Le Machine Learning : numérique non supervisé et symbolique peu supervisé, une chance pour l'analyse sémantique automatique des langues peu dotées

Hammou Fadili[1,2]

[1]Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France

[2](Pôle Systèmes d'Information et du Numérique, Programme Maghreb) de la FMSH Paris
190, avenue de France 75013, Paris, France

Hammou.fadili@cnam.fr / fadili@msh-paris.fr

Résumé : Les données non structurées dominent l'univers de la production et de la publication des données, et en représentent, d'après plusieurs études, plus de 80%. Ce type de contenus, constitue la partie riche et précieuse en termes de données, d'informations et de connaissances ; donc nécessaire à intégrer et à prendre en considération dans les processus d'analyses et d'exploitations des données. L'analyse des données non structurées reste une discipline difficile, car elle repose sur de nombreux (pré-)traitements numériques (formalisation, normalisation, corpus, annotations, etc.) de la langue naturelle, faisant souvent défaut, surtout dans le cas des langues peu dotées. Dans cet article nous présentons une approche, bien adaptée aux cas où on ne dispose pas ou que de peu de données traitées. Elle est basée sur des méthodes d'apprentissages numériques non supervisés indépendants de la langue et symboliques peu supervisés, permettant d'exploiter directement des données brutes ou seulement des petites quantités de données traitées, comme base d'apprentissage pour l'interprétation des données. En l'appliquant à un cas concret d'une langue peu dotée, nous avons pu, montrer l'utilité et surtout l'opportunité que ces technologies pourraient constituer pour contourner les problèmes dont souffre ce type de langues, facilitant ainsi leur accès dans le monde de l'analyse sémantique automatique des données non structurées. Cette étude a été validée à travers des expérimentations confirmant de bons résultats pour l'approche.

Mots clés : Machine Learning, apprentissages non/peu supervisés, contextes, relations sémantiques, modèles des données, ressources langagières, LSA, LDA, langues peu dotées.

1. Introduction

D'une manière générale, l'analyse et l'interprétation sémantiques des données textuelles, n'est pas encore totalement prise en charge par les systèmes actuels. Ceci est dû aux problèmes et aux difficultés liées à la modélisation et à la formalisation de la complexité de la langue naturelle dans sa globalité, à ses très nombreux cas possibles, à ses nombreuses exceptions, etc. Ces difficultés sont encore plus accentuées, dans le cas des langues peu dotées, faute de ressources informatiques et de corpus traités suffisants.

Or, dans le domaine de l'analyse sémantique des données, il existe plusieurs technologies indépendantes de la langue traitée, ou ayant des capacités de généralisation à partir de petites bases d'apprentissage. Elles sont généralement basées, sur des notions d'apprentissages numériques non supervisés et/ou symboliques peu supervisés, ne nécessitent pas / ou peu de (pré)traitements préalables. Ces méthodes sont donc applicables à n'importe quelle langue, constituant ainsi une chance pour l'analyse sémantique automatique des langues peu dotées.

Afin de valider ces approches, nous avons testé des technologies, déjà étudiées pour le Français et l'Anglais, sur une langue peu dotée. Ceci afin de montrer, d'une part, leur indépendance par rapport aux langues et d'autre part, la possibilité de leur application à des langues peu dotées. Pour cela, nous avons exploité des dictionnaires existants pour enrichir et instancier une première partie du modèle de

données proposé pour l'apprentissage. L'autre partie a été complétée en exploitant des technologies basées sur des apprentissages non supervisés : LSA (Latente Semantic Analysis) pour la génération de la sémantique latente et LDA (Latent Dirichlet Allocation) pour la génération des thèmes du contexte général. D'autres outils classiques pour gérer le POS, effectuer le streaming, générer le contexte local, etc., ont été également exploités.

Dans cet article, la première partie sera consacrée à rappeler les motivations de ce travail, la deuxième partie sera consacrée à la description de notre contribution : la définition du modèle d'apprentissage, la préparation des données, présentation des technologies exploitées et/ou améliorées, description de l'architecture du système, ainsi que son évaluation. La dernière partie conclura l'article.

2. Motivation

Dans ce paragraphe, nous allons rappeler quelques éléments sur l'analyse sémantique des données non structurées d'une manière générale. Ceci, afin de montrer les difficultés qu'on rencontre dans le cas des langues peu dotées et les solutions qui peuvent être adoptées.

