TOPIC-SPACE BASED SETUP OF A NEURAL NETWORK FOR THEME IDENTIFICATION OF HIGHLY IMPERFECT TRANSCRIPTIONS

Mohamed Morchid, Richard Dufour, Georges Linarès

LIA - University of Avignon (France)

{mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr

ABSTRACT

This paper presents a method for speech analytics that integrates topic-space based representation into a feed-forward artificial neural network (FFANN), working as a document classifier. The proposed method consists in configuring the FFANN'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 FFANN'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 FFANN's topology with the LDA-based configuration, in comparison to classical empirical setup.

***Index Terms*—** Artificial neural network, Latent Dirichlet allocation, Weights initialization, Hidden layer

1. INTRODUCTION

Numerous speech analytics methods rely on the mapping of automatic transcriptions into topic spaces obtained by unsupervised analysis of large text corpus, such as latent semantic analysis (LSA) [1] or latent Dirichlet allocation (LDA) [2]. This mapping aims at abstracting word-level representations, that may be impacted by disfluent speech and transcription errors. Speech analytics module operates on these topic spaces, typically by applying classification or identification methods. Most of the time, content representation and analysis modules are independently optimized: representation spaces are designed to be as expressive and compact as possible, since analysis module is optimized according to the final-task objective function. In this work, we propose a holistic approach where topic-space and analysis-system are jointly optimized. This method relies on a LDA-based topic models and a feed-forward artificial neural network (FFANN) whose hidden layer is configured according to a pre-estimated LDA-based topic-space.

Neural networks are now a standard approach for signal and speech processing, but efficiency is usually obtained by heavy tuning and learning processes. This difficulty in designing and training efficient FFANN architecture is due to the fact that training is a stochastic optimization process that is highly dependent to many factors such as training data distribution or initial conditions. A crucial point of the learning process is the choice of the initial configuration (including weight and topology), which may dramatically affect the training time [3] as well as the FFANN performance. Many previous works related to training speed-up consisted in adapting the momentum and the learning rate [4, 5, 6, 7, 8, 9]. Some of them focused on weights and biases initialization, typically by applying pre-processing based on data analysis or fast clustering methods [10, 11, 12].

In this paper, we propose a FFANN setup method that is evaluated in the context of a documents classification task that involves automatic transcriptions of telephone conversations from the RATP customer care service (Paris Public Transportation Authority) [13]. Telephone conversations are a particular case of human-human interaction whose automatic processing raises problems. In particular, the speech recognition step required to obtain the transcription of the speech contents may have poor performance, due to unexpected speaker's behavior and large training/test mismatch. Globally, the speech signal may be strongly impacted by various sources of variability: environment and channel noises, acquisition devices, etc.

The theme identification system should deal with problems related both to recognition errors and to class proximity. To deal with these problems, dialogues are mapped into a topic space abstracting the ASR outputs. In a classical scheme, the classification would operate in these topic spaces. Here, we investigate the impact of our LDA-based FFANN setup method. Firstly, different features as input neurons are compared using classical term-frequency and topic-based features. In addition, we propose to evaluate different FFANN initialization weights using a classical ini-

This work was funded by the ContNomina project supported by the French National Research Agency contract ANR-12-BS02-0009.

tialization with small uniform random values, and using our original initialization with thematic-based priors.

The rest of the paper is organized as follows. Section 2 presents previous works related to word representation and FFANN initialization. The basic concepts of FFANN and the thematic features are described in Section 3. Sections 4 and 5 report experiments and results, before concluding in Section 6.

2. RELATED WORK

Dialogue classification is a particular case of text categorization. Many approaches considered the document as a mixture of latent topics such as latent semantic analysis (LSA) [14] or latent Dirichlet allocation (LDA) [2]. Topic-based methods have demonstrated their performance on various tasks, such as sentence [15] or keyword [16] extraction.

In particular, LDA considers that a theme is associated to each occurrence of a word composing the document, rather than associate a topic to the complete document. As a result, topics of a document may change from a word to another one. However, word occurrences are connected by a latent variable which controls the global match of the distribution of the topics in the document. These latent topics are characterized by a distribution of associated word probabilities. LDA models generally outperform LSA on Information Retrieval tasks [17]. In this paper, probabilities of hidden topic features, estimated with LDA, are considered for possibly capturing word dependencies expressing the semantic contents of a given conversation.

