Initialisation de Réseaux de Neurones à l'aide d'un Espace Thématique

Mohamed Morchid, Richard Dufour, Georges Linarès
Laboratoire Informatique d'Avignon
prénom.nom@univ-avignon.fr

Résumé. Ce papier présente une méthode de traitement de documents parlés intégrant une représentation fondée sur un espace thématique dans un réseau de neurones artificiels (ANN) employé comme classifieur de document. La méthode proposée consiste à configurer la topologie d'un ANN ainsi que d'initialiser les connexions de celui-ci à l'aide des espaces thématiques appris précédemment. Il est attendu que l'initialisation fondée sur les probabilités thématiques permette d'optimiser le processus d'optimisation des poids du réseau ainsi qu'à accélérer la phase d'apprentissage tout en amélioration la précision de la classification d'un document de test. Cette méthode est évaluée lors d'une tâche de catégorisation de dialogues parlés entre des utilisateurs et des agents du service d'appels de la Régie Autonome Des Transports Parisiens (RATP). Les résultats montrent l'intérêt de la méthode proposée d'initialisation d'un réseau, avec un gain observé de plus de 4 points en termes de bonne classification comparativement à l'initialisation aléatoire. De plus, les expérimentations soulignent que les performances sont faiblement dépendantes de la topologie du ANN lorsque les poids de la couche cachée sont initialisés au moyen des espaces de thèmes issus d'une allocation latente de Dirichlet ou *latent Dirichlet Allocation* (LDA) en comparaison à une initialisation empirique.

Abstract.
Neural Network Initialization using a Topic Space
This paper presents a method for speech analytics that integrates topic-space based representation into an artificial neural network (ANN), working as a document classifier. The proposed method consists in configuring the ANN's topology and in initializing the weights according to a previously estimated topic-space. Setup based on thematic priors is expected to improve the efficiency of the ANN's weight optimization process, while speeding-up the training process and improving the classification accuracy. This method is evaluated on a spoken dialogue categorization task which is composed of customer-agent dialogues from the call-centre of Paris Public Transportation Company. Results show the interest of the proposed setup method, with a gain of more than 4 points in terms of classification accuracy, compared to the baseline. Moreover, experiments highlight that performance is weakly dependent to ANN's topology with the LDA-based configuration, in comparison to classical empirical setup.

Mots-clés : Réseau de neurones artificiels, Allocation latente de Dirichlet, Initialisation de poids.

Keywords: Artificial neural network, Latent Dirichlet allocation, Weights initialization.

1 Introduction

Plusieurs méthodes d'analyse de documents parlés projettent leurs transcriptions automatiques dans un espace de thèmes [1] obtenu par le biais d'un apprentissage non-supervisé sur des corpus de documents de grande taille. L'objectif de cette projection est d'abstraire la représentation de surface des termes composant le document transcrit car ceux-ci peuvent rendre l'analyse directe des documents difficile (erreurs de transcription, disfluences...). Le module d'analyse de documents opère alors dans ces espaces thématiques lors de tâches de classification ou d'identification. Souvent, la représentation du contenu du document et le module d'analyse sont traités ou optimisés indépendamment : l'espace de représentation est conçu pour être le plus expressif et le plus compact possible, alors que le module d'analyse est optimisé par la fonction objective de la tâche finale. Dans ce travail, nous proposons une approche holistique où les espaces de thèmes et le système d'analyse de contenu sont optimisés conjointement. Les réseaux de neurones artificiels (ANN) sont maintenant une approche standard pour le traitement de la langue mais nécessitent un processus lourd et coûteux pour l'estimation des paramètres lors de la phase d'apprentissage. Cette difficulté d'élaboration de l'architecture du ANN est essentiellement