2.1. Analyse des données textuelles

Le processus d'analyse des données est généralement composé de deux phases :

- Phase de préparation des données
- Phase de l'analyse sémantique

Phase de préparation des données (filtrage des données)

Cette phase correspond aux prétraitements linguistiques des données textuelles, nécessaires, aux traitements des couches supérieures dont celles du traitement automatique de la sémantique. A ce niveau, le processus consiste en général, en l'analyse et en la génération d'un réseau de mots et de relations, qu'on appelle un réseau morphosyntaxique, dépourvu de sens, représentant seulement les éléments constituant du texte. La génération d'un tel réseau de mot suit en « général » le processus suivant :

- Segmentation
    - Tokénisation (tokenization) : découpage du texte en unités lexicales élémentaires (tokens)
    - Segmentation (Sentence Segmentation) : découpage du texte en phrases
- Analyse morphologique
    - Racinisation (stemming): recherche de la racine
    - Lemmatisation (lemmatization) : regroupement des formes de mots qui appartiennent à la même famille
- Analyse syntaxique
    - Etiquetage grammatical (Part-of-speech tagging : POS) : détermine la catégorie lexicale (nom, verbe, adj, adv, etc.), annotation des catégories grammaticales
    - Analyse de groupes grammaticaux (Chunking) : détermine les groupes grammaticaux (nominaux, verbaux, etc).

Phase d'analyse sémantique

Cette phase exploite les résultats de la phase précédente et des traitements sophistiqués pour la déduction de la sémantique dans le contexte. Ceci complète le processus général de l'analyse textuelle :

- Analyse sémantique
    - Exploitation des étapes précédentes du processus et des modèles symboliques ontologiques et/ou mathématiques statistiques pour l'analyse du sens.

Cette phase sera décrite d'une manière détaillée, à travers notre approche, un peu plus tard dans cet article. Les résultats d'analyses des deux phases sont en général stockés sous formats structurés comme en xml, csv, etc., échangeables et exploitables par des applications tierces.

2.2. Cas des langues peu dotées

D'une manière générale, les langues peu dotées, sont des langues qui souffrent de plusieurs problèmes liés à leur traitement automatique : problèmes liés à la graphie, au manque d'un système d'écriture stable, au manque de ressources informatiques et linguistiques. Le manque de ressources langagières concerne les dictionnaires, thésaurus, corpus traités, etc. ; le manque d'outils numériques concerne les outils du traitement automatique de la langue naturelle : analyseurs morphologiques, syntaxiques, sémantique, etc. Tous ces éléments rendent difficile, voire impossible l'analyse sémantique des langues peu dotées.

Afin de contourner ces problèmes, nous proposons l'exploitation, de solutions, indépendantes / peu dépendantes de la langue traitée et ayant fait leur preuve pour d'autres langues mieux dotées (Anglais, Français, etc.).

3. Notre approche

Notre approche permet de contourner les problèmes rencontrés pour les langues peu dotées. Elle est basée sur des apprentissages des modèles de représentation pour le traitement automatique de la sémantique du texte. D'une manière générale, il y a les approches dites numériques qui exploitent les fréquences des mots et les modèles mathématiques statistiques et probabilistes ; puis il y a les approches dites symboliques qui exploitent la structure des données, leur description et des systèmes de règles. Dans le premier cas, un texte est représenté dans un espace vectorielle, dont les dimensions sont des termes sélectionnés à partir d'un vocabulaire donné (un sous-ensemble des mots du texte), par un vecteur V des fréquences des mots du texte. Dans le deuxième cas, le texte est représenté par les mots le constituant ainsi que des propriétés et des règles issues des différents traitements linguistiques ou des formalisations spécifiques telles les ontologies. D'une manière générale, les modèles symboliques exploitant les données d'experts sont efficaces, mais sont difficiles à mettre en place, sauf pour des domaines très restreints. La difficulté réside dans le fait qu'on ne peut pas modéliser et décrire d'une manière complète la totalité d'un domaine donné. Cela demande beaucoup de ressources et de connaissances, donc beaucoup de travail. Les méthodes statistiques quant à elles n'ont pas besoin de toutes ces informations pour spécifier et décrire un modèle d'analyse ; elles exploitent seulement les données d'analyse et des modèles mathématiques. Dans notre cas, nous avons combiné les deux méthodes : on exploite des techniques d'apprentissage non supervisé pour préparer une partie ou la totalité des données qu'on combine avec un petit échantillon de données sur lequel on applique des techniques d'apprentissage peu-supervisé.