Neural networks constitute a classical framework for classification or prediction tasks. One of the most popular model is the feed-forward multilayer perceptron, usually trained by the backpropagation algorithm or one of its numerous variations [18, 19, 20, 21]. Backpropagation is a gradient-descend optimization technique that offers fast convergence properties but which is also highly dependent to the initial conditions, mainly empirically chosen. This issue was addressed by many authors in the past. [22] proposed an algorithm in which the backpropagation process is employed to compute the weights bounds. Another work determined a range of weights for a given task [23]. Then, the network has to solve this problem with integer weights in that range as well as possible. Generally, most of the methods proposed relied on data analysis, machine-learning or a priori knowledge [24, 11, 12].

Our proposal is to setup initial FFANN weights and the hidden layer size according to a topic-space model estimated by LDA. In this scheme, each cell of the hidden layer represents a topic, input-hidden layer weights being initialized by thematic priors. Then, classical back-propagation training is achieved. This last step may be viewed as a joint optimization of thematic-layer representation and class discrimination.

3. PROPOSED APPROACH

A feed-forward neural network (FFANN) is composed of three different components (or layers) as presented in Figure 1: input layer (x), hidden layer(s) (θ) and output layer (y). A FFANN containing one hidden layer fully connected to input and output ones is used in this paper.

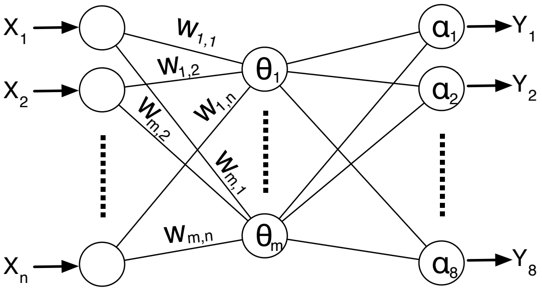

Fig. 1. Example of a FFANN architecture.

The first experiment consists in evaluating the impact of different features given as input of the FFANN: term-frequency features and topic-based features, both described in section 3.2. The number of neurons contained into the input layer (x) corresponds to the number of features (*i.e.* number of words or number of topics). The weights of the hidden layer (8 neurons=8 classes in the documents corpora) are initialized randomly. The second experiment seeks to evaluate the impact of the weight initialization, with a classical random initialization and a new one that uses the topic-based features. Finally, for both experiments, the number of neurons composing the output layer is equal to the number of themes related to the DECODA corpus (*i.e.* 8 in our experiments).

3.1. FFANN basic concepts

3.1.1. Activation function

The activation function used during the experiments is the classical *hyperbolic-tangent* function:

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

One can find out more information about transfer functions in [25].

3.1.2. Feed-forward learning process

The feed-forward algorithm is composed of 3 steps: forward, learning and update phases.
Forward phase
Let N_l be the number of neurons contained into the layer l

($1 \leq l \leq M$) and M the number of layers of the FFANN. $\theta_{n,l}$ is the bias of the neuron n ($1 \leq n \leq N_l$) from the layer l. Given a set of p input patterns x_i ($1 \leq i \leq p$) and a set of labels y_i associated to each x_i, as described in Figure 1, the output $\gamma_{n,l}$ of the neuron n from the layer l is given by:

$$\begin{aligned} \gamma_{n,l} &= \alpha\Big(\sum_{m=0}^{N_{l-1}} w^l_{nm} \times \gamma_{m,l-1}\Big) + \theta_{n,l} \\ &= \alpha_{n,l} \qquad (2) \end{aligned}$$

