à l'ensemble T des textes du corpus considéré.

Nous associons aux signifiants sortant de l'analyse syntaxique leurs signifiés représentés comme vecteurs, éléments de l'espace vectoriel des concepts.

2.1.2. Utilisation d'un thésaurus

L'ensemble Arrivée de *Vs* est un espace vectoriel de concepts, sa dimension est N ($N > 1$). Chaque vecteur unitaire e_i de la droite dimensionnelle D_i est l'image d'au moins un concept parmi les Nc concepts, $Nc = 990$, décrits par le thésaurus de Roget qui est la base conceptuelle utilisée ici. Cependant, pour rester cohérent avec la définition d'un espace vectoriel qui exige l'orthogonalité des droites dimensionnelles, on a $N < Nc$, car il existe des concepts du thésaurus qui sont opposés, par exemple, les numéros **189. Présence** et **190. Absence** du même thésaurus.

Ce livre de concepts de référence est divisé en deux parties.

1. La première partie numérote, classe et donne une liste de tous les concepts principaux trouvés par Roget et augmentés par Kirkpatrick. Il y en a 990 dans l'édition de 1987. Après le numéro et le nom du concept suivent des synonymes ou des lexies dont le sens est proche ou dérivé.

2. La seconde partie répertorie une très grande majorité des mots (lexies simples) et des groupes de mots (lexies complexes) usuels ou moins usuels, et en fait une projection sur les concepts référencés dans la première partie. On obtient, pour une lexie, une liste de couples (concept, numéro de concept). Soit l'exemple :

Entité :
.(Existence, 01).
.(Substance, 02).
.(Tout, 52).
.(Unité, 88).

2.1.3. Une composition d'applications

L'introduction de ces concepts généraux donnés par Roget et augmentés par Kirkpatrick en 1987 montre que l'application *Vs* recherchée est la composée de deux applications, *Sens* et *vect*, décrites ci-après et représentées par la figure 1 donnée en annexe.

La première application, *Sens*, est définie de L vers l'ensemble S des sens du mot donnés par la deuxième partie du thésaurus de Roget.

Cette application associe à chaque lexie son signifié, *i.e.* la liste des couples (concept, numéro de concept) donnés dans la seconde partie du thésaurus décrite plus haut.

La deuxième application *vect* a pour Source l'ensemble S et pour But l'espace vectoriel C des concepts vectoriels. On peut noter que le mot *Entité* a un poids sémantique différent suivant le concept donné. Ainsi, pour une *Entité*, l'**existence** semble plus importante que la **substance**. C'est pourquoi des coefficients vectoriels a_i seront donnés aux différents vecteurs images $vect(x_i)$, pour tout x_i dans S, ces coefficients étant définis comme ceci :

- association forte sur le concept : Coefficient vectoriel associé : +1 ;
- association faible sur le concept : Coefficient vectoriel associé : +1/2 ;
- aucune association sur le concept : Coefficient vectoriel associé : 0.

Soit, pour le mot *entité* :

$$\overrightarrow{vect}(Sens(entite)) = +1.\overrightarrow{vect}(existence) + \frac{1}{2}.\overrightarrow{vect}(substance) + \ldots$$

D'où, l'application *Vs* peut être comme ceci :

$$\overrightarrow{Vs} : \mathcal{L} \xrightarrow{\text{Sens}} S \xrightarrow{\overrightarrow{vect}} C \text{ avec } Dim(C) = N \text{ et } 1 < N < N_C.$$

Soit x élément de L,

$$\overrightarrow{Vs}(x) = \overrightarrow{vect}(Sens(x)) = \sum_{1 \le i \le N} a_i.\overrightarrow{vect}(c_i) \text{ avec } c_i \in \{\text{concepts du ROGET's}\}.$$

Soit e_i le vecteur unitaire directeur de la droite dimensionnelle D_i de l'espace des concepts C, on a :

$$\overrightarrow{vect}(c_i) = \sum_{1 \le i \le N} a'_i . \vec{e}_i$$

En posant $b_i = a_i . a'_i$, on a :

$$\vec{Vs}(x) = \sum_{1 \le i \le N} b_i . \vec{e}_i .$$

2.1.4. Extension de l'application Vecteur Sémantique aux phrases et aux textes

Maintenant que le vecteur Vs(x) est construit, pour tout x dans L, voici comment est étendue l'application Vecteur Sémantique d'abord aux phrases puis, dans un deuxième temps, aux textes. Soit la phrase «L'homme mange les pommes». La structure syntaxique[2] créée par l'analyseur morphologique et syntaxique à l'aide du système SYGMART (Chauché 1984) est donnée par la figure 2 (en annexe).