3.1. Contexte de l'étude

Cette étude s'inscrit dans le contexte général, de l'analyse sémantique des données textuelle des langues peu dotées. Elle est basée sur des apprentissages non/peu supervisés, et a été exploitée pour la détection de la sémantique dans la langue Amazigh. Pour se faire, on a juste traduit un mini corpus du Français sur lequel on a appliqué les différents algorithmes. Dans ce qui suit, une description des éléments du processus, exploités et/ou améliorés pour les besoins de l'étude.

3.2. Dictionnaires

En terme de dictionnaires, le manque de corpus annotées, d'une part, et le soucis de rendre notre démarche générique (sans se limiter à un domaine particulier) , d'autre part, nous ont obligé à mettre en place une stratégie basée sur l'exploitation d'un dictionnaire de langue et des apprentissages automatiques. On a exploité des méthodes d'apprentissages mixtes (peu supervisés et non supervisés) et des technologies de prétraitements (préparation) des données, pour instancier le modèle étendu de données et construire les bases de validation, d'entrainement et de test.

3.3. Contexte sémantique global

La notion de contexte sémantique dont on parle ici, concerne tous les paramètres pouvant influencer le sens des mots dans le texte. Ce contexte peut être considéré comme une agrégation de paramètres appartenant aux ensembles suivants :

- contexte textuel : le contexte du texte, domaine, thème, univers du discours,…
- contexte local : l'ensemble des mots situés de part et autre du mot, on l'appelle aussi la fenêtre textuelle
- contexte linguistique : le rôle du mot dans la phrase
- contexte d'utilisation : le domaine d'utilisation, but d'utilisation…
- contexte de l'auteur : les sens particuliers des mots suivant certains auteurs
- contexte du lecteur (point de vue d'analyse) : la compréhension particulière du sens de certains mots par certains lecteurs
- tous autres paramètres pouvant influencer le sens des mots.

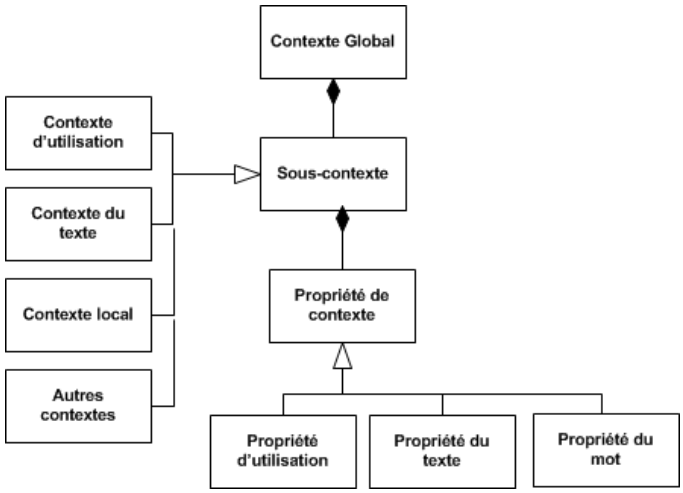

FIG. 1 - Méta-modèle du contexte globale

Ces éléments définissant le contexte sémantique global, sont au moins de trois types :

- Ceux contrôlés par l'utilisateur : domaine, but…
- Ceux véhiculés par le texte : thèmes, contexte textuel, local, linguistique…
- Ceux pour lesquels on peut faire des choix (cibler une lecture particulière du texte), i.e. qu'on peut fixer au départ ou extraire du texte lui-même : lecteurs, auteurs…

Des ontologies peuvent être utilisées pour modéliser et représenter des éléments des contextes. La partie du contexte connue à l'avance, doit être instanciée par l'utilisateur dès le départ ; l'autre partie dépendante du texte, doit être calculée et générée automatiquement au moment de l'analyse (contexte textuel, local, etc.).

3.4. Relations sémantiques

On considère ici toutes les relations influençant le sens que peuvent entretenir deux mots. On peut citer :

- les relations paradigmatiques : les relations de sens que peuvent entretenir deux mots
- les relations syntagmatiques : les relations de collocation modifiant le sens des mots
- les relations thématiques : les relations que peuvent entretenir des mots par rapport à une thématique
- les relations de variation : relations entre un mot et ses variations
- les relations racinisation : relation entre un mot et sa racine
- les relations sigles : relations d'équivalences entre les entités et les sigles, les initiales, et autres annotations…
- toutes les autres relations utiles pour la détection du sens des mots

En s'inspirant d'un travail de Igor Mel'cuk (Mel'cuk, 1997), nous avons recensé une soixantaine de relations sémantiques (Fadili, 2013).