Learning phase

The error e observed between the expected outcome y and the result of the forward phase γ is then evaluated as follows:

$$e^l_n = y_n - \gamma_{n,M} \qquad (3)$$

for the output layer M, and

$$e^l_n = \sum_{m=1}^{N_{l+1}} w_{m,n} \times \delta_{m,l+1} \ , \qquad (4)$$

for the hidden layer. The gradient δ is computed with:

$$\delta_{n,l} = e^l_n \times \alpha_{n,l} \qquad (5)$$

Update phase

When errors between the expected outcome and the result are computed, the weights $w^l_{n,m}$ and the bias $\theta_{n,l}$ have to be respectively updated to $w^{l\star}_{n,m}$ and $\theta^{\star}_{n,l}$:

$$\begin{aligned} w^{l\star}_{n,m} &= w^l_{n,m} + \epsilon\delta_{n,l} \times \alpha_{n,l} \qquad (6) \\ \theta^{\star}_{n,l} &= \theta_{n,l} + \epsilon\delta_{n,l} \ . \qquad (7) \end{aligned}$$

3.2. FFANN input features from documents

The proposed FFANN setup method is evaluated in the theme identification task of conversations from the DECODA project [13]. A FFANN needs a set of features x_i as input and a class (*i.e.* theme) y_i associated to the given dialogue as output. Two different document representations based on respectively the classical term-frequency of discriminative words contained into the document, and a more abstract representation based on a LDA topic space, are presented in the next sections.

3.2.1. Term-frequency features using discriminative terms

A discriminative word subset $\mathbf{V}$ of size 166 is composed as described in [26] and based on TF-IDF-Gini criteria.

Note that a same word t can be present in different themes, but with different scores depending of its relevance in the theme.

For each dialogue d, a set of semantic features x^d is determined. The k^{th} feature x^d_k is composed with the number of occurrences of the word t_k ($|t_k|$) in d and the score Δ of t_n in the discriminative word set $\mathbf{V}$ defined as:

$$x^d_k = |t_k| \times \Delta(t_k) \qquad (8)$$

This set of features is used as the FFANN input. More details about discriminative term-frequency representation are in [26].

3.2.2. Topic-based features from a latent Dirichlet allocation

Several techniques, such as Variational Methods [2], Expectation-propagation [27] or Gibbs Sampling [28], have been proposed to estimate the parameters describing a LDA hidden space. The Gibbs Sampling, reported in [28] and detailed in [29], is used to estimate LDA parameters and to represent a new dialogue d with the r^{th} topic space of size T. This model extracts a features set of d from the topic-based representation. The k^{th} feature is computed as follows:

$$x^d_k = \theta^r_{k,d} \ , \qquad (9)$$

where $\theta^r_{k,d} = P(z^r_k|d)$ is the probability of topic z^r_k (1 ≤ k ≤ T) produced by the dialogue d in the r^{th} topic space of size T.

4. EXPERIMENTS

4.1. Dataset

The corpus is a set of human-human telephone conversations in the customer care service (CCS) of the RATP Paris transportation system. This corpus comes from the DECODA project [13] and is used to perform experiments on conversation theme identification. It is composed of 1,242 telephone conversations, which corresponds to about 74 hours of signal. The data set was split in 8 categories or themes as described in Table 1.

Table 1. DECODA dataset.

Class	*Number of samples*		
label	**training**	**development**	**testing**
problems of itinerary	145	44	67
lost and found	143	33	63
time schedules	47	7	18
transportation cards	106	24	47
state of the traffic	202	45	90
fares	19	9	11
infractions	47	4	18
special offers	31	9	13
Total	**740**	**175**	**327**

The LIA-Speeral Automatic Speech Recognition (ASR) system [30] is used to extract textual content of dialogues from the DECODA corpus. Acoustic model parameters were estimated from 150 hours of speech in telephone conditions. The vocabulary contains 5,782 words. A 3-gram language model (LM) was obtained by adapting a basic LM with the training set transcriptions. This system reaches an overall Word Error Rate (WER) of 45.8% on the training set, 59.3% on the development set, and 58.0% on the test set A "stop list" of 126 words[1] was used to remove unnecessary words (mainly function words) which results in a Word Error Rate (WER) of 33.8% on the training, 45.2% on the development, and 49.5% on the test. These high WER are mainly due to speech disfluencies and to adverse acoustic environments (for example, calls from noisy streets with mobile phones).

4.2. Experimental protocol

First experiments, described in Section 5.1, compare two sets of features for a document using classical term-frequency based features, described in Section 3.2.1, and topic-based features from a LDA approach, detailed in Section 3.2.2. These representations are used as input of the FFANN. This represents the classical way to train FFANN with document content features.

