An I-vector Based Approach to Compact Multi-Granularity Topic Spaces Representation of Textual Documents

Mohamed Morchid[†], Mohamed Bouallegue[†], Richard Dufour[†], Georges Linarès[†], Driss Matrouf[†] 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

Abstract

Various studies highlighted that topic-based approaches give a powerful spoken content representation of documents. Nonetheless, these documents may contain more than one main theme, and their automatic transcription inevitably contains errors. In this study, we propose an original and promising framework based on a compact representation of a textual document, to solve issues related to topic space granularity. Firstly, various topic spaces are estimated with different numbers of classes from a Latent Dirichlet Allocation. Then, this multiple topic space representation is compacted into an elementary segment, called c-vector, originally developed in the context of speaker recognition. Experiments are conducted on the DECODA corpus of conversations. Results show the effectiveness of the proposed multi-view compact representation paradigm. Our identification system reaches an accuracy of 85%, with a significant gain of 9 points compared to the baseline (best single topic space configuration).

1 Introduction

Automatic Speech Recognition (ASR) systems frequently fail on noisy conditions and high Word Error Rates (WER) make the analysis of the automatic transcriptions difficult. Speech analytics suffer from these transcription issues that may be overcome by improving the ASR robustness and/or the tolerance of speech analytic systems to ASR errors. This paper proposes a new method to improve the robustness of speech analytics by combining a semantic multi-model approach and a noise reduction technique based on the i-vector paradigm.

This method is evaluated in the application framework of the RATP call centre (Paris Public Transportation Authority), focusing on the theme identification task (Bechet et al., 2012).

Telephone conversations are a particular case of human-human interaction whose automatic processing raises problems, especially due to the speech recognition step required to obtain the transcription of the speech contents. First, the speaker's behavior may be unexpected and the training/test mismatch may be very large. Second, the speech signal may be strongly impacted by various sources of variability: environment and channel noises, acquisition devices, etc.

Telephone conversation issues

Topics are related to the reason why the customer called. Various classes corresponding to the main customer's requests are considered (*lost and founds, traffic state, timelines*, etc). In addition to classical issues in such adverse conditions, the topic-identification system should deal with problems related to class proximity. For example, a *lost & found* request is related to itinerary (*where was the object lost?*) or timeline (*when?*), that could appear in most of the classes. In fact, these conversations involve a relatively small set of basic concepts related to transportation issues. Figure 1 shows an example of a dialogue which is manually labeled by the agent as an issue related to an *infraction*. However, words in bold suggest that this conversation could be related to a *transportation card*. Thus, we assume that a dialogue representation should be seen as a multi-view problem to substantiate the claims regarding the multi-theme representation of a given dialogue.

On the other hand, multi-view approaches introduce additional variability due to the diversity of the views. This variability is also due to the vocabulary used by both agent and customer

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 443–454,
October 25-29, 2014, Doha, Qatar.

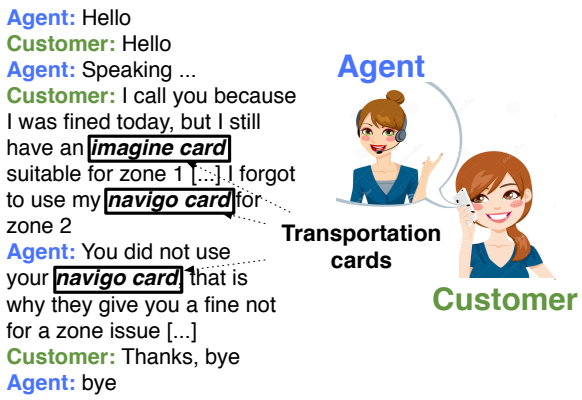


Figure 1: Example of a dialogue from the DECODA corpus labeled by the agent as an *infraction* issue which contains more than one theme (*infraction* + *transportation cards*).

during a telephone conversation. Indeed, an agent have to follow an predefined scenario of conversation. Thus, the agent can find the main reason for the call which corresponds to the theme.

Proposed solutions

An efficient way to tackle both ASR robustness and class ambiguity could be to map dialogues into a topic space abstracting the ASR outputs. Then, dialogue categorization is achieved in this topic space. Numerous unsupervised methods for topic-space estimation were proposed in the past. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been largely used for speech analytics; one of its main drawbacks is the tuning of the model, that involves various meta-parameters such as the number of classes (that determines the model granularity), word distribution methods, temporal spans... If the decision process is highly dependent on these features, the system's performance could be quite unstable.

Classically, this abstract representation involves selecting the right number of classes composing the topic space. This decision is crucial since topic model perplexity, which expresses its quality, is highly dependent on this feature. Furthermore, the multi-theme context of the study (see Figure 1) involves a more complex dialogue representation. In this paper, we propose to deal with these two drawbacks by using a compact representation from multiple topic spaces. This model is based on a robust multi-view representation of the textual documents.

A multi-view representation of a dialogue introduces both a *relevant* variability needed to represent different contexts of the dialogue, and a *noisy* variability related to topic space processing. Thus, a topic-based representation of a dialogue is built from the dialogue content itself. For this reason, the mapping process of a dialogue into several topic spaces generates a noisy variability related to the difference between the dialogue and the content of each class. In the same way, the relevant variability comes from the common content between the dialogue and the classes composing the topic space.