Syn, Anti, Conv$_{ij}$, Contr, Epit, Gener, Figur, S$_0$, A$_0$, V$_0$, Adv$_0$, S$_i$, S$_{instr}$, S$_{med}$, S$_{mod}$, S$_{loc}$, S$_{res}$, Able$_i$, Qual$_i$, Sing, Mult, Cap, Equip, A$_i$, Adv$_i$, Imper, Result, Centr, Magn, Plus, Minus, Ver, Bon, Pos$_i$, Loc$_{in}$, Loc$_{ab}$, Loc$_{ad}$, Instr, Propt, Copul, Pred, Oper$_i$, Func$_i$, Labor$_{ij}$, Incep, Cont, Fin, Caus, Perm, Liqu, Real$_i$, Fact$_i$, Labreal$_{ij}$, Involv, Manif, Prox, Prepar$_i$, Degrad, Son, Obstr$_i$, Stop$_i$, Excess$_i$, Sympt$_{ijk}$.

FIG 2 - relations sémantiques

Dans le cas des relations sémantiques, on a besoin de dictionnaires modélisant les différents liens de sens entre les mots. On peut se baser sur des dictionnaires existants pour représenter une partie des relations. Ces dictionnaires souvent incomplets, doivent êtres enrichis automatiquement en utilisant des technologies basées sur des statistiques et des règles d'apprentissage ; afin de détecter de nouvelles relations, nécessaires à l'amélioration de l'interprétation automatique.

3.5. Modèle de données pour l'apprentissage

Pour l'analyse du texte, nous avons besoin d'une représentation la plus complète possible d'un mot et d'un texte, qui permet de capter toutes les informations discriminantes et de déduire le sens exact du mot dans le texte. Dans le présent travail, nous proposons un modèle exprimant essentiellement deux notions essentielles, la notion de contexte et la notion des relations sémantiques. Nous supposons, que la définition la plus large possible du contexte et la détermination de tous les mots liés d'un mot permettent de déterminer d'une manière non ambiguë son sens réel.

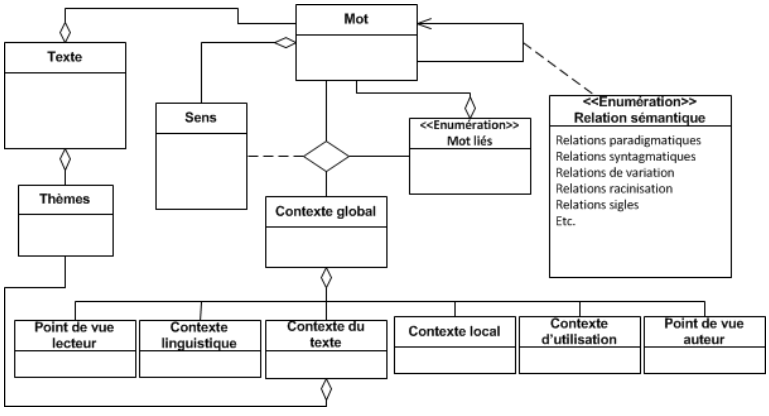

FIG. 3 - Modèle de données pour l'apprentissage

3.6. Sémantique latente pour l'instanciation d'une partie du modèle d'apprentissage

Dans ce paragraphe, nous allons présenter des approches statistiques pour l'extraction de la sémantique latente. Ces approches ont été exploitées pour compléter l'instanciation du modèle des données étendu et du contexte globale.

LSA : Analyse sémantique latente (Latente Semantic Analysis, LSA) (Deerwester et al. 1990) consiste à découvrir de manière statistique la sémantique latente dans un corpus. Elle permet d'établir des relations entre les termes (termes en plus de leur contexte) en exploitant une matrice de cooccurrences. Grâce à LSA, on peut exprimer par exemple une relation de synonymie entre deux mots qui apparaissent souvent ensemble et une relation de polysémie si un mot apparait souvent dans plusieurs contextes différents. L'obtention de la sémantique latente s'obtient par la construction de la matrice approchée X_k de la matrice des cooccurrences X (apparition des mots du texte dans les différents contextes locaux) qu'on décompose en une matrice diagonale de valeurs singulières et en deux matrices orthogonales $X = U\sum V^T$, puis qu'on réduit au nombre des valeurs singulières k non nulles $X_k = U_k\sum_k U_k^T$.