The weights w of the hidden layer have to be initialized (see Figure 1). In a second phase, experiments compare the classical random weights initialization to our proposed weights initialization based on thematic priors from a LDA model.

Experiments are conducted using the automatic transcriptions only (ASR). The FFANNs are learned and tested with the DECODA corpus (see Table 1). The cross validation (learning process with training corpus and validation in each iteration with the development set) is used to find out the best configuration point (*i.e.* numbers of iterations).

5. RESULTS

5.1. Comparison of FFANN input features

First experiments aim to define an efficient set of features, that describes the document content, as an input x of the FFANN. Section 3.2 details the two considered sets of features: a classical one based on term-frequency (see Section 3.2.1), and a more sophisticated and abstract one based on topic-based features (see Section 3.2.2).

As presented in Section 3.1, the FFANN considered is composed with three layers: input (x from a set of term-frequency or of topic-based features), hidden (1 layer with 8 neurons) and output (number of themes contained in the DECODA corpora$=8$) layers. FFANN's weights are randomly initialized during these initial experiments for both features set configurations.

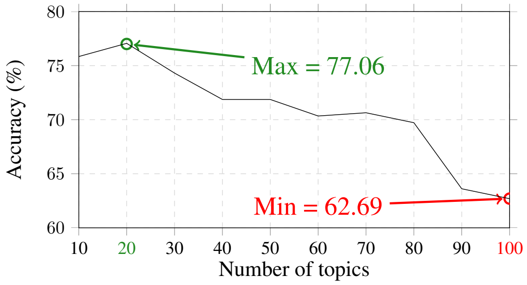

Fig. 2. Theme classification accuracies (%) using various topic-based representations on the DECODA test set.

The first experiment (term-frequency features) allowed us to obtain a theme classification accuracy of 75.84% (8 neurons in the hidden layer). The second experiment uses as an input the topic priors obtained from a LDA. Since the LDA model configuration may impact the classification performance [31], we propose to evaluate the FFANN performance by varying the number of topics. Figure 2 presents the accuracies obtained with different topic-based features configurations (10 to 100 topics) from LDA algorithm, still with 8 neurons in the hidden layer.

The first remark is that the best accuracy obtained is 77.06% with a possible gain, compared to classical term-frequency set of features, of 1.22 points. Nonetheless, results obtained with LDA priors as input are quite unstable, and vary greatly depending on the number of topics (difference of $77.06 - 62.69 \simeq 15$ points). The next section will then take advantage of both representations, by using term-frequency features as input, while initializing the FFANN weights with LDA topic priors.

5.2. Weights initialization

The previous section has compared, as input of the FFANN, a term-frequency and a topic-based features set, considering that hidden-layer weights are randomly initialized. The purpose of the next experiments is to work out the difficult choice of the FFANN initialization weights by using term-frequency features as input, while initializing the FFANN hidden-layer weights with LDA topic priors. This original initialization is compared to a random one. To do so, two FFANNs are built with the same architecture: input layer neurons x_n are the term-frequencies of discriminative terms t_n composed with $|\mathbf{V}|$ ($\mathbf{V} = 166$ discriminative words, *i.e.* 166 neurons), the hidden layer is composed with $|T|$ neurons ($|T| =$ number of classes contained into the LDA topic space $10 \leq |T| \leq 100$) [26] , while the output layer contains 8 neurons ( ).

The topic-based weights initialization consists in considering each neuron from the hidden layer as a LDA topic z_m. Then the weights are considered as the LDA topic priors be-

[1] http://code.google.com/p/stop-words/

tween the discriminative word t_n and the neuron in the hidden layer z_m:

$$w_{m,n} = P(t_n|z_m) \tag{10}$$

Two weights initialization approaches are then compared: Figure 3 shows accuracies obtained with weights randomly initialized while Figure 4 presents accuracies obtained with topic-based weights initialization.

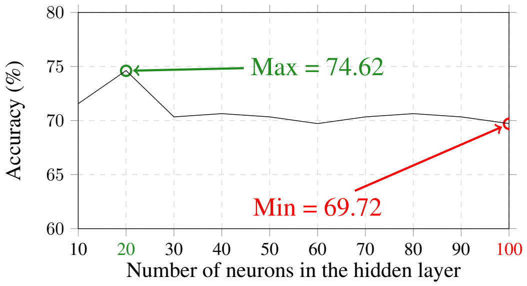

Fig. 3. Theme classification accuracies (%) on test set using various number of neurons randomly initialized (hidden layer).

