Subspace Gaussian Mixture Models for Dialogues Classification

Mohamed Bouallegue[†], *Mohamed Morchid*[†], *Richard Dufour*[†],
Driss Matrouf[†], *Georgès Linares*[†] *and Renato De Mori*[†‡]

[†]LIA, University of Avignon, France
[‡]McGill University, School of Computer Science, Montreal, Quebec, Canada
{firstname.lastname}@univ-avignon.fr, rdemori@cs.mcgill.ca

10.21437/Interspeech.2014-426

Abstract

The main objective of this paper is to identify themes from dialogues of telephone conversations in a real-life customer care service. In order to capture significant semantic content in spite of high expression variability, features are extracted in a large number of hidden spaces constructed with a Latent Dirichlet Allocation (LDA) approach. Multiple views of a spoke document can then be represented with several hidden topic models. Nonetheless, the model diversity due to the multi-model approach introduces a new type of variability. An approach is proposed based on features extracted in a common homogenous subspace with the purpose of reducing the multi-span representation variability. A *Gaussian Mixture Model subspace* model, inspired by previous work on speaker identification, is proposed for theme identification. This representation, novel for theme classification, is compared with the direct application of multiple topic-model representations. Experiments are reported using a corpus collected in the call center of the Paris Transportation Service. Results show the effectiveness of the proposed representation paradigm with a theme identification accuracy of 78.8%, showing a significant improvement with respect to previous results on the same corpus.

Index Terms: Human/Human conversation analysis, theme identification, LDA features, GMM subspace, Latent Dirichlet Allocation.

1. Introduction

This paper introduces a new approach to theme identification of conversations between an agent and a casual user in a call centre.

In these types of real-life applications, the customer behavior may exhibit significant variability in the way problems are expressed. Modeling this type of linguistic variability is one of the concerns addressed in the proposed approach.

Other sources of variability due to environment and channel noises, distortions introduced by acquisition devices, also have to be taken into account. These sources of variability impact the performance of Automatic Speech Recognition (ASR) systems resulting in high Word Error Rates (WER) in some conversations. The extraction of useful speech analytics may be adversely affected by these errors even if classification uncertainty can be alleviated by different types of redundancy in the expression of themes and problems. In order to improve theme classification accuracy, efforts can be made to improve ASR robustness and interpretation robustness to ASR errors. In this paper, we propose a new method to improve the robustness of a classification application by combining a semantic multi-model approach using a homogenous dialogue representation subspace.

The proposed approach is evaluated in the application framework of the RATP call centre (Paris Public Transportation Authority), focusing on the topic identification task [1]. Themes are related to the main customer request during a call. In this task, themes are, for example, *lost and founds, traffic state, schedules, costs*. Theme identification should also take into account semantic relations between themes. For example, a *lost & found* customer request can be related to an itinerary (*where the object was lost?*) or time (*when?*). These conversations involve a relatively small set of basic concepts related to transportation issues, and classification has to resolve the ambiguity arising from mentions of facts of different themes while only one of these themes has to be identified as the main concern of the customer call.

An efficient way to improve both ASR and interpretation robustness consists in capturing syntactic dependencies expressing relevant semantic content by representing dialogues in a hidden topic space. A popular unsupervised method for such a task is based on Latent Dirichlet Allocation (LDA) [2]. A major problem when using a LDA based approach is the selection of various meta-parameters such as the number of hidden topics (that determines the model granularity), the word distribution assumptions, the temporal spans, and others. If the decision process is highly dependent on these choices, the system performance could be quite unstable.

Our proposal is to estimate a large set of topic spaces by varying the number of topics of LDA models. The mapping of a document into each of the resulting spaces could be considered as a particular view of the spoken contents. In the topic identification context, this multiple representation of the same dialogue could improve the tolerance of the identification system to the recognition errors, to the class proximity and to the LDA meta-parameter dependency.

On the other hand, this multi-view diversity introduces an additional variability. We propose to reduce this variability by representing dialogues in a homogenous feature space. This process is performed in the vocabulary domain, assuming that the whole vocabulary space is modeled by a global Gaussian Mixture Model, called Universal Background Model (UBM). Each theme of dialogue is modeled by a GMM obtained by adapting the parameters of a UBM using the features of the dialogue belonging to this theme. The adaptation is performed in a super-vector space where the concatenations of the GMM means are represented.

This work was funded by the SUMACC and ContNomina projects supported by the French National Research Agency (ANR) under contracts ANR-10-CORD-007 and ANR-12-BS02-0009.