1. Il sera par la suite appelé "thème" un ensemble de termes regroupés dans une même classe dans l'espace LDA.

due au fait que la phase d'apprentissage est un processus d'optimisation stochastique dépendant de plusieurs facteurs comme la distribution des termes dans le corpus d'apprentissage ou les conditions d'initialisation du réseau. Le choix de la configuration initiale du réseau est un point crucial lors de la phrase d'apprentissage (poids des couches cachées, topologie...). Ce choix peut considérablement affecter le temps d'apprentissage (Adhikari & Joshi, 1956) ainsi que les performances du ANN (optimum local). Plusieurs études adaptent le momentum et le taux d'apprentissage pour accélérer l'apprentissage (Beale, 1972; Møller, 1993; Powell, 1977; Nguyen & Widrow, 1990; Drago & Ridella, 1992; Thimm & Fiesler, 1997) ou se concentrent sur l'initialisation des poids ou du biais en appliquant un pré-traitement fondé sur l'analyse de données ou des méthodes de classification (Breukelen & Duin, 1998; Kathirvalavakumar & Subavathi, 2011; Dahl *et al.*, 2012).

Dans ce papier, nous proposons une méthode d'initialisation d'un ANN évaluée lors d'une tâche d'identification de catégories de conversations transcrites automatiquement et issues de la RATP (Bechet *et al.*, 2012) potentiellement bruitées. Afin de gérer cette difficulté, un espace de thèmes permettant une abstraction des transcriptions en sortie du système de reconnaissance automatique de la parole (SRAP) est utilisé. Dans un schéma classique, la classification devrait s'opérer dans ces espaces thématiques. Ici, nous étudions l'impact de notre méthode d'initialisation d'un réseau de neurones (ANN) s'appuyant sur un espace de thèmes issu d'une allocation latente de Dirichlet (LDA). Dans un premier temps, nous comparons différentes entrées du ANN en utilisant la fréquence des termes de la conversation puis les probabilités issues d'espaces de thèmes. Nous proposons ensuite d'évaluer différentes initialisations des poids de la couche cachée d'un ANN : une initialisation aléatoire suivant une loi uniforme classique et notre initialisation originale au moyen des probabilités estimées avec une LDA.

La partie 2 présente les études précédentes liées à la représentation de documents ainsi qu'aux méthodes d'initialisation des ANN. Les concepts de base d'un ANN ainsi que les caractéristiques thématiques sont décrits dans la partie 3. La partie 4 présente les expériences ainsi que les résultats avant de conclure dans la partie 5.

2 Travaux antérieurs

L'approche classique à base de fréquences de mots TF-IDF (Robertson, 2004), a été très largement utilisée afin d'extraire les mots discriminants contenus dans des textes. D'autres approches ont proposé de considérer le document comme un mélange de thèmes cachés telles que *Latent Semantic Analysis* (LSA) (Deerwester *et al.*, 1990) ou encore *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003), permettant un niveau de représentation plus élevé. Les performances de ces méthodes ont pu être démontrées sur de nombreuses tâches. En particulier, dans l'approche LDA, un thème est associé à chacune des occurrences d'un terme contenu dans le document (et non un thème pour l'ensemble du document). Il en résulte que les thèmes appartenant à un document peuvent changer d'un terme à l'autre. Dans ce papier, les probabilités des thèmes cachés, estimées avec une LDA, capturent les dépendances possibles entre les termes pour permettre de modéliser le contenu sémantique d'une conversation donnée. Les réseaux de neurones (ANN) constituent un environnement standard aujourd'hui pour des tâches de classification ou de prédiction. L'un des modèles les plus populaires est le *feed-forward multilayer perceptron*, habituellement entraîné à l'aide de l'algorithme de rétro-propagation du gradient ou une de ses nombreuses variantes (Cazorla & Escolano, 2003; Hagan *et al.*, 1996). La rétro-propagation est une technique d'optimisation de descente du gradient offrant des propriétés de convergence rapide mais qui est fortement dépendante des conditions d'initialisation souvent choisies empiriquement. Ce problème est d'actualité et est abordé par plusieurs chercheurs. (Feuring, 1996) propose d'employer l'algorithme de rétro-propagation pour calculer les bornes de l'intervalle des valeurs potentiellement optimales pour les poids d'un ANN pour une tâche donnée (Draghici, 2002). Ensuite, le ANN doit résoudre ce problème avec des poids dont la valeur est un entier dans cet intervalle. Généralement, la plupart des méthodes proposées reposent sur l'analyse de données, des méthodes d'apprentissage automatique ou sur des connaissances *a priori* (Kathirvalavakumar & Subavathi, 2011; Dahl *et al.*, 2012).