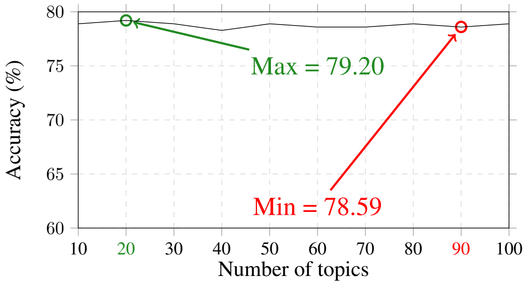

Fig. 4. Theme classification accuracies (%) on test set using various topic model sizes to initialize weights of FFANN hidden layer with thematic priors.

By comparing these two curves, one can clearly figure out that the results obtained using the weights initialized with the topic-based priors outperform those achieved with the weights randomly initialized, regardless the number of neurons considered in the hidden layer. Indeed, the best accuracy reaches 74.62% for the random initialization while the topic-based weights initialization achieves a maximum accuracy of 79.20% (gain $= 4.58$ points). Finally, the proposed approach allows us to improve the FFANN using topic-based features (LDA) as input with a gain of $79.20 - 77.06 = 2.14\%$ as shown in Table 2.

Results presented in Figure 4 are also more consistant (the difference between the minimum and maximum accuracies reaches 0.6 points) in comparison to the robustness of a neural network with weights randomly initialized presented in Figure 3 (difference $= 4.9$ points). This approach then allows us to achieve better results than classical random weights initialization, but more importantly, eliminates the difficult choice of the number of neurons in the hidden layer.

Table 2. Best theme classification accuracies with different input types (word frequency and LDA) and hidden layer initialization methods (random and LDA weights initialization).

Input	Init	# neurons	#*n*	Accuracy
Word freq.	random	8	X	75.84
LDA	random	20	20	77.06
Word freq.	LDA	20	20	**79.20**

Figure 5 presents the cross-validation accuracies obtained with random and topic-based initialization on the development set. One can easily point out that the topic-based weighting approach allows us to achieve better accuracies (78% and 82% for respectively the random and the topic-based weights initialization) with a lower number of iterations (356 and 219 iterations for respectively random and topic-based weights initialization). A gain of 137 iterations is then observed, which corresponds to a gain of 38.5% in terms of processing time.

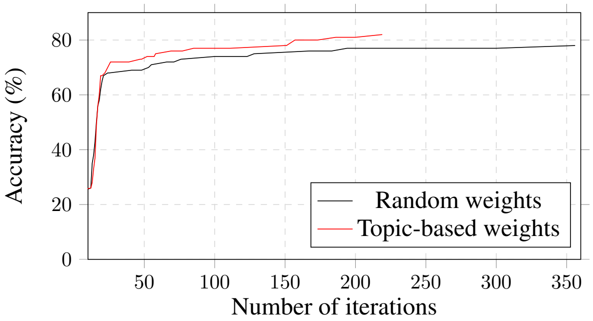

Fig. 5. Theme classification cross-validation accuracies on dev. set using random and topic-based weights initialization.

6. CONCLUSIONS

This paper presents an original setup of a feed-forward artificial neural network (FFANN) using LDA priors. Experiments have shown the interest of the use of latent variables (topic-based priors) to initialize the weights of a neural-network during a classification task: LDA provides relevant representation of noisy contents that are, during the training phase, optimized according to the task-related objective function. This method outperforms the standard sequential scheme based on a first LDA-based representation followed by a FFANN-based classification process. The gain is about 4 points in terms of accuracy, while the training time is considerably reduced (the observed speed gain is about 38% absolute). We plan now to evaluate this approach using deep neural networks and hierarchical topic spaces.

7. REFERENCES

[1] S. Dumais, “Latent semantic indexing (lsi) and trec-2,” *NIST SPECIAL PUBLICATION SP*, pp. 105–105, 1994.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] BP Adhikari and DD Joshi, “Distance discrimination et resume exhaustif,” *Publ. Inst. Statist. Univ. Paris*, vol. 5, pp. 57–74, 1956.