We propose to reduce the noisy variability by using a factor analysis technique, which was initially developed in the domain of speaker identification. In this field, the factor analysis paradigm is used as a decomposition model that enables to separate the representation space into two subspaces containing respectively useful and useless information. The general Joint Factor Analysis (JFA) paradigm (Kenny et al., 2008) considers multiple variabilities that may be cross-dependent. Therefore, JFA representation allows us to compensate the variability within sessions of a same speaker. This representation is an extension of the GMM-UBM (Gaussian Mixture Model-Universal Background Model) models (Reynolds and Rose, 1995). (Dehak et al., 2011) extract a compact super-vector (called an i-vector) from the GMM super-vector. The aim of the compression process (i-vector extraction) is to represent the super-vector variability in a low dimensional space. Although this compact representation is widely used in speaker recognition systems, this method has not been used yet in the field of text classification.

In this paper, we propose to apply factor analysis to compensate noisy variabilities due to the multiplication of LDA models. Furthermore, a normalization approach to condition dialogue representations (multi-model and i-vector) is presented. The two methods showed improvements for speaker verification: within Class Covariance Normalization (WCCN) (Dehak et al., 2011) and Eigen Factor Radial (EFR) (Bousquet et al., 2011). The latter includes length normalization (Garcia-Romero and Espy-Wilson, 2011). Both methods dilate the total variability space as a means of reducing the within-class variability. In our multi-model representation, the within class variability is redefined according to both dialogue content

(vocabulary) and topic space characteristics (word distributions among the topics). Thus, the speaker is represented by a theme, and the speaker session is a set of topic-based representations (frames) of a dialogue (session).

The paper is organized as follows. Section 2 presents previous related works. The dialogue representation is described in Section 3. Section 4 introduces the i-vector compact representation and presents its application to text documents. Sections 5 and 6 report experiments and results. The last section concludes and proposes some perspectives.

2 Related work

In the past, several approaches considered a text document as a mixture of latent topics. These methods, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Bellegarda, 1997), Probabilistic LSA (PLSA) (Hofmann, 1999) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003), build a higher-level representation of the document in a topic space. ¿ Document is then considered as a bag-of-words (Salton, 1989) where the word order is not taken into account. These methods have demonstrated their performance on various tasks, such as sentence (Bellegarda, 2000) or keyword (Suzuki et al., 1998) extraction.

In opposition to a multinomial mixture model, 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. Therefore, a document can change topics 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. PLSA and LDA models have been shown to generally outperform LSA on IR tasks (Hofmann, 2001). Moreover, LDA provides a direct estimate of the relevance of a topic given a word set. 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.

Topic-based approaches involve defining a number of topics composing the topic space. The choice of the "right" number of topics is a crucial step, especially when the documents may contain multiple themes. Many studies have tried to find a relevant method to deal with this issue. (Arun et al., 2010) proposed to use a Singular Value Decomposition (SVD) to represent the separability between the words contained in the vocabulary. Then, if the singular values of the topic-word matrix $\mathbf{M}$ equal the norm of the rows of $\mathbf{M}$, this means that the vocabulary is well separated among the topics. This method has to be evaluated with the Kullback-Liebler divergence metric for each topic space. However, this process would be time consuming for thousands of representations of a dialogue.

(Teh et al., 2004) proposed the Hierarchical Dirichlet Process (HDP) method to find the "right" number of topics by assuming that the data has a hierarchical structure. The HDP models were then compared to the LDA ones on the same dataset. (Zavitsanos et al., 2008) presented a method to *learn* the right depth of an ontology depending of the number of topics of LDA models. The study presented by (Cao et al., 2009) is quite similar to (Teh et al., 2004). The authors consider the average correlation between pairs of topics at each stage as the right number of topics.

All these methods assume that a document can have only one representation since they consider that finding the optimal topic model is the best solution. Another solution would be to consider a set of topic models to represent a document. Nonetheless, a multi-topic-based representation of a dialogue can involve a noisy variability due to the mapping of a dialogue in each topic space. Indeed, a dialogue does not share its content (*i.e.* words) with each class composing the topic space. Thus, a variability is added during the mapping process. Another weakness of the multi-view representation is the relation between classes in a topic space. (Blei and Lafferty, 2006) show that classes into a LDA topic space are correlated. Moreover, (Li and McCallum, 2006) consider a class as a node of an acyclic graph and as a distribution over other classes contained in the same topic space.

3 Multi-view representation of automatic dialogue transcriptions in a homogeneous space

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 relevant features that represent semantic contents 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) (Blei et al., 2003), as described in section 2.

Given a training 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 training corpus annotated with themes is available, supervised LDA (Griffiths and Steyvers, 2004) is not suitable for the proposed approach. LDA is used only for producing different feature sets used involved in statistical variability models.

In order to estimate the parameters of different hidden spaces, a set of discriminative words V is constructed as described in (Morchid et al., 2014a). Each theme t contains a set of specific words. Note that the same word may appear in several discriminative word sets. All the selected words are then merged without repetition to form V.

Several techniques, such as Variational Methods (Blei et al., 2003), Expectation-propagation (Minka and Lafferty, 2002) or Gibbs Sampling (Griffiths and Steyvers, 2004), have been proposed for estimating the parameters describing a LDA hidden space. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) (Geman and Geman, 1984) and gives a simple algorithm for approximate inference in high-dimensional models such as LDA (Heinrich, 2005). 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 $\overrightarrow{\beta}$.