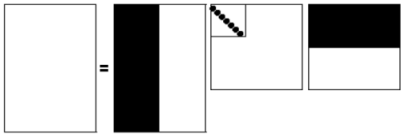

FIG. 4 - Réduction des dimensions dans LSA

$\boldsymbol{X_k}$ ne contient que la sémantique des mots la plus importante d'un point de vue statistique pour le document (la sémantique latente des mots dans le corpus). Cette matrice peut être utilisée pour calculer la distance entre les termes ou entre les contextes locaux, en utilisant la distance du cosinus entre les vecteurs lignes et colonnes respectivement. L'intersection d'une ligne et d'une colonne détermine la pertinence du terme associé à la ligne par rapport au contexte associé à la colonne.

LDA : Allocation latente de Dirichlet (Latent Dirichlet Allocation, LDA) (Blei & Lafferty, 2009) consiste à découvrir les thèmes latents d'un corpus correspondant à une distribution spécifique de mots fréquemment groupés. LDA fait partie d'une catégorie de modèles appelés "topic models", qui cherchent à découvrir des structures thématiques cachées dans les textes se présentant comme un mélange de thèmes. La répartition des mots par rapport aux thèmes se fait de la manière suivante : après avoir instancié le nombre de thèmes et une répartition aléatoire de tous les mots sur tous les thèmes, on ajuste l'appartenance des mots aux thèmes, en calculant les probabilités *p(theme|document)* * *p(mot|theme)* pour tous les thèmes et pout tous les mots. Dans notre cas, LDA a été utilisée pour associer un contexte à un document à partir des mots qui le composent. Elle a été également utilisée pour établir des liens entre les mots et les thèmes. Ces deux notions ont été exploitées pour augmenter la caractérisation du modèle des données pour l'apprentissage (décrit un peu plus haut dans ce document).

Ces 2 méthodes sont non-supervisées et ont souvent besoin d'être employées avec une décomposition en valeur singulière, pour réduire leur dimensionnalité, éviter les matrices creuses, et conserver un maximum de la significativité.

3.7. Apprentissages du sens

L'approche d'apprentissages définie et exploitée pour l'interprétation et la déduction du sens des mots dans le texte, est basée sur un réseau profond, permettant de modéliser avec un haut niveau d'abstraction des modèles de données articulés autour de transformations non linéaires. On l'a conçu pour permettre des apprentissages mixtes : les premières couches ou couches basses utilisent des

apprentissages non supervisés et les dernières couches des apprentissages supervisées ou peu supervisées (couches de décision). En d'autres termes, les premières couches ont été consacrées à l'extraction des caractéristiques à partir des informations latentes (apprentissage non supervisé) et les couches suivantes à la prise de décision (apprentissage peu supervisé).

3.8. Architecture

Le processus consiste en deux étapes principales :

- préparation des données,
- application des algorithmes d'apprentissage.

La première étape est elle-même composée en deux phases :

- La première phase consiste à générer les entrées suivant le modèle des données. Le but est de créer un vecteur par mot, enrichi par des propriétés discriminantes du sens. Pour cela, on exploite les technologies du POS, tokenisation, stemming et spliting pour générer la catégorie grammaticale et le contexte local, puis les technologies de gestion de la sémantique latente LSA et LDA pour générer le contexte du texte (thèmes) et certaines relations sémantiques latentes entre les mots.

- La deuxième phase consiste à exploiter un dictionnaire de langue pour extraire tous les Synset du mot, et enrichir le vecteur généré dans la première phase par, la description du Synset (gloss=définition+exemple) et tous les mots liés par les relations sémantiques. Au bout de cette deuxième phase, on obtient un vecteur globalement contextualisé et enrichi sur lequel on peut appliquer les algorithmes de classification et déduire le sens réelle, comme sortie du système.

FIG. 5 – Architecture

3.9. Evaluation

Nous avons exploité les fonctionnalités des outils Automap et Weka pour générer un fichier .arff (format exigé par Weka) d'instances suivant le modèle de données et pour implémenter un perceptron multicouche pour la détection du sens. Afin de rendre le système « honnête : les données de validation différentes de celles de test», nous avons séparé les données générées en trois parties comme suivant :

- 30% des données pour la validation, ceci afin d'optimiser les hyper-paramètres du système : le pas d'apprentissage, le type de la fonction d'activation et le nombre de couches.
- 60% pour l'entrainement, ceci afin d'estimer les meilleurs coefficients *(w_i)* de la fonction du réseau de neurones, minimisant l'erreur entre les sorties réelles et les sorties désirées.
- et 30% pour les tests, ceci afin d'évaluer les performances du système.