[4] EML Beale, “A derivation of conjugate gradients,” *Numerical methods for nonlinear optimization*, pp. 39–43, 1972.

[5] Martin Fodslette Møller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural networks*, vol. 6, no. 4, pp. 525–533, 1993.

[6] Michael James David Powell, “Restart procedures for the conjugate gradient method,” *Mathematical programming*, vol. 12, no. 1, pp. 241–254, 1977.

[7] Derrick Nguyen and Bernard Widrow, “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*. IEEE, 1990, pp. 21–26.

[8] Gian Paolo Drago and Sandro Ridella, “Statistically controlled activation weight initialization (scawi),” *Neural Networks, IEEE Transactions on*, vol. 3, no. 4, pp. 627–631, 1992.

[9] Georg Thimm and Emile Fiesler, “High-order and multilayer perceptron initialization,” *Neural Networks, IEEE Transactions on*, vol. 8, no. 2, pp. 349–359, 1997.

[10] Martijn Van Breukelen and Robert P. W. Duin, “Neural network initialization by combined classifiers,” in *Proceedings of the 14th International Conference on Pattern Recognition*, Washington, DC, USA, 1998, ICPR ’98, pp. 215–, IEEE Computer Society.

[11] HThangairulappan Kathirvalavakumar and Subramanian Jeyaseeli Subavathi, “A new weight initialization method using cauchys inequality based on sensitivity analysis,” *Journal of Intelligent Learning Systems and Applications*, vol. 3, no. 1, pp. 242–248, 2011.

[12] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 20, no. 1, 2012.

[13] Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot, “Decoda: a call-centre human-human spoken conversation corpus,” LREC’12, 2012.

[14] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[15] Jerome R. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[16] Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi, “Keyword extraction using term-domain interdependence for dictation of radio news,” in *17th international conference on Computational linguistics*. ACL, 1998, vol. 2, pp. 1272–1276.

[17] Thomas Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.

[18] Miguel A Cazorla and Francisco Escolano, “Two bayesian methods for junction classification,” *Image Processing, IEEE Transactions on*, vol. 12, no. 3, pp. 317–327, 2003.

[19] Martin T Hagan, Howard B Demuth, Mark H Beale, et al., *Neural network design*, vol. 1, Pws Boston, 1996.

[20] Patricia Melin, Claudia Gonzalez, Diana Bravo, Felma Gonzalez, and Gabriela Martinez, “Modular neural networks and fuzzy sugeno integral for face and fingerprint recognition,” in *Applied Soft Computing Technologies: The Challenge of Complexity*, pp. 603–618. Springer, 2006.

[21] VV Phansalkar and PS Sastry, “Analysis of the back-propagation algorithm with momentum,” *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 505–506, 1994.

[22] Thomas Feuring, “Learning in fuzzy neural networks,” in *Neural Networks, 1996., IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 1061–1066.

[23] Sorin Draghici, “On the capabilities of neural networks using limited precision weights,” *Neural networks*, vol. 15, no. 3, pp. 395–414, 2002.

[24] Zhe Chen, Tian-Jin Feng, and Zweitze Houkes, “Incorporating a priori knowledge into initialized weights for neural classifier,” in *IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000*, Como, Italy, 2000, vol. 2, pp. 291–296.

[25] Włodzisław Duch and Norbert Jankowski, "Survey of neural transfer functions," *Neural Computing Surveys*, vol. 2, no. 1, pp. 163–212, 1999.

[26] Mohamed Morchid, Georges Linarès, Marc El-Beze, and Renato De Mori, "Theme identification in telephone service conversations using quaternions of speech features," in *Interspeech*. ISCA, 2013.

[27] Thomas Minka and John Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[28] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[29] Gregor Heinrich, "Parameter estimation for text analysis," *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[30] Georges Linarès, Pascal Nocéra, Dominique Massonie, and Driss Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.

[31] Mohamed Morchid, Richard Dufour, Pierre-Michel Bousquet, Mohamed Bouallegue, Georges Linarès, and Renato De Mori, "Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule," in *ICASSP*. IEEE, 2014.