Gibbs Sampling allows us both to estimate the LDA parameters in order to represent a new dialogue d with the r^{th} topic space of size n, and to obtain a feature vector $V_d^{z^r}$ of the topic representation of d. The j^{th} feature $V_d^{z_j^r} = P(z_j^r|d)$ (where $1 \leq j \leq n$) is the probability of topic z_j^r to be generated by the unseen dialogue d in the r^{th} topic space of size n (see Figure 2) and $V_{z_j^r}^w = P(w|z_j^r)$ is the vector representation of a word into r.

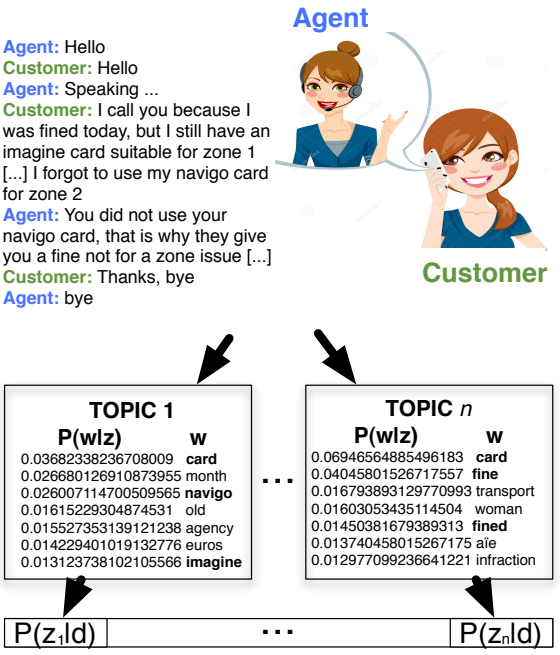


Figure 2: Example of a dialogue d mapped into a topic space of size n.

In the LDA technique, topic z_j, j is drawn from a multinomial over θ which is drawn from a Dirichlet distribution over $\overrightarrow{\alpha}$. Thus, a set of p topic spaces are learned using LDA by varying the number of topics n to obtain p topic spaces of size n. The number of topics n varies from 10 to $3,010$. Thus, a set of $3,000$ topic spaces is estimated. This is high enough to generate, for each dialogue, many feature sets for estimating the parameters of a variability model.

The next process allows us to obtain a homogeneous representation of transcription d for the r^{th} topic space r. The feature vector $V_d^{z^m}$ of d is mapped to the common vocabulary space V composed with a set of $|V|$ discriminative words (Morchid et al., 2014a) of size 166, to obtain a new feature vector $V_{d,r}^w = \{P(w|d)_r\}_{w \in V}$

of size $|V|$ for the r^{th} topic space r of size n where the i^{th} ($0 \leq i \leq |V|$) feature is:

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

where $\langle \cdot, \cdot \rangle$ is the inner product, δ being the frequency of the term w_i in d, $V_{z_j^r}^{w_i} = P(w_i|z_j)$ and $V_d^{z_j^r} = P(z_j|d)$ evaluated using Gibbs Sampling in the topic space r.

4 Compact multi-view representation

In this section, an i-vector-based method to represent automatic transcriptions is presented. Initially introduced for speaker recognition, i-vectors (Kenny et al., 2008) have become very popular in the field of speech processing and recent publications show that they are also reliable for language recognition (Martınez et al., 2011) and speaker diarization (Franco-Pedroso et al., 2010). I-vectors are an elegant way of reducing the imput space dimensionality while retaining most of the relevant information. The technique was originally inspired by the Joint Factor Analysis framework (Kenny et al., 2007). Hence, i-vectors convey the speaker characteristics among other information such as transmission channel, acoustic environment or phonetic content of speech segments. The next sections describe the i-vector extraction process, the application of this compact representation to textual documents (called c-vector), and the vector transformation with the EFR method and the Mahalanobis metric.

4.1 Total variability space definition

I-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech super-vectors according to a linear-Gaussian model. The speech (of a given speech recording) super-vector $\mathbf{m}_s$ of concatenated GMM means is projected in a low dimensionality space, named Total Variability space, with:

$$\mathbf{m}_{(h,s)} = m + \mathbf{T}\mathbf{x}_{(h,s)} \ , \qquad (1)$$

where m is the mean super-vector of the UBM[1]. $\mathbf{T}$ is a low rank matrix ($MD \times R$), where M is the number of Gaussians in the UBM and D is the cepstral feature size, which represents a basis of the reduced total variability space. $\mathbf{T}$ is named *Total Variability matrix*; the components of $\mathbf{x}_{(h,s)}$ are the total factors which represent the coordinates of the speech recording in the reduced total variability space called i-vector (i for $i$dentification).

4.2 From i-vector speaker identification to c-vector textual document classification

The proposed approach uses i-vectors to model transcription representation through each topic space in a homogeneous vocabulary space. These short segments are considered as basic semantic-based representation units. Indeed, vector V_d^w represents a segment or a session of a transcription d. In the following, (d, r) will indicate the dialogue representation d in the topic space r. In our model, the segment super-vector $\mathbf{m}_{(d,r)}$ of a transcription d knowing a topic space r is modeled:

$$\mathbf{m}_{(d,r)} = m + \mathbf{T}\mathbf{x}_{(d,r)} \qquad (2)$$

where $\mathbf{x}_{(d,r)}$ contains the coordinates of the topic-based representation of the dialogue in the reduced total variability space called c-vector (c for $c$lassification).