Soit J, l'ensemble des lexies présentes dans une phrase $phrase_j$.

$$\vec{Vs}(phrase_j) = \sum_{lexie \in \mathcal{J}} \left(\frac{1}{2}\right)^{i_{lexie}} . \vec{Vs}(lexie) ,$$

Si la lexie est un *nom* ou un *adjectif*, i_{lexie} est le nombre de GN (ou de GNPREP) moins 1, qui existent en remontant dans la structure syntaxique de reconnaissance de la phrase à partir du premier GN trouvé jusqu'à la racine PH de la phrase. Soit dans l'exemple, les lexies *homme* et *pommes* ont même i_{lexie} c'est-à-dire 0. Par contre, le mot *campagne* a pour valeur de i_{lexie}, 1. Cette dernière lexie aura moins d'influence que le mot *homme* sur le vecteur sémantique de la phrase.

Si la lexie est un *verbe*, i_{lexie} est traitée dans l'indexation du verbe. Cela dit, des travaux linguistiques (Schmitt *et al.* 1992) montrent que ce sont les groupes nominaux (GN ou GNPREP) qui sont les plus pertinents dans la signification de la phrase, donc les verbes auront une i_{lexie} plus grande que celle des noms ou adjectifs de même niveau dans l'arborescence. D'où, le verbe *mange(r)* a une i_{lexie} de 1.

Il est évident que dans une phrase complexe avec des relatives et des conjonctives, cela n'est pas aussi trivial. Il faut traiter les relatives et les conjonctives localement, avant d'en tenir compte dans la phrase qui les inclut.

L'idée, pour réaliser cette extension, est de prendre en compte l'importance des groupes de lexies signifiantes principales par rapport à celles secondaires qui sont là pour spécifier les premières : Dans *bateau à voile*, *bateau* est la lexie principale, et *voile* est là pour spécifier de quel type de bateau il s'agit, ce qui introduira des notions de *vent*, etc. Il ne faut pas que ces notions aient plus de poids que celles amenées par le groupe principal.

Pour les vecteurs sémantiques des textes, il reste à faire la somme des vecteurs sémantiques des phrases présentes dans le texte.

Voici donc étendue l'application *Vs* :

$$\vec{Vs} : \mathcal{L} \cup \mathcal{P} \cup \mathcal{T} \xrightarrow{\text{Sens}} S \xrightarrow{\overrightarrow{vect}} C.$$

2.1.5. Cosinus de deux vecteurs sémantiques

Le cosinus de l'angle de deux vecteurs sémantiques V_1 et V_2 est par définition le produit scalaire de ces deux vecteurs, soit :

$$\cos(\widehat{\vec{V}_1, \vec{V}_2}) = \frac{\vec{V}_1 . \vec{V}_2}{\| \vec{V}_1 \| . \| \vec{V}_2 \|}$$

Définition de la synonymie par rapport à notre espace de concepts :

Un mot lexie$_1$ est *synonyme* d'une autre lexie$_2$ si leurs sens sont les mêmes ou sont proches l'un de l'autre. Ce qui donne dans notre espace vectoriel de concepts :

Un mot lexie$_1$ est *synonyme* d'une autre lexie$_2$ si et seulement si les vecteurs sémantiques associés V_1 et V_2 sont tels que :

$$\cos(\widehat{\vec{V}_1, \vec{V}_2}) = 1 \textit{ ou tend vers } 1 \; (\forall \varepsilon, 0 < \varepsilon < 1, \cos(\widehat{\vec{V}_1, \vec{V}_2}) > \varepsilon).$$

Le produit scalaire des vecteurs concernés fait le produit des coefficients non nuls situés sur les droites dimensionnelles, c'est-à-dire les coefficients des concepts communs aux deux vecteurs car les autres s'annulent. Cela revient à calculer les poids des concepts communs à deux entités qui peuvent être une lexie, une phrase, ou un texte. Le *cosinus* est représenté graphiquement sur la figure 3 située en annexe.