3 Approche proposée

Un réseau de neurones (ANN) ou *feed-forward neural network* est composé de trois couches comme présenté dans la figure 1 : une couche d'entrée (x), une ou plusieurs couche(s) cachée(s) (θ) et une couche de sortie (y). Un ANN contient une couche cachée totalement connectée aux couches d'entrée et de sortie dans ce papier.

La première expérimentation consiste à évaluer l'impact de différents jeux de caractéristiques d'entrée d'un ANN issus de la fréquence classique des termes et issus d'espaces thématiques (voir partie 3.2). Le nombre de neurones contenus dans la couche d'entrée (x) correspond au nombre de caractéristiques (*i.e.* nombre de termes ou nombre de thèmes LDA). La seconde expérimentation cherche à évaluer l'impact, dans une tâche de catégorisation, de l'initialisation des poids de la

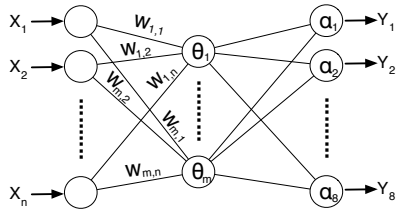

FIGURE 1 – Exemple d'architecture d'un ANN.

couche cachée soit aléatoirement, soit au moyen de probabilités estimées durant une LDA. Dans ce schéma, les neurones de la couche d'entrée représentent le vocabulaire et chacun des neurones de la couche cachée représente un thème LDA, les entrées de la couche cachée étant initialisées avec les probabilités de ces thèmes sachant les termes représentés par les neurones de la couche d'entrée. Ensuite, l'algorithme d'apprentissage reposant sur la rétro-propagation du gradient est réalisé. Cette dernière étape peut être vue comme un optimisation conjointe de la représentation de la couche thématique et de discrimination de la catégorie (*i.e.* classe) à associer à la conversation.

3.1 Concepts de base d'un réseau de neurones artificiels (ANN)

3.1.1 Fonction d'activation

La fonction d'activation utilisée durant les expérimentations est la fonction classique de *tangente hyperbolique* :

$$\alpha(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

Plus d'informations concernant les fonctions de transfert en général sont disponibles dans (Duch & Jankowski, 1999).

3.1.2 Algorithme d'apprentissage du *feed-forward*

Trois étapes sont nécessaires : calcul de la sortie, rétro-propagation de l'erreur et mise-à-jour des poids et biais.

Phase de calcul des sorties
Soit N_l le nombre de neurones contenus dans la couche l ($1 \leq l \leq M$) et M le nombre de couches du ANN. $\theta_{n,l}$ est le biais du neurone n ($1 \leq n \leq N_l$) de la couche l. Soit un ensemble de p exemples d'entrée x_i ($1 \leq i \leq p$) et un ensemble de classes y_i associées à chacun des x_i. La sortie $\gamma_{n,l}$ du neurone n de la couche l (voir figure 1) est donnée par :

$$\gamma_{n,l} = \alpha_{n,l} = \alpha(\sum_{m=0}^{N_{l-1}} w_{nm}^{l} \times \gamma_{m,l-1}) + \theta_{n,l} \tag{2}$$