Let $\mathbf{N}_{(d,r)}$ and $\mathbf{X}_{(d,r)}$ be two vectors containing the zero order and first order dialogue statistics respectively. The statistics are estimated against the UBM:

$$\mathbf{N}_r[g] = \sum_{t \in r} \gamma_g(t); \ \{\mathbf{X}_{(d,r)}\}_{[g]} = \sum_{t \in (d,r)} \gamma_g(t) \cdot t \qquad (3)$$

where $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g for the observation t. In the equation, $\sum_{t \in (d,r)}$ represents the sum over all the frames belonging to the dialogue d.

Let $\overline{\mathbf{X}}_{(d,r)}$ be the state dependent statistics defined as follows:

$$\{\overline{\mathbf{X}}_{(d,r)}\}_{[g]} = \{\mathbf{X}_{(d,r)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{(d,r)} \mathbf{N}_{(d,r)}[g] \qquad (4)$$

Let $\mathbf{L}_{(d,r)}$ be a $R \times R$ matrix, and $\mathbf{B}_{(d,r)}$ a vector

[1]The UBM is a GMM that represents all the possible observations.

Algorithm 1: Estimation algorithm of **T** and latent variable **x**.

For each dialogue d mapped into the topic space r: $x_{(d,r)} \leftarrow 0$, $\mathbf{T} \leftarrow random$;
Estimate statistics: $\mathbf{N}_{(d,r)}$, $\mathbf{X}_{(d,r)}$ (eq.3);
for $i = 1$ *to* $nb\_iterations$ **do**
  for *all* d *and* r **do**
    Center statistics: $\overline{\mathbf{X}}_{(d,r)}$ (eq.4);
    Estimate $\mathbf{L}_{(d,r)}$ and $\mathbf{B}_{(d,r)}$ (eq.5);
    Estimate $\mathbf{x}_{(d,r)}$ (eq.6);
  end
  Estimate matrix **T** (eq. 7 and 8) ;
end

of dimension R, both defined as:

$$\mathbf{L}_{(d,r)} = \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(d,r)}[g] \cdot \{\mathbf{T}\}^t_{[g]} \cdot \mathbf{\Sigma}^{-1}_{[g]} \cdot \{\mathbf{T}\}_{[g]}$$
$$\mathbf{B}_{(d,r)} = \sum_{g \in \text{UBM}} \{\mathbf{T}\}^t_{[g]} \cdot \mathbf{\Sigma}^{-1}_{g} \cdot \{\overline{\mathbf{X}}_{(d,r)}\}_{[g]}, \tag{5}$$

By using $\mathbf{L}_{(d,r)}$ and $\mathbf{B}_{(d,r)}$, $\mathbf{x}_{(d,r)}$ can be obtained using the following equation:

$$\mathbf{x}_{(d,r)} = \mathbf{L}^{-1}_{(d,r)} \cdot \mathbf{B}_{(d,r)} \tag{6}$$

The matrix **T** can be estimated line by line, with $\{\mathbf{T}\}^i_{[g]}$ being the i^{th} line of $\{\mathbf{T}\}_{[g]}$ then:

$$\mathbf{T}^i_{[g]} = \mathbf{L}\mathbf{U}^{-1}_g \cdot \mathbf{R}\mathbf{U}^i_g, \tag{7}$$

where $\mathbf{R}\mathbf{U}^i_g$ and $\mathbf{L}\mathbf{U}_g$ are given by:

$$\mathbf{L}\mathbf{U}_g = \sum_{(d,r)} \mathbf{L}^{-1}_{(d,r)} + \mathbf{x}_{(d,r)}\mathbf{x}^t_{(d,r)} \cdot \mathbf{N}_{(d,r)}[g]$$
$$\mathbf{R}\mathbf{U}^i_g = \sum_{(d,r)} \{\overline{\mathbf{X}}_{(d,r)}\}^{[i]}_{[g]} \cdot \mathbf{x}_{(d,r)} \tag{8}$$

Algorithm 1 presents the method adopted to estimate the multi-view variability dialogue matrix with the above developments where the standard likelihood function can be used to assess the convergence. One can refer to (Matrouf et al., 2007) to find out more about the implementation of the factor analysis.

C-vector representation suffers from 3 raised c-vector issues: (i) the c-vectors x of equation 2 have to be theoretically distributed among the normal distribution $\mathcal{N}(0, I)$, (ii) the "radial" effect should be removed, and (iii) the full rank total factor space should be used to apply discriminant transformations. The next section presents a solution to these 3 problems.

4.3 C-vector standardization

A solution to standardize c-vectors has been developed in (Bousquet et al., 2011). The authors proposed to apply transformations for training and test transcription representations. The first step is to evaluate the empirical mean $\overline{\mathbf{x}}$ and covariance matrix **V** of the training c-vector. Covariance matrix **V** is decomposed by diagonalization into:

$$\mathbf{P}\mathbf{D}\mathbf{P}^T \tag{9}$$

where **P** is the eigenvector matrix of **V** and **D** is the diagonal version of **V**. A training i-vector $\mathbf{x}_{(d,r)}$ is transformed in $\mathbf{x}'_{(d,r)}$ as follows:

$$\mathbf{x}'_{(d,r)} = \frac{\mathbf{D}^{-\frac{1}{2}}\mathbf{P}^T(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}{\sqrt{(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}} \tag{10}$$

The numerator is equivalent by rotation to $\mathbf{V}^{-\frac{1}{2}}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})$ and the Euclidean norm of $\mathbf{x}'_{(d,r)}$ is equal to 1. The same transformation is applied to the test c-vectors, using the training set parameters $\overline{\mathbf{x}}$ and mean covariance **V**as estimations of the test set of parameters.

