A LDA-Based Topic Classification Approach from Highly Imperfect Automatic Transcriptions

Mohamed Morchid, Richard Dufour, Georges Linarès

LIA - University of Avignon (France)
{mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr

Abstract

Although the current transcription systems could achieve high recognition performance, they still have a lot of difficulties to transcribe speech in very noisy environments. The transcription quality has a direct impact on classification tasks using text features. In this paper, we propose to identify themes of telephone conversation services with the classical Term Frequency-Inverse Document Frequency using Gini purity criteria (TF-IDF-Gini) method and with a Latent Dirichlet Allocation (LDA) approach. These approaches are coupled with a Support Vector Machine (SVM) classification to resolve theme identification problem. Results show the effectiveness of the proposed LDA-based method compared to the classical TF-IDF-Gini approach in the context of highly imperfect automatic transcriptions. Finally, we discuss the impact of discriminative and non-discriminative words extracted by both methods in terms of transcription accuracy.

Keywords: Speech analytics; Topic identification; Latent Dirichlet Allocation

1. Introduction

The application considered in this paper concerns the automatic analysis of telephone conversations between an agent and a customer in a customer care service of the Paris transportation system. The agent follows a conversation protocol to address customer requests or complains. One purpose of the application is to identify themes that appear in the conversation. A conversation may contain more than one semantically related theme, but not all of them are relevant for the application task. For example, a customer may inquiry about an object lost on a transportation mean that was late. In such a case, the loss is a much more relevant theme than the traffic state. In this situation, agents annotate a conversation with what they consider the major theme of the customer request. This leads to annotate a theme for each conversation.

This paper presents a system for the automatic extraction of themes from conversations acquired during the daily operation of a call centre in Paris. The system generates hypotheses about the most relevant theme of each conversation. The major difficulty of this classification task concerns the unpredictable behavior of the customers. Conversations may contain very noisy segments and are decoded by an Automatic Speech Recognition (ASR) component.

In the context of Information Retrieval (IR) tasks, the main feature used is the *word term frequency*. This specific feature allows to obtain a subset of discriminative words for a considered class (a "theme" in this study). The term "discriminative" is associated to a word if it permits to discern a class from the others. Finally, this set of discriminative words should permit to compose a vector representation of conversation themes in the semantic space.

While the term frequency is a performant feature in the context of manually written texts, its application to automatic transcriptions seems to be more difficult since transcription errors are inevitable. Indeed, these errors would lead to an incorrect representation of the discriminative words. For this reason, the projection of the automatically transcribed words in a more abstracted space could increase the robustness to the ASR errors.

In this paper, we propose to compare two unsupervised representations of discriminative words to automatically identify themes of telephone conversations in different configurations of highly imperfect transcriptions. The classical Term Frequency-Inverse Document Frequency with Gini purity criteria (TF-IDF-Gini) method (Robertson, 2004) is firstly applied to extract discriminative words for each theme to identify from transcriptions. We secondly propose to explore a topic space representation of discriminative words with the use of the Latent Dirichlet Allocation (LDA) approach (Blei et al., 2003). Each representation is finally used to train a Support Vector Machine (SVM) classifier to automatically associate a theme to a conversation.

We also propose in this article a discussion about the classification performance impact of discriminative and non-discriminative words chosen by both methods in terms of transcription accuracy.

2. Related work

Recent reviews for spoken conversation analysis, speech analytics, topic identification and segmentation can be found in (Tur and De Mori, 2011), (Melamed and Gilbert, 2011), (Hazen, 2011) and (Purver, 2011) respectively. The classical Term Frequency-Inverse Document Frequency (TF-IDF) (Robertson, 2004) has been widely used for extracting discriminative words from texts. Works also found improvements associating TF-IDF with the Gini purity criteria (Dong et al., 2011).

Other approaches proposed to consider the 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. All of these methods are commonly used in the Information Retrieval (IR) field. They consider documents as a bag-of-words (Salton, 1989) without taking ac-

count of word order; nevertheless, they demonstrated their performance on various tasks.

LDA is a generative model which considers a document, seen as a bag-of-words, as a mixture probability of latent topics. 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 with the complete document. Thereby, a document can change of topics from a word to another. However, the word occurrences are connected by a latent variable which controls the global respect of the distribution of the topics in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them. 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 knowing a word set.

Support Vector Machines (SVM) are a set of supervised learning techniques. Knowing a sample, SVMs determine a separation plan between parts of the sample called *support vector*. Then, a separating hyperplane that maximizes the *margin* between the support vectors and the hyperplane separator (Vapnik, 1963) is calculated. SVMs were used for the first time by (Boser et al., 1992) both in regression (Müller et al., 1997) and in classification (Joachims, 1999) tasks. The SVM popularity is due to the good results achieved in these two specific tasks and the low number of parameters requiring adjustment.

A LDA-based approach combined with a SVM classification process has recently been studied in various domains, such as biology (hua Yeh and hsing Chen, 2010), text classification (Zrigui et al., 2012), stylometry (Arun et al., 2009), audio information retrieval (Kim et al., 2009), social event detection (Morchid et al., 2013a) or image detection (Tang et al., 2009). To our knowledge, a combined LDA-SVM approach has not yet been applied to theme classification of highly imperfect automatic transcriptions but was used in the context of keyword and keyphrase extraction in automatic transcriptions (Sheeba and Vivekanandan, 2012). The TF-IDF extraction method coupled with a SVM classification, which constitutes our baseline system, has been widely studied in text classification such as (Lan et al., 2005; Georgescul et al., 2006).