2.2. En pratique

2.2.1. Le déroulement

En sortie de l'analyse syntaxique développée sur Sygmart par Chauché (1984), les arbres syntaxiques correspondant aux textes traités sont obtenus. Pendant la même phase, à chaque mot traité correspondent un ou plusieurs vecteurs sémantiques chargés à partir d'un lexique, correspondant à un ou plusieurs signifiés.

La rentrée des *Vs* dans ce lexique s'explique ainsi : dans un premier temps, les tuples relatifs aux lexies sont rentrés tels qu'ils sont donnés dans le Roget. Puis en première approximation, dans le *Vs* de la lexie, les concepts opposés sont corrigés pour ne conserver que l'un des deux, pris arbitrairement initialement. Par exemple, on ne conserve que **189. Présence** au lieu de **190. Absence.** Toutes les lexies contenant le concept **Absence** voient leur composante **189** du *Vs* sur la dimension du concept **Présence** s'opposer automatiquement. Le résultat est alors rentré dans ce lexique. Les tuples de vecteurs sémantiques pour chaque mot sont rentrés manuellement.

Tous les *Vs* des lexies sont récupérés pour former alors les *Vs* des phrases puis des textes, comme décrit plus haut dans cet article. Tous les *Vs* sont normés, ce sont donc des vecteurs moyens.

Le *Vs* moyen d'un texte (ou d'une phrase) représente les directions conceptuelles, et la norme permet de mettre tous les textes (ou les phrases) quelle que soit leur taille, à la même dimension. Ce qui permet une comparaison entre vecteurs associés à des textes (ou des phrases) de taille différente.

Le principe des comparaisons des *Vs* repose sur le cosinus de 2 vecteurs, donc le cosinus d'un angle.

À chaque entité (lexie, phrase, texte) est associée une liste des couples (valeur de cosinus, index de l'entité) triée par valeur décroissante de cosinus des vecteurs associés à l'entité.

En regroupant les comparaisons dans des tableaux sont obtenus des tableaux de comparaison mots / mots, mots / phrases, phrases / phrases, phrases / textes et textes / textes.

Donc, pour chaque entité, le premier élément le plus proche de l'entité correspond à l'entité la plus proche sémantiquement.

2.2.2. Levée d'ambiguïtés

En prenant tous les *Vs* différents des lexies et des phrases, le calcul du *Vs* moyen d'un texte est fait en tenant compte de tous les sens possibles de chaque mot.

Dans le cas d'ambiguïtés de sens des mots, les directions communes dans le *Vs* moyen du texte se renforcent, et les directions non significatives pour le *Vs* moyen du texte émergent de la masse par leur différence de poids et peuvent alors être neutralisées.

Comme le sens le plus probable d'un mot ambigu est celui qui est le plus proche de celui du sens moyen du texte dans lequel il intervient, on peut enlever les sens des mots ambigus qui sont peu corrélés à la direction moyenne, et cela permet un recalcul plus fin du *Vs* moyen du texte *a posteriori*.

3. CONCLUSION

L'approche choisie sur la désambiguïsation sémantique est basée sur l'association des lexies signifiantes à leurs signifiés représentés comme des vecteurs dans un espace vectoriel de concepts élémentaires. Cet espace est la représentation vectorielle des concepts généraux donnés par le thésaurus de Roget, édition de 1987.

Ces vecteurs sémantiques associés aux lexies permettent, en tenant compte de la syntaxe des phrases, de construire des vecteurs sémantiques associés aux phrases.

Enfin, les textes eux-mêmes peuvent être représentés par des vecteurs sémantiques qui sont alors les sommes des vecteurs sémantiques des phrases des textes. Cette association permet non seulement de montrer l'existence ou non de concepts dans un texte, mais aussi de créer une pondération dans chaque dimension conceptuelle présente par l'addition de deux facteurs :

- la force des sens dans le mot ;
- la nature et la place occupée par ce mot dans la phrase, donnée par la structure syntaxique des textes.

Les poursuites de ce travail se feront sur une recherche de distance entre les mots des textes traités plutôt que d'une similarité qui ne peut nous donner une classification des mots par rapport aux différents concepts. Il semble évident que les concepts élémentaires ont une influence les uns sur les autres. Comment en tenir compte dans notre espace vectoriel ?