La génération du fichier .arff, s'est fait grâce, dans un premier, à la plateforme Automap, qui nous a permis d'instancier une partie du modèle (POS, Contexte local, etc.) via un fichier .csv. Puis dans un deuxième temps, grâce à la plateforme Weka, qui nous a permis d'appliquer sur le texte les technologies non supervisées (LSA et LDA) pour générer l'autre partie du modèle des données. La plateforme Weka a ensuite été utilisée pour l'implémentation d'un réseau de neurones de type « perceptron multicouche » appliqué aux instances pour l'évaluation.

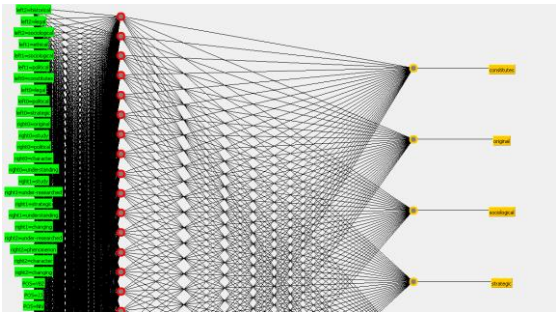

FIG. 6 - Application du perceptron multicouche

Ci-après quelques résultats des tests :

Correctly Classified Instances	71.4286%
Incorrectly Classified Instances	28.5714%

TAB. 2 – Vue d'ensemble

	Precision	Recall	F-Measure	ROC Area	Class
	0	0	0	0.917	iyya
	1	1	1	1	anezli
	1	1	1	1	tasnametti
	1	1	1	1	astrategic
	0.5	1	0.667	0.917	tisfrass
	0.5	1	0.667	0.917	yann
	0	0	0	0.917	azamul
Weighted Avg.	0.571	0.714	0.619	0.952	

TAB. 3 – Mesures

a	b	c	d	e	f	g	<-- classified as
0	0	0	0	0	1	0	| a=iyya
0	1	0	0	0	0	0	| b= anezli
0	0	1	0	0	0	0	| c= tasnametti
0	0	0	1	0	0	0	| d=astrategic
0	0	0	0	1	0	0	| e=tisfras
0	0	0	0	0	1	0	| f=yann
0	0	0	0	1	0	0	| g= azamul

TAB. 4 – Matrice de confusion

Les résultats des tests montrent les bonnes performances de l'approche. Le système à réussit grâce à l'apprentissage sur une partie des instances à déduire le sens réel de tous les mots qui étaient mal classés au départ.

4. Conclusion

Cet article propose une approche d'apprentissage indépendante de la langue traitée, pour instancier le modèle de données d'apprentissage et lui appliquer des méthodes d'apprentissages peu supervisés. Elle a l'avantage de contourner les problèmes majeurs rencontrés dans l'analyse des données non structurées dans le contexte des langues peu dotées, à savoir le manque d'outils et de données annotées, sémantiquement et automatiquement exploitables. Les expériences menées sur des exemples confirment d'une part l'apport considérable du modèle proposé pour la détection du sens réel des mots dans le texte et d'autre part, l'application de l'approche sur les langues peu dotées.

Références

Akoka J. et al., 2014. A Semantic Approach for Semi-Automatic Detection of Sensitive Data. *Information Resources Management Journal*, vol. 27, n° 4, pp. 23-44.

Blei D. et Lafferty J., 2009. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining, *Classification, Clustering, and Applications* . Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Cybenko, G., 1989. Approximation by Superpositions of a Sigmoidal Function. *Math. Control Signals Systems*, 2, 303-314.

Deerwester, S. et al., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407.

Fadili H., 2013. Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexico-semantic relations and the concept of the context., *IEEE-AICCSA, Ifrane (Morroco)*, May.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward nets. *Neural Networks.* 4, 231–242.

Rumelhart, D., Hinton G., Williams R., 1986. Learning internal representation by error propagation. In Parallel Distributed Processing, *Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press ,Vol. 1, pp. 318–362.

Turney D., 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.