This approach, called Subspace Gaussian Mixture Model (SGMM), has already been proposed in [3] for HMM-state modeling for speech recognition. This approach has some similarities to Eigenvoices [4] and Cluster Adaptive Training [5], and some relationship to the Joint Factor Analysis approach used in speaker identification [5]. The SGMM proposed in this paper is conceived for reducing the variability of features obtained from LDA hidden spaces and used for theme identification in human/human conversations.

This paper is organized as follows. The dialogue representation is described in section 2. In section 3, the proposed approach to model the theme of the dialogue by using the SGMM is presented. Sections 4 and 5 report experimental results, while section 6 concludes this work.

2. Dialogue representation

The purpose of the considered application is the identification of the major theme of a human/human telephone conversation in the customer care service (CCS) of the RATP Paris transportation system. The approach considered in this paper focuses on modeling the variability between different dialogues expressing the same theme t. For this purpose, it is important to select features that represent semantic contents relevant for the theme of a dialogue. An attractive set of features for capturing possible semantically relevant word dependencies is obtained with Latent Dirichlet Allocation (LDA) [2]. Given a train set of conversations D, a hidden topic space is derived and a conversation d is represented by its probability in each topic of the hidden space. Estimation of these probabilities is affected by a variability inherent to the estimation of the model parameters. If many hidden spaces are considered and features are computed for each hidden space, it is possible to model the estimation variability together with the variability of the linguistic expression of a theme by different speakers in different real-life situations. Even if the purpose of the application is theme identification and a train corpus annotated with themes is available, supervised LDA [6] is not suitable for the proposed approach since LDA is used only for producing different feature sets used for computing statistical variability models.

In order to estimate the parameters of different hidden spaces, a vocabulary V or theme discriminative words is constructed as described in [7, 8, 9]. For each theme t, a set of 50 theme specific words is identified. The same word may appear in more than one theme vocabulary selection. All the selected words are then merged without repetition to form V made of 166 words.

Several techniques, such as Variational Methods [2], Expectation-propagation [10] or Gibbs Sampling [6], have been proposed for estimating the parameters describing an LDA hidden space. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [11] and gives a simple algorithm for approximate inference in high-dimensional models such as LDA [12]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as: $p(W|\overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{w \in W} p(\overrightarrow{w}|\overrightarrow{\alpha}, \overrightarrow{\beta})$ for the whole data collection W knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

The Gibbs Sampling for estimating LDA was firstly reported in [6]. A more comprehensive description of this method can be found in [12]. One can refer to these papers for a better understanding of this sampling technique. This method is used both to estimate the LDA parameters and to infer an unseen dialogue d with the n^{th} topic space of size q. The Gibbs sampling allows to obtain a feature vector $V_d^{z^n}$ of d where the k^{th} feature $V_d^{z_k^n} = P(z_k^n|d)$ (where $1 \leq k \leq q$) is the probability of topic z_k^n knowing the unseen dialogue d in the n^{th} topic space of size q.

A set of p topic spaces are learned using LDA by varying the number of topics q to obtain p topic spaces of size q. The number of topics q varies from 10 to $3,010$. Thus, a set of $3,000$ topic spaces are learned from a LDA.

The next process allows to obtain a homogeneous representation of the dialogue d for the n^{th} topic space. The feature vector $V_d^{z^n}$ of the dialogue d is mapped into the common vocabulary space V composed with a set of discriminative words [7, 8, 9] for each theme to obtain a new feature vector $V_d^{z^n}$ of size 166 for the n^{th} topic space of size q where the i^{th} $(0 \leq i \leq 166)$ feature is:

$$\begin{aligned} V_d^{w_i} &= P(w_i|d) \\ &= \sum_{j=1}^{q} P(w_i|z_j)P(z_j|d) \\ &= \sum_{j=1}^{q} V_{z_j^n}^{w_i} \times V_d^{z_j^n} \\ &= \left\langle \overrightarrow{V_{z^n}^{w_i}}, \overrightarrow{V_d^{z^n}} \right\rangle \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, w_i is one of the 166 discriminative word, $V_{z_j^n}^{w_i} = P(w_i|z_j)$ and $V_d^{z_j^n} = P(z_j|d)$ evaluated by the Gibbs Sampling [13].