Figure 3 shows the transformation steps: Figure 3-(a) is the original training set; Figure 3-(b) shows the rotation applied to the initial training set around the principal axes of the total variability when $\mathbf{P}^T$ is applied; Figure 3-(c) shows the standardization of c-vectors when $\mathbf{D}^{-\frac{1}{2}}$ is applied; and finally, Figure 3-(d) shows the c-vector $\mathbf{x}'_{(d,r)}$ on the surface area of the unit hypersphere after a length normalization by a division of $\sqrt{(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_{(d,r)} - \overline{\mathbf{x}})}$.

5 Experimental Protocol

The proposed c-vector representation of automatic transcriptions is evaluated in the context of the theme identification of a human-human telephone conversation in the customer care service (CCS) of the RATP Paris transportation system. The metric used to identify of the best theme is the Mahalanobis metric.

5.1 Theme identification task

The DECODA project corpus (Bechet et al., 2012) was designed to perform experiments on the identification of conversation themes. It is composed of 1,514 telephone conversations, corresponding to about 74 hours of signal, split into a training

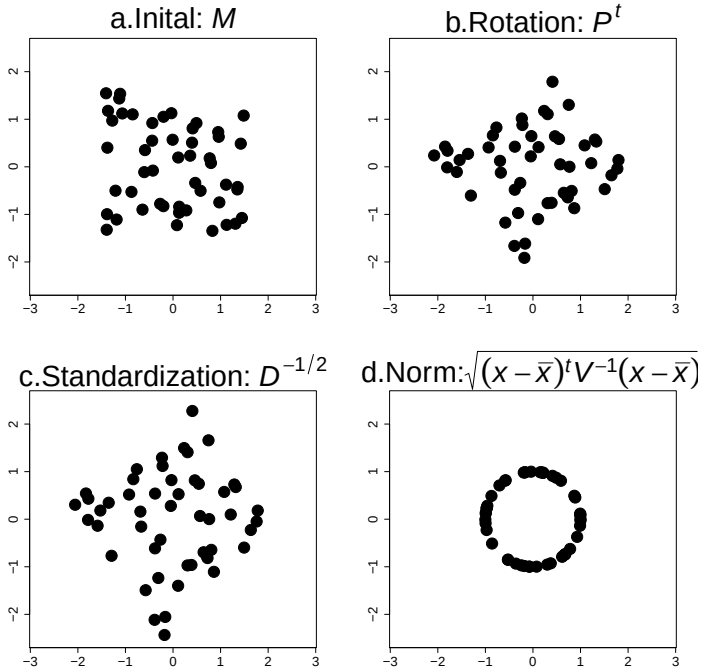


Figure 3: Effect of the standardization with the EFR algorithm.

set (740 dialogues), a development set (175 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*.

An LDA model allowed us to elaborate 3,000 topics spaces by varying the number of topics from 10 to 3,010. A topic space having less than 10 topics is not suitable for a corpus of more than 700 dialogues (training set). For each theme $\{C_i\}_{i=1}^{8}$, a set of 50 specific words is identified. All the selected words are then merged without repetition to compose V, which is made of 166 words. The topic spaces are made with the LDA Mallet Java implementation[2].

The LIA-Speeral ASR system (Linarès et al., 2007) is used for the experiments. 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. A "stop list" of 126 words[3] 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).

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

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

5.2 Mahalanobis metric

Given a new observation x, the goal of the task is to identify the theme belonging to x. Probabilistic approaches ignore the process by which c-vectors were extracted and they pretend instead they were generated by a prescribed generative model. Once a c-vector is obtained from a dialogue, its representation mechanism is ignored and it is regarded as an observation from a probabilistic generative model. The Mahalanobis scoring metric assigns a dialogue d with the most likely theme C. Given a training dataset of dialogues, let $\mathbf{W}$ denote the within dialogue covariance matrix defined by:

$$\begin{aligned}\mathbf{W} &= \sum_{k=1}^{K} \frac{n_t}{n} \mathbf{W_k} \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{i=0}^{n_t} \left(x_i^k - \overline{x_k}\right) \left(x_i^k - \overline{x_k}\right)^t \end{aligned} \quad (11)$$

where $\mathbf{W_k}$ is the covariance matrix of the k^{th} theme C_k, n_t is the number of utterances for the theme C_k, n is the total number of dialogues, and $\overline{x_k}$ is the centroid (mean) of all dialogues x_i^k of C_k.

Each dialogue does not contribute to the covariance in an equivalent way. For this reason, the term $\frac{n_t}{n}$ is introduced in equation 11. If homoscedasticity (equality of the class covariances) and Gaussian conditional density models are assumed, a new observation x from the test dataset can be assigned to the most likely theme $C_{k_{\text{Bayes}}}$ using the classifier based on the Bayes decision rule:

$$C_{k_{\text{Bayes}}} = \arg\max_k \{\mathcal{N}(x \mid \overline{x_k}, \mathbf{W})\}$$
$$= \arg\max_k \left\{ -\frac{1}{2} (x - \overline{x_k})^t \mathbf{W}^{-1} (x - \overline{x_k}) + a_k \right\}$$

where $\mathbf{W}$ is the within theme covariance matrix defined in eq. 11; $\mathcal{N}$ denotes the normal distribution and $a_k = \log(P(C_k))$. It is noted that, with these assumptions, the Bayesian approach is similar to Fisher's geometric approach: x is assigned to the class of the nearest centroid, according to the Mahalanobis metric (Xing et al., 2002) of $\mathbf{W}^{-1}$:

$$C_{k_{\text{Bayes}}} = \arg\max_k \left\{ -\frac{1}{2} ||x - \overline{x_k}||^2_{\mathbf{W}^{-1}} + a_k \right\}$$