Phase d'apprentissage
L'erreur e observée entre la sortie attendue y et le résultat de la phase de calcul des sorties γ est évaluée comme suit :

$$e_n^l = y_n - \gamma_{n,M} \tag{3}$$

$$e_n^l = \sum_{m=1}^{N_{l+1}} w_{m,n} \times \delta_{m,l+1} \tag{4}$$

pour respectivement la couche de sortie M (3) et les couches cachées (4). Le gradient δ est calculé ainsi : $\delta_{n,l} = e_n^l \times \alpha_{n,l}$

Phase de mise-à-jour
Lorsque les erreurs entre la sortie attendue et le résultat sont calculées, les poids $w_{n,m}^l$ et les biais $\theta_{n,l}$ doivent alors être respectivement mis-à-jour comme $w_{n,m}^{l^\star}$ et $\theta_{n,l}^\star$:

$$w_{n,m}^{l^\star} = w_{n,m}^l + \epsilon\delta_{n,l} \times \alpha_{n,l} \tag{5}$$

$$\theta_{n,l}^\star = \theta_{n,l} + \epsilon\delta_{n,l}\ . \tag{6}$$

3.2 Caractéristiques issues du document pour l'ANN

La méthode proposée d'initialisation du ANN est évaluée lors d'une tâche de catégorisation de conversations issues du corpus du projet DECODA (Bechet *et al.*, 2012). Un ANN a besoin d'un ensemble de caractéristiques en entrée x_i et de classes (*i.e.* catégories) y_i associées à un dialogue en sortie. Deux représentations différentes du document fondées respectivement sur la fréquence classique des termes discriminants contenus dans le document, et une représentation plus abstraite issue d'un espace de thèmes LDA, sont présentées dans les parties suivantes.

Fréquence de termes discriminants
Un ensemble de mots discriminants $\mathbf{V}$ de taille 166 est composé avec le critère de TF-IDF-Gini. Pour chacun des dialogues d, un ensemble de caractéristiques x^d est déterminé. La $k^{\text{ème}}$ caractéristique x^d_k est composée du nombre d'occurrences du mot t_k ($|t_k|$) dans d et le score Δ de t_n dans la liste de termes discriminants $\mathbf{V}$: $x^d_k = |t_k| \times \Delta(t_k)$

Espace de thèmes LDA
L'échantillonnage de Gibbs, présenté dans (Griffiths & Steyvers, 2004), est utilisé pour estimer les paramètres LDA et pour représenter un nouveau document dans l'espace des thèmes r de taille T. Ce modèle extrait un ensemble de caractéristiques de d depuis l'espace de représentation en thèmes. La $k^{\text{ème}}$ caractéristique est composée comme suit : $x^d_k = \theta^r_{k,d}$, où $\theta^r_{k,d} = P(z^r_k|d)$ est la probabilité du thème z^r_k (1 $\leq$k$\leq$ T) soit généré par le dialogue d dans l'espace de thème r^{th}.

4 Expériences

4.1 Protocole expérimental

Les expériences sur l'identification du thème d'une conversation sont menées sur le corpus du projet DECODA (Bechet *et al.*, 2012). Ce corpus est composé de 1 242 conversations téléphoniques (environ 74 heures de signal) découpées en un corpus d'apprentissage (740 dialogues), un corpus de développement (175 dialogues) et un corpus de test (327 dialogues). Ces dialogues ont été manuellement annotés selon 8 thèmes : *problème d'itinéraire*, *objet perdu et trouvé*, *horaire*, *carte de transport*, *état du traffic*, *prix du ticket*, *infraction* et *offre spéciale*. Les expérimentations conduites utilisent des transcriptions automatiques issues d'un système de reconnaissance automatique de la parole (SRAP) décrit dans (Morchid *et al.*, 2014) et obtenues avec le système Speeral (Linarès *et al.*, 2007). Enfin, le processus de validation croisée (apprentissage sur le corpus d'entraînement et validation à chacune des itérations avec l'ensemble de développement) est employé pour trouver la meilleure configuration (*i.e.* nombre d'itérations).