3. Subspace Gaussian Mixture Models

The approach to theme identification in human/human conversations explores the possibility of making decisions using a statistical variability model. The proposed solution is inspired by the use of Subspace Gaussian Mixture Models (SGMM) proposed for speaker verification. For this task, the speaker variability and the session variability are modeled with mixtures of gaussians. For theme identification, the variability of the sentences, used for expressing a theme, is modeled in a similar way as the speaker variability, and the variability of different feature estimations for each theme is modeled in a similar way as the variability observed in different sessions in which the same person speaks.

In the model for theme identification, mixtures of gaussians of vectors are considered. The vector has 166 elements, one for each discriminative word in the vocabulary V. For a given conversation d, there are 3,000 instances of the vector. The elements of an instance are the posterior probabilities of a discriminative word given the document d. The probabilities are computed using the hidden topics of a specific LDA hidden space.

For the sake of clarity, the use of SGMM for speaker verification is briefly reviewed in the next subsection.

3.1. SGMM for speaker verification

In the context of a speaker verification system based on GMM-UBM, a SGMM paradigm was introduced in order to model the speaker characteristics and the session variability at the same time, but as two distinct components [14].

The global GMM-UBM is defined as follows: UBM=$(\alpha_g, m_g, \Sigma_g)$, where α_g, m_g and Σ_g are respectively the weight, the mean and the covariance matrix of the

g^{th} gaussian. The GMM-UBM represents all speakers. The specific speaker model is then derived from the UBM and the available training data. Only the GMM means are adapted. The other GMM parameters (variances and weights) are taken from the UBM without any modification. In SGMM, the vector obtained by concatenating all Gaussian means is named *super-vector*. Let D be the dimension of the feature space. The dimension of a super-vector mean is $M \times D$, where M is the number of gaussians in the GMM. A speaker model can be decomposed into three different components: a speaker and session independent component, a speaker dependent component, and a session dependent component. Let (h, s) indicates the session h of the speaker s. In the SGMM, the means super-vector random variable of the speaker is written as follows:

$$\mathbf{m}_{h,s} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{h,s} \quad (1)$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent super-vector mean, $\mathbf{D}$ is a $MD \times MD$ diagonal matrix, $\mathbf{y}_s$ is the speaker vector (a MD vector), $\mathbf{U}$ is the session variability matrix of low rank R (a $MD \times R$ matrix) and $\mathbf{x}_{(h,s)}$ are the channel factors (theoretically, $\mathbf{x}_{(h,s)}$ is independent to s). We assume that $\mathbf{y}_s$ and $\mathbf{x}_{(h,s)}$ are normally distributed among $N(0, I)$. D satisfies the equation $I = \tau D^t \Sigma^{-1} D$ where τ is the relevance factor required in the standard MAP adaptation. In the training phase, the $\mathbf{U}$ matrix, the $\mathbf{y}_s$ and $\mathbf{x}_{(h,s)}$ have to be estimated. The $\mathbf{U}$ matrix is estimated on a large number of speakers, each one having several sessions; the $\mathbf{y}_s$ component is estimated on all the sessions that belong to speaker s; the $\mathbf{x}_{(h,s)}$ component is estimated on session h. The algorithm that presents the adopted strategy to estimate different components of the equation 1 is detailed in [15].

Once the speaker variability and the session variability is modeled, the session component is ignored, and $\mathbf{m} + \mathbf{D}\mathbf{y}_s$ is kept as model for speaker s.

3.2. SGMM for dialogue classification

In this subsection, the application of the SGMM approach to automatic theme identification is described. SGMMs use LDA based features following evidence provided in [9] that using a single LDA space limits the negative impact of ASR errors.

Nonetheless, the projection of the dialogues in a topic space generates a variability due to the estimation of the LDA parameters. Their estimation with the Gibbs Sampling [16] takes into account all words contained in the vocabulary. Indeed, for an unseen dialogue d, the estimation of the probability that a topic z was generated by d adds a residual semantic variability due to the fact that $p(z|d)$ is estimated for all words in the vocabulary, and not only for the words contained in d. This variability may impact the dialogue classification performance.