6 Experiments and results

The proposed c-vector approach is applied to the same classification task and corpus proposed in (Morchid et al., 2014a; Morchid et al., 2014b; Morchid et al., 2013) (state-of-the-art in text classification in (Morchid et al., 2014a)). Experiments are conducted using the multiple topic spaces estimated with an LDA approach. From these multiple topic spaces, a classical way is to find the one that reaches the best performance. Figure 4 presents the theme classification performance obtained on the development and test sets using various topic-based representation configurations with the EFR normalization algorithm (*baseline*).

For sake of comparison, experiments are conducted using the automatic transcriptions only (ASR) only. The conditions indicated by the abbreviations between parentheses are considered for the development (Dev) and the test (Test) sets.

Only homogenous conditions (ASR for both training and validations sets) are considered in this study. Authors in (Morchid et al., 2014a) notice that results collapse dramatically when heterogenous conditions are employed (TRS or TRS+ASR for training set and ASR for validation set).

First of all, we can see that this baseline approach reached a classification accuracy of 83% and 76%, respectively on the development and the test sets. However, we note that the classification performance is rather unstable, and may completely change from a topic space configuration to another. The gap between the lower and the higher classification results is also important, with a difference of 25 points on the development set (the same trend is observed on the test set). As a result, finding the best topic space size seems crucial for this classification task, particularly in the context of highly imperfect automatic dialogue transcriptions containing more than one theme.

The topic space that yields the best accuracy with the baseline method ($n = 15$ topics) is presented in Figure 5. This figure presents each of the 15 topics and their 10 most representative words (highest $P(w|z)$). Several topics contain more or less the same representative words, such as topics 3, 6 and 9. This figure points out some interesting topics that allow us to distinguish a theme from the others. For example:

- topics 2, 10 and 15 represent some words related to *itinerary problems*,
- the *transportation cards* theme is mostly represented in topic 4 and 15 (*Imagine* and *Navigo* are names of transportation cards),
- the words which represent the *time schedules* theme are contained in topic 5,6,7 and less in topic 9,
- *state of the traffic* could be discussed with words such as: *departure, line, service, day*. These words and others are contained in topic 13,
- topics 4 and 12 are related to the *infractions* theme with to words *fine*, *pass*, *zone* or *ticket*,
- but topic 12 could be related to theme *fares* or *special offers* as well .

Table 1 presents results obtained with the proposed c-vector approach coupled with the EFR algorithm. We can firstly note that this compact representation allows it to outperform the best topic space configuration (*baseline*), with a gain of 9.4 points on the development data and of 9 points on the test data. Moreover, if we consider the different c-vector configurations with the development