4.2 Résultats

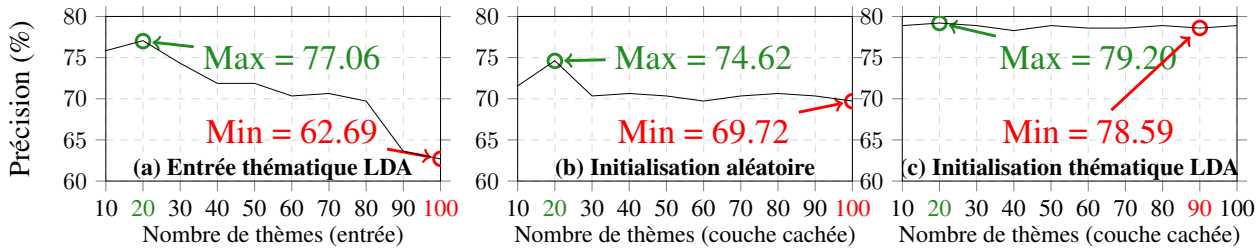

FIGURE 2 – Précision de la classification (%) en faisant varier le nombre de thèmes LDA en entrée du ANN (a), avec la couche cachée initialisée aléatoirement (b) et avec les poids de l'espace thématique (c).

Les premières expérimentations comparent deux ensembles de caractéristiques d'un document utilisant la représentation classique en fréquence de termes et utilisant un espace de thèmes LDA (voir parties suivantes). Ces représentations sont utilisées comme les entrées du ANN permettant l'apprentissage de celui-ci avec des documents textuels. Le ANN considéré est composé de trois couches : entrée (x issue de fréquences de termes ou de l'espace de thèmes LDA), cachée (1 couche de 8 neurones) et sortie (nombre de thèmes du corpus DECODA$=$ 8). Les poids du ANN sont initialisés aléatoirement durant ces expérimentations initiales. Dans un second temps, nous comparons l'initialisation classique des poids w de la couche cachée à l'aide d'une variable aléatoire et notre méthode issue d'espaces de thèmes LDA. Les réseaux de neurones sont appris en utilisant, dans ce cas, la représentation thématique des conversations téléphoniques.

Entrée	# Neurones	#n	Précision
Fréquence de termes	8	X	75,84 %
LDA	20	8	77,06 %
Fréqu. de termes + initialisation LDA	20	20	**79,20 %**

TABLE 2 – Meilleures précisions lors de l'identification de thèmes.

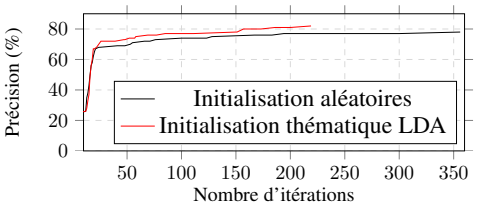

FIGURE 3 – Précision lors de la validation croisée (%) avec les deux approches d'initialisation de la couche cachée.

Comparaison des caractéristiques d'entrée du ANN

La première expérimentation utilisant les caractéristiques fondées sur la fréquence des termes permet d'obtenir une précision lors de la tâche de classification de 75.84 % (8 neurones dans la couche cachée). La seconde expérimentation utilise comme ensemble d'entrée les probabilités issues d'espaces de thèmes LDA. Sachant que la configuration des modèles LDA peut agir sur les performances de classification (Morchid *et al.*, 2014), nous proposons d'évaluer la performance du ANN en faisant varier le nombre de thèmes dans l'espace abstrait. La figure 2-(a) présente la précision obtenue avec différentes configurations de l'espace de thèmes (10 à 100 thèmes) au moyen de l'algorithme LDA, toujours avec 8 neurones dans la couche cachée. La première remarque est que la meilleure précision obtenue est de 77,06 % avec un gain possible de 1,22 points comparativement à la représentation fondée sur la fréquence de termes. Cependant, les résultats obtenus avec les probabilités issues de modèles LDA comme entrée sont instables et dépendent du nombre de thèmes dans l'espace (difference de $77,06 - 62,69 \simeq 15$ points). La partie suivante tire un avantage des deux représentations, en utilisant les caractéristiques fondées sur la fréquence des termes (plus stable) comme entrée du ANN tout en initialisant le ANN à l'aide des probabilités issues de LDA.