In order to apply SGMMs to model the different themes of dialogues, the GMM-UBM is estimated using all dialogues of all themes in the train set. A theme corresponds to a speaker and a dialogue to a session (see section 3.1). Note that $m_{t,d}$ is the super-vector corresponding to the theme (t, d). The SGMM, in this new context, can be written as:

$$m_{(t,d)} = m + Dy_t + Ux_{(t,d)} \quad (2)$$

where $\mathbf{m}_{(t,d)}$ is the theme-dialogue dependent super-vector mean, $\mathbf{D}$ is a $MD \times MD$ diagonal matrix, $\mathbf{y}_t$ is the theme vector (a MD vector), $\mathbf{U}$ is the dialogue variability matrix of low rank R (a $MD \times R$ matrix) and $\mathbf{x}_{(t,d)}$ are the dialogue factors (theoretically $\mathbf{x}_{(t,d)}$ is independent to t). To obtain the theme model, the component of dialogue variability $Ux_{(t,d)}$ is ignored, leaving $\mathbf{m} + \mathbf{D}\mathbf{y}_t$ to model the variability that may adversely affect classification.

In order to obtain the final model for each theme, different weight sets for each theme are estimated using all the dialogues of each theme in the train set. Estimation is performed by re-estimation of the weights of the GMM-UBM with a simple iteration process using the EM (Expectation-Maximization) algorithm.

Let λ_g be the weight of the gaussian g in the GMM-UBM. The weight λ_g^t of this gaussian in the theme t is calculated as follows:

$$\lambda_g^t = \frac{\sum_{x \in t} P(g|x)}{N_t} \quad (3)$$

where x is a dialogue belonging to a specific theme and $P(g|x)$ is the feature probability computed with Gaussian g on features obtained from x as follows:

$$P(g|x) = \frac{\lambda_g * f(x|g)}{\sum_{g'} \lambda_g * f(x|g')} \quad (4)$$

N_t is the number of representative vectors of the theme t and $f(x|g)$ is the likelihood for the frame x given the gaussian g.

In our theme models, the variances remain unchanged with respect to the GMM-UBM.

4. Experimental Protocol

The experiments on theme identification are performed using the DECODA project corpus [1]. This corpus is composed of 1,514 telephone conversations, corresponding to about 74 hours of speech. It has been split into a train set (740 dialogues), a development set (447 dialogues) and a test set (327 dialogues), and manually annotated with 8 conversation themes: *problems of itinerary*, *lost and found*, *time schedules*, *transportation cards*, *state of the traffic*, *fares*, *infractions* and *special offers*.

A set of $n = 3000$ LDA hidden spaces was created with a variable number of hidden topics as described in section 2)using the LDA Mallet Java implementation[1]. For each of the 8 themes, the size of the discriminative word vocabulary was empirically set to 50 words with experiments on the development set. These sets were merged leading to a vocabulary of 166 different words.

The ASR system used for the experiments is the LIA-Speeral system [17]. Model parameters were estimated with maximum *a-posteriori* probability (MAP) adaptation from 150 hours of speech in telephone condition. The vocabulary contains 5,782 words. A 3-gram language model (LM) was obtained by adapting a basic LM with the train set transcriptions. A "stop list" of 126 words[2] was used to remove unnecessary words (mainly function words) which results in a WER of 33.8% on the train, of 45.2% on the development, and of 49.5% on the test. These high WER are mainly due to speech disfluencies and to adverse acoustic environments for some dialogues when, for example, users are calling from noisy streets with mobile phones.

[1] http://mallet.cs.umass.edu/

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

5. Results

Experiments were performed with the development set to choose the number of gaussians for the GMM-UBM and the rank of the matrix U modeling the variability among dialogues of all themes.

Gaussian mixtures for the GMM-UBM, with respectively 16, 32, 64 and 128 gaussians, and ranks of the U matrix (see equation 2) with sizes 20, 50, 100, 150, have been considered.

The results obtained with the development set are reported in Table 1.

U \ g	16	32	64	128
20	82.45	83.62	81.28	83.62
50	84.21	**85.96**	83.04	84.21
100	81.87	84.79	83.04	83.62
150	81.28	84.79	82.45	83.62

Table 1: *Classification performance of the SGMM method with different GMM-UBM and matrix U sizes on the development data.*

The highest classification accuracy, close to 86%, is obtained with a model of 32 gaussians and a matrix U of rank 50.

The fact that the best results are obtained with only 32 gaussians is probably due to the relatively small size of the train and development sets. The results also show that a rank 50 of a matrix U is sufficient to model the variability between dialogues.

The SGMM setting obtained with experiments on the development set was applied to the theme identification on the test data. The results are reported in Table 2. The best classification performance on the test set is obtained with 128 gaussians and a matrix U of rank 50. Nonetheless, a high performance (78.19%) is obtained with 32 gaussians and a matrix U of rank 50.