Et enfin, il nous faut approfondir les conséquences de l'intervention des modalités dans la signification des phrases.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV^es Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).
1. Ce travail a été financé avec l'aide du Gouvernement Territorial de la Polynésie Française.
2. C'est l'arborescence de dérivation de la grammaire d'analyse syntaxique qui reconnaît ou non les phrases passées en entrée.

RÉFÉRENCES

CHAUCHÉ, J. (1984) : «Le système Sygmart», *COLING' 84*, Stanford.
KIRKPATRICK, B. (1987) : *Roget's Thesaurus of English Words and Phrases*, Penguin Books.
NIWA, Y. et Y. NITTA (1995) : «Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries», *The Computation and Language E-Print Archive*, sur Internet.
RASTIER, F. (1987) : *Langages : sémantique et intelligence artificielle*, Bernard Willerval Jouve, 14192 édition.
RESNIK, Ph. (1995) : «Using Information Content to Evaluate Semantic Similarity in a Taxonomy», *IJCAI, 95*.
SALTON, G. (1988) : «Term-weighting Approaches in Automatic Text Retrieval».
SALTON, G. et M. J. MCGILL (1983) : *Introduction to Modern Information Retrieval*, McGraw-Hill Computer Science Series, New York, McGraw-Hill.
SCHMITT, L., OLIVAN, E., LANDI, B., ROYAUTÉ, J. et J. DUCLOY (1992) : «STDI : une station de travail pour une indexation assistée», *Natural Language Processing and its Applications*, Avignon.
SOWA, J. F. (1984) : *Conceptual Structures*, Addison Wesley.
YAROESKY, D. (1992) : «Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora», *COLING' 92*, Nantes, 23-28 août 1992.

ANNEXE

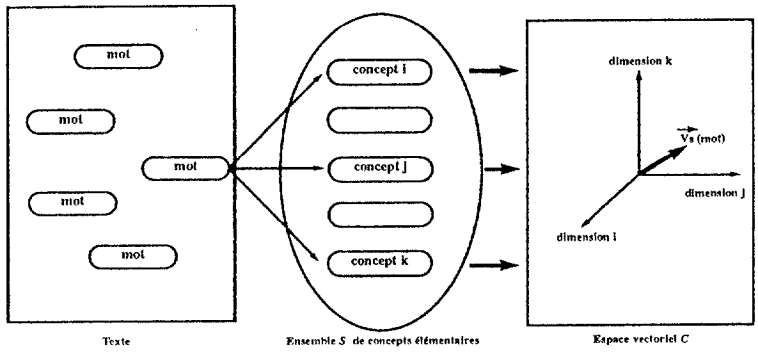

Figure :
Représentation graphique de l'application VS

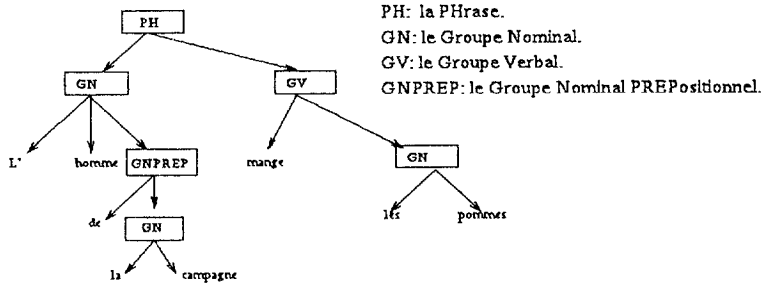

Figure 2 :
Structure syntaxique de la phrase *L'Homme de la compagne mange les pommes*

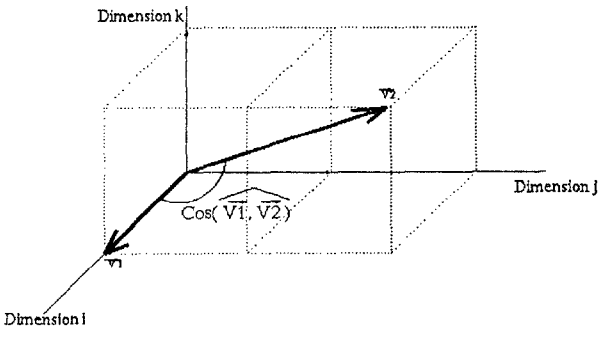

Figure 3 :
Représentation spatiale du cosinus