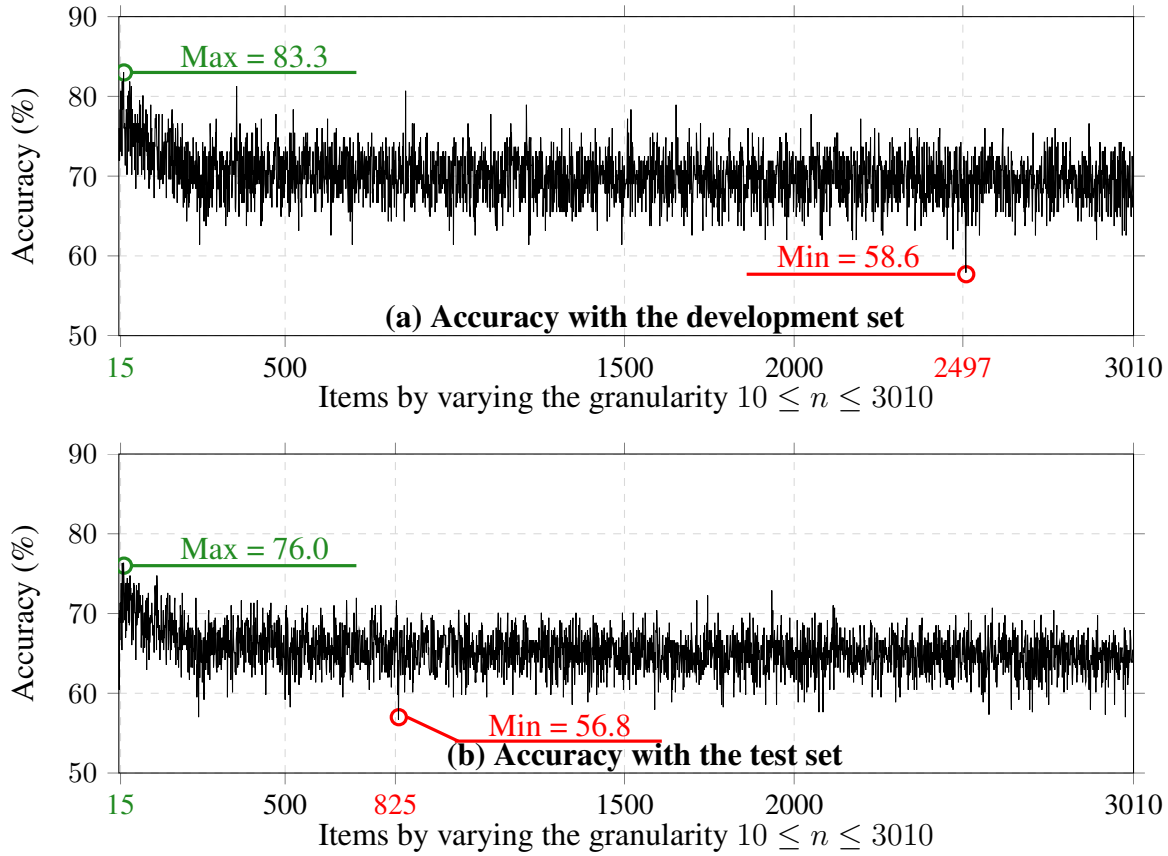


Figure 4: Theme classification accuracies using various topic-based representations with EFR normalization (baseline) on the development and test sets (X-coordinates start at 10 indeed, but to show the best configuration point (15), the origine (10) has been removed).

Table 1: Theme classification accuracy (%) with different c-vectors and GMM-UBM sizes.

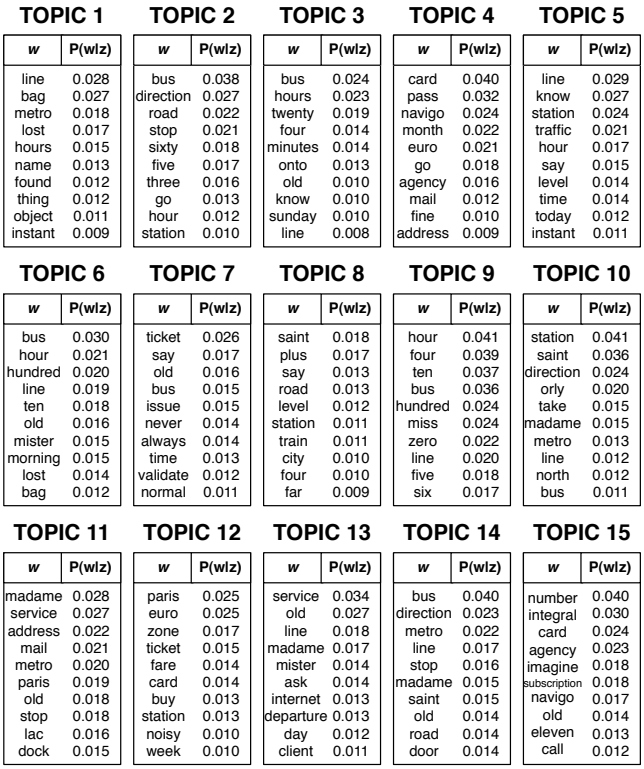
	DEV				*TEST*			
c-vector size	**Number of Gaussians in GMM-UBM**							
	32	*64*	*128*	*256*	*32*	*64*	*128*	*256*
60	88.8	*86.5*	91.2	90.6	85.0	*82.6*	83.5	84.7
100	91.2	**92.4**	92.4	87.7	86.0	**85.0**	83.5	84.7
120	89.5	92.2	89.5	87.7	85.0	83.5	85.4	84.1

Table 2: Maximum (Max), minimum (Min) and Difference ($Max - Min$) theme classification accuracies (%) using the baseline and the proposed c-vector approaches.

	Max		**Min**		**Difference**	
Method	*DEV*	*TEST*	*DEV*	*TEST*	*DEV*	*TEST*
baseline	83.3	76.0	58.6	56.8	14.7	20.8
c-vector	**92.4**	**85.0**	86.5	82.6	5.9	2.4

and test sets, the gap between accuracies is much smaller: classification accuracy does not go below 82.6%, while it reached 56% for the worst topic-based configuration. Indeed, as shown in Table 2, the difference between the maximum and the minimum theme classification accuracies is of 20% using the baseline approach while it is only of 2.4% using the c-vector method.

We can conclude that this original c-vector approach allows one to better handle the variabilities



TOPIC 1		TOPIC 2		TOPIC 3		TOPIC 4		TOPIC 5	
w	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)
line	0.028	bus	0.038	bus	0.024	card	0.040	line	0.029
bag	0.027	direction	0.027	hours	0.023	pass	0.032	know	0.027
metro	0.018	road	0.022	twenty	0.019	navigo	0.024	station	0.024
lost	0.017	stop	0.021	four	0.014	month	0.022	traffic	0.021
hours	0.015	sixty	0.018	minutes	0.014	euro	0.021	hour	0.017
name	0.013	five	0.017	onto	0.013	go	0.018	say	0.015
found	0.012	three	0.016	old	0.010	agency	0.016	level	0.014
thing	0.012	go	0.013	know	0.010	mail	0.012	time	0.014
object	0.011	hour	0.012	sunday	0.010	fine	0.010	today	0.012
instant	0.009	station	0.010	line	0.008	address	0.009	instant	0.011