U \ g	16	32	64	128
20	74.14	76.01	75.70	76.32
50	74.45	78.19	76.32	79.12
100	75.38	**78.81**	76.01	77.57
150	75.70	77.57	76.01	77.57

Table 2: *Classification performance of the SGMM method with different GMM-UBM and matrix U sizes on the test data.*

For the sake of comparison, the following classical methods of gaussian mixture modeling were also evaluated: Expectation-Maximization (EM) algorithm [18] and Maximum a posteriori (MAP) adaptation [19].

The model parameters (means, weights and variances of the gaussian mixture of the themes) of each theme were estimated with the EM algorithm while for SGMMs a common variance for all theme models is used.

Table 3 shows the results obtained with the model estimated with the EM maximization algorithm on the development and test data. It appears that the classification performance obtained with the EM algorithm is lower than the one obtained with SGMMs. These results could be explained by the fact that the limited size of training data is not sufficient to correctly estimate GMM themes independently.

Dataset \ g	16	32	64	128
DEV	80.11	80.70	78.36	79.60
TEST	58.47	63.55	61.68	57.00

Table 3: *Classification performance of the Expectation-Maximization algorithm on the development and test data.*

In a second experiment, the performance obtained with the SGMM model was compared to performance obtained with MAP adaptation. Indeed, the GMM-UBM is adapted with this method using the data of each theme in order to obtain a GMM for each theme. The obtained results, detailed in Table 4, show that this method gives good performance compared to EM models, but it is still less efficient than the proposed SGMM approach. This can be explained by the fact that the MAP adaptation does not take into account the inter-variability between dialogues (that is modeled with the SGMM approach).

Dataset \ g	16	32	64	128
DEV	83.62	84.79	83.04	83.62
TEST	76.32	77.88	77.88	77.90

Table 4: *Classification performance of the MAP adaptation on the development and test data.*

Overall, the experiments show that classification with the SGMM approach outperforms classification obtained with other methods using gaussian mixture modeling in this theme classification task. The results obtained by the SGMM approach are also superior to the best classification results previously obtained on the DECODA corpus [7](SSQ method), with a gain of 4.9 points on the test data (same ASR condition).

6. Conclusions

The possibility of using a large number of different LDA hidden topic spaces has been considered for extracting features suitable for estimating parameters of SGMM models.

The novelty of this approach is that the SGMM representation is used for modeling semantic contents of human/human conversations with topic-based representation features. With this technique, conversation themes are modeled with mixture of gaussians. With the SGMM approach, the dialogue variability can be estimated in a subspace of low dimension, the variability component being ignored in the space model. This method allows to obtain a clean model for the themes, and significantly improves the classification results in terms of average accuracy among imperfect transcriptions. These encouraging results promote to further investigate the adaptation of speech processing techniques to natural language processing for problems such as document categorization or keyword extraction.

7. References

[1] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "Decoda: a call-centre human-human spoken conversation corpus." LREC'12, 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *ICASSP*, 2010, pp. 4330–4333.

[4] J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," vol. 8, 2000, pp. 695–707.

[5] M. J. F. Gales, "Multiple-cluster adaptive training schemes," in *IN PROC. ICASSP*, 2001.

[6] T. L. Griffiths and M. 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.

[7] M. Morchid, G. Linarès, M. El-Beze, and R. De Mori, "Theme identification in telephone service conversations using quaternions of speech features," in *INTERSPEECH*, 2013.

[8] M. Morchid, R. Dufour, P.-M. Bousquet, M. Bouallegue, G. Linarès, and R. De Mori, "Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule," in *ICASSP*, 2014.

[9] M. Morchid, R. Dufour, and G. Linarès, "A LDA-based topic classification approach from highly imperfect automatic transcriptions," in *LREC'14*, 2014.

[10] T. Minka and J. 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.

[11] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.

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

[13] M. Morchid, R. Dufour, and G. Linarès, "Thematic representation of short text messages with latent topics: Application in the twitter context," in *PACLING*, 2013.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, vol. 1, March 18-23, 2005, pp. 637–640.

[15] D. Matrouf, N. Scheffer, B. G. B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *INTERSPEECH*, 2007, pp. 1242–1245.

[16] J. G. Scott and J. Baldridge, "A recursive estimate for the predictive likelihood in a topic model."

[17] G. Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977. [Online]. Available: http://web.mit.edu/6.435/www/Dempster77.pdf

[19] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," vol. 2. IEEE Transactions on Speech and Audio Processing, 1994, pp. 291–298.