Initialisation des poids

La partie précédente a comparé un ensemble de caractéristiques en entrée du ANN en considérant que les poids de la couche cachée sont initialisés aléatoirement. L'objectif des expérimentations suivantes est de résoudre le choix difficile des poids initiaux du ANN en utilisant comme entrée la représentation en fréquence de termes tout en initialisant les poids à l'aide des probabilités thématiques issues de LDA. La méthode originale d'initialisation est comparée à une initialisation aléatoire des poids. Pour ce faire, deux ANNs sont construits avec la même architecture : couche d'entrée composée de x_n neurones avec la fréquence de termes discriminants t_n (vocabulaire $\mathbf{V} = 166$ mots discriminants, *i.e.* 166 neurones), couche cachée composée de $|T|$ neurones ($|T|$ =nombre de classes contenues dans l'espace de thèmes LDA $10 \leq |T| \leq 100$) et couche de sortie contenant 8 neurones (8 catégories dans le corpus DECODA). L'initialisation des poids de la couche cachée, fondée sur les thèmes LDA, consiste à considérer chacun des neurones de la couche cachée comme un thème LDA z_m. Ainsi, les poids sont considérés comme les probabilités thématiques du terme discriminant t_n sachant le thème représenté par le neurone dans la couche cachée z_m :

$$w_{m,n} = P(t_n|z_m) \tag{7}$$

La figure 2-(b) montre la précision obtenue avec une initialisation aléatoire alors que la figure 2-(c) présente les précisions obtenues avec une initialisation fondée sur les espaces de thèmes. En comparant ces deux courbes, nous constatons clairement que les résultats obtenus en initialisant des poids à l'aide de l'espace de thèmes LDA, sont meilleurs que ceux obtenus en initialisant les poids aléatoires, quel que soit le nombre de neurones contenus dans la couche cachée. En effet, la meilleure précision est de 74,62 %, obtenue pour une initialisation aléatoire, alors que l'initialisation à l'aide des probabilités thématiques LDA atteint une précision maximum de 79,20 % (gain de 4.58 points). Finalement, la méthode proposée (entrée=TF-IDF-Gini et poids=LDA) permet d'améliorer les performances du ANN en utilisant des caractéristiques d'entrée fondées sur les espaces de thèmes (entrée=LDA et poids=aléatoire) avec un gain de $79,20 - 77,06 = 2,14\%$ comme le montre le tableau 2. Les résultats présentés dans la figure 2-(b) sont également plus consistants (la différence entre la valeur minimum et maximum atteinte en termes de précision atteint 0,6 point) en comparaison avec la robustesse du réseau de neurones initialisé aléatoirement présenté dans la figure 2-(a) (différence de 4,9 points). Cette approche permet donc d'atteindre de meilleurs résultats qu'une initialisation classique aléatoire, mais plus important, élimine le choix difficile du nombre de neurones dans la couche cachée (résultats équivalents lorsque le nombre de neurones dans la couche cachée varie). La figure 3 présente la précision lors de la phase de validation croisée (ensemble de développement) avec une initialisation aléatoire et une initialisation fondée sur les espaces de thèmes LDA. Nous pouvons aisément constater que l'initialisation des poids à l'aide des probabilités thématiques permet d'obtenir une meilleure précision (78 % et 82 % pour respectivement une initialisation aléatoire et fondée sur l'espace LDA) avec un nombre d'itérations plus faible (356 et 219 itérations pour respectivement une initialisation aléatoire et fondée sur l'espace LDA). Un gain de 137 itérations

est alors observé, ce qui correspond à un gain en termes de temps de traitement (apprentissage du ANN) de 38.5% .