TOPIC 6		TOPIC 7		TOPIC 8		TOPIC 9		TOPIC 10	
w	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)
bus	0.030	ticket	0.026	saint	0.018	hour	0.041	station	0.041
hour	0.021	say	0.017	plus	0.017	four	0.039	saint	0.036
hundred	0.020	old	0.016	say	0.013	ten	0.037	direction	0.024
line	0.019	bus	0.015	road	0.013	bus	0.036	orly	0.020
ten	0.018	issue	0.015	level	0.012	hundred	0.024	take	0.015
old	0.016	never	0.014	station	0.011	miss	0.024	madame	0.015
mister	0.015	always	0.014	train	0.011	zero	0.022	metro	0.013
morning	0.015	time	0.013	city	0.010	line	0.020	line	0.012
lost	0.014	validate	0.012	four	0.010	five	0.018	north	0.012
bag	0.012	normal	0.011	far	0.009	six	0.017	bus	0.011

TOPIC 11		TOPIC 12		TOPIC 13		TOPIC 14		TOPIC 15	
w	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)	*w*	P(w|z)
madame	0.028	paris	0.025	service	0.034	bus	0.040	number	0.040
service	0.027	euro	0.025	old	0.027	direction	0.023	integral	0.030
address	0.022	zone	0.017	line	0.018	metro	0.022	card	0.024
mail	0.021	ticket	0.015	madame	0.017	line	0.017	agency	0.023
metro	0.020	fare	0.014	mister	0.014	stop	0.016	imagine	0.018
paris	0.019	card	0.014	ask	0.014	madame	0.015	subscription	0.018
old	0.018	buy	0.013	internet	0.013	saint	0.015	navigo	0.017
stop	0.018	station	0.013	departure	0.013	old	0.014	old	0.014
lac	0.016	noisy	0.010	day	0.012	road	0.014	eleven	0.013
dock	0.015	week	0.010	client	0.011	door	0.014	call	0.012

Figure 5: Topic space (15 topics) that obtains the best accuracy with the baseline system (see Fig. 4).

contained in dialogue conversations: in a classification context, better accuracy can be obtained and the results can be more consistent when varying the c-vector size and the number of Gaussians.

7 Conclusions

This paper presents an original multi-view representation of automatic speech dialogue transcriptions, and a fusion process with the use of a factor analysis method called i-vector. The first step of the proposed method is to represent a dialogue in multiple topic spaces of different sizes (*i.e.* number of topics). Then, a compact representation of the dialogue from the multiple views is processed to compensate the vocabulary and the variability of the topic-based representations. The effectiveness of the proposed approach is evaluated in a classification task of theme dialogue identification. Thus, the architecture of the system identifies conversation themes using the i-vector approach. This compact representation was initially developed for speaker recognition and we showed that it can be successfully applied to a text classification task. Indeed, this solution allowed the system to obtain better classification accuracy than with the use of the classical best topic space configuration. In fact, we highlighted that this original compact version of all topic-based representations of dialogues, called c-vector in this work, coupled with the EFR normalization algorithm, is a better solution to deal with dialogue variabilities (high word error rates, bad acoustic conditions, unusual word vocabulary, etc). This promising compact representation allows us to effectively solve both the difficult choice of the right number of topics and the multi-theme representation issue of particular textual documents. Finally, the classification accuracy reached 85% with a gain of 9 points compared to usual baseline (best topic space configuration). In a future work, we plan to evaluate this new representation of textual documents in other information retrieval tasks, such as keyword extraction or automatic summarization systems.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. 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.

References

R. Arun, Venkatasubramaniyan Suresh, C.E. Veni Madhavan, and Musti Narasimha Murty. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining*, pages 391–402. Springer.

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

Jerome R. Bellegarda. 1997. A latent semantic analysis framework for large-span language modeling. In *Fifth European Conference on Speech Communication and Technology*.

Jerome R. Bellegarda. 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

David M. Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Pierre-Michel Bousquet, Driss Matrouf, and Jean-François Bonastre. 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In *Interspeech*, pages 485–488.

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7):1775–1781.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Javier Franco-Pedroso, Ignacio Lopez-Moreno, Doroteo T Toledano, and Joaquin Gonzalez-Rodriguez. 2010. Atvs-uam system description for the audio segmentation and speaker diarization albayzin 2010 evaluation. In *FALA VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop*, pages 415–418.

Daniel Garcia-Romero and Carol Y Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

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

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*, page 21. Citeseer.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. 2007. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447.

Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel. 2008. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations.

Georges Linarès, Pascal Nocéra, Dominique Massonie, and Driss Matrouf. 2007. The lia speech recognition system: from 10xrt to 1xrt. In *Text, Speech and Dialogue*, pages 302–308. Springer.

David Martınez, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka. 2011. Language recognition in ivectors space. *Interspeech*, pages 861–864.

Driss Matrouf, Nicolas Scheffer, Benoit G.B. Fauve, and Jean-Francois Bonastre. 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Interspeech*, pages 1242–1245.

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

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

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

Mohamed Morchid, Richard Dufour, and Georges Linarès. 2014b. A LDA-Based Topic Classification Approach from Highly Imperfect Automatic Transcriptions. In *LREC*.

Douglas A. Reynolds and Richard C. Rose. 1995. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Gerard Salton. 1989. Automatic text processing: the transformation. *Analysis and Retrieval of Information by Computer*.

Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. 1998. Keyword extraction using term-domain interdependence for dictation of radio news. In *17th international conference on Computational linguistics*, volume 2, pages 1272–1276. ACL.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*.

Eric P. Xing, Michael I. Jordan, Stuart Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512.

Elias Zavitsanos, Sergios Petridis, Georgios Paliouras, and George A. Vouros. 2008. Determining automatically the size of learned ontologies. In *ECAI*, volume 178, pages 775–776.