5 Conclusion

Ce papier présente une configuration originale de poids initiaux d'un réseau de neurones artificiels (ANN) au moyen des probabilités issues d'un espace thématique LDA. Les expérimentations ont montré l'intérêt de l'utilisation de variables latentes (probabilités thématiques) pour initialiser les poids du réseau de neurones durant une tâche de classification. LDA fournit ainsi une représentation robuste et pertinente de contenus bruités durant la phase d'apprentissage, optimisée selon la fonction objective liée à la tâche de classification. Cette méthode obtient de meilleurs résultats qu'un schéma classique fondé sur une représentation thématique du document à l'aide de LDA suivie par une classification à l'aide d'un ANN. Le gain est d'environ 4 points en termes de précision alors que le temps d'apprentissage est considérablement réduit (ce gain est d'environ 38%). Nous envisageons, dans des travaux futurs, d'évaluer cette approche en utilisant des réseaux de neurones profonds ainsi que des espaces de représentation hiérarchiques.

Références

ADHIKARI B. & JOSHI D. (1956). Distance discrimination et resume exhaustif. *Publ. Inst. Statist. Univ. Paris*, **5**, 57–74.

BEALE E. (1972). A derivation of conjugate gradients. *Numerical methods for nonlinear optimization*, p. 39–43.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC'12*.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.

BREUKELEN M. V. & DUIN R. P. W. (1998). Neural network initialization by combined classifiers. In *Proceedings of the 14th International Conference on Pattern Recognition*, ICPR'98, p. 215–.

CAZORLA M. A. & ESCOLANO F. (2003). Two bayesian methods for junction classification. *Image Processing, IEEE Transactions on*, **12**(3), 317–327.

DAHL G. E., YU D., DENG L. & ACERO A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, **20**(1).

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, **41**(6), 391–407.

DRAGHICI S. (2002). On the capabilities of neural networks using limited precision weights. *Neural networks*, **15**(3).

DRAGO G. P. & RIDELLA S. (1992). Statistically controlled activation weight initialization (scawi). *Neural Networks, IEEE Transactions on*, **3**(4), 627–631.

DUCH W. & JANKOWSKI N. (1999). Survey of neural transfer functions. *Neural Computing Surveys*, **2**(1), 163–212.

FEURING T. (1996). Learning in fuzzy neural networks. In *Neural Networks, 1996., IEEE International Conference on*, volume 2, p. 1061–1066 : IEEE.

GRIFFITHS T. L. & STEYVERS M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, **101**(Suppl 1), 5228–5235.

HAGAN M. T., DEMUTH H. B., BEALE M. H. *et al.* (1996). *Neural network design*, volume 1. Pws Boston.

KATHIRVALAVAKUMAR H. & SUBAVATHI S. J. (2011). A new weight initialization method using cauchy's inequality based on sensitivity analysis. *Journal of Intelligent Learning Systems and Applications*, **3**(1), 242–248.

LINARÈS G., NOCÉRA P., MASSONIE D. & MATROUF D. (2007). The lia speech recognition system : from 10xrt to 1xrt. In *Text, Speech and Dialogue*, p. 302–308 : Springer.

MØLLER M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, **6**(4).

MORCHID M., DUFOUR R., BOUSQUET P.-M., BOUALLEGUE M., LINARÈS G. & DE MORI R. (2014). Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *ICASSP*.

NGUYEN D. & WIDROW B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, p. 21–26 : IEEE.

POWELL M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical programming*, **12**(1).

ROBERTSON S. (2004). Understanding inverse document frequency : on theoretical arguments for idf. *Journal of Documentation*, **60**(5), 503–520.

THIMM G. & FIESLER E. (1997). High-order and multilayer perceptron initialization. *Neural Networks, IEEE Transactions on*, **8**(2), 349–359.