3. Theme identification methods

This section presents the proposed theme classification system using discriminative words extracted from highly imperfect transcriptions. The system is composed of two main parts. The first one creates a vector representation of words with two different unsupervised approaches: a term frequency Okapi/BM25 vector (Robertson, 2004) with the TF-IDF-Gini method (Dong et al., 2011) and a topic space representation with the LDA approach (Blei et al., 2003). The second part uses the extracted vectors to learn SVM classifiers. Figure 1 presents the global architecture of the proposed classification system using manual (TRS) and automatic (ASR) transcriptions.

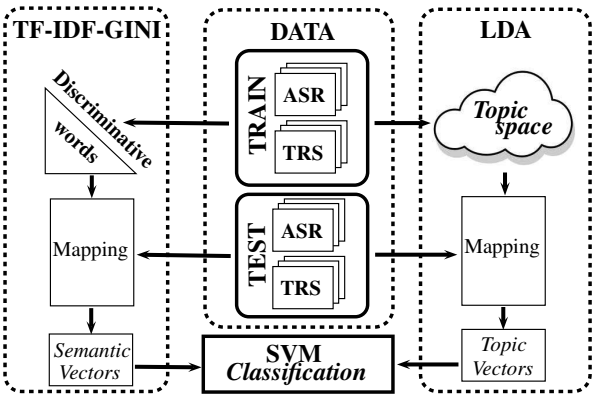


Figure 1: General approach of the classification system.

3.1. Description of dialogue features

To perform the classification task, a features representation of each dialogue is needed. Thus, the next sections describe two different representations of a dialogue using a discriminative terms list and a topic space.

3.1.1. Discriminative terms

Let's consider a corpus D of dialogues d with a word vocabulary $\mathbf{V} = \{w_1, \ldots, w_N\}$ of size N where d is seen as a bag-of-words (Salton, 1989). A term of $\mathbf{V}$ is chosen according to its importance δ_t in the theme t by calculating its Term Frequency (TF), its Inverse Document Frequency (IDF) (Robertson, 2004) and the Gini purity criteria (Dong et al., 2011) that is common for all the themes. This set of scores δ composes the frequency model f:

$$\delta_t^w = tf_t(w) \times idf_t(w) \times gini_t(w)$$

Then the words with highest scores Δ for all themes $\mathbf{T}$ are extracted and constitute a discriminative word subset $\mathbf{V_\Delta}$ (each theme $t \in \mathbf{T}$ has its own score δ_t) and its own frequency γ in the model f (Morchid et al., 2013b):

$$\gamma_f^t = \frac{\#d \in t}{\#d \in D}$$

Note that a same word w can be present in different themes, but with different scores (TF-IDF-Gini) depending of its relevance in the theme:

$$\begin{aligned}\Delta(w) = P(w|f) &= \int_t P(w|t)P(t|f)\,\mathrm{dt} \\ &= \sum_{t\in\mathbf{T}} P(w|t)P(t|f) \\ &= \sum_{t\in\mathbf{T}} \delta_t^w \times \gamma_f^t \\ &= \left\langle \overrightarrow{\delta^w}, \overrightarrow{\gamma^f} \right\rangle \end{aligned} \quad (1)$$

3.1.2. Semantic representation

For each dialogue $d \in D$, a semantic feature vector V_d^s is determined. The n^{th} ($1 \leq n \leq |\mathbf{V_\Delta}|$) feature $V_d^s[n]$, is composed with the number of occurrences of the word

w_n ($|w_n|$) in d and the score Δ of w_n (see eq 1) in the discriminative word set $\mathbf{V_\Delta}$:

$$V_d^s[n] = |w_n| \times \Delta(w_n) \quad (2)$$

3.1.3. Topic representation

The topic representation is performed using a Latent Dirichlet Allocation (LDA) based approach (see section 2.). A thematic space m of n topics is then obtained with, for each theme z, the probability of each word w of $\mathbf{V}$ knowing z ($P(w|z) = V_z^w$) and for the entire model m, the probability of each theme z knowing the model m ($P(z|m) = V_m^z$). For every dialogue d of a corpus D, a first parameter θ is drawn according to a Dirichlet law of parameter α. A second parameter ϕ is drawn according to the same Dirichlet law of parameter β. Then, to generate every word w of the document d, a latent topic z is drawn from a multinomial distribution on θ. Knowing this topic z, the distribution of the words is a multinomial of parameters ϕ. The parameter θ is drawn for all the documents from the same *prior* parameter α. This allows to obtain a parameter binding the documents all together (Blei et al., 2003).

Mapping of conversations/topic space

The Gibbs sampling algorithm (Griffiths and Steyvers, 2002) was used to infer a dialogue d with the n topics of the thematic space m. This algorithm is based on the Markov Chain Monte Carlo (MCMC) method. Thus, the Gibbs sampling allows to obtain samples of the distribution parameters θ knowing a word w of a test document and a given topic z. A feature vector V_z^d of the topic representation of d is then obtained. The k^{th} feature $V_z^d[k]$ (where $1 \leq k \leq n$) is the probability of the topic z_k knowing the dialogue d:

$$V_z^d[k] = P(z_k|d) \quad (3)$$

3.2. SVM classification

In this part, classifiers are trained with the vector representation of words to automatically assign the most relevant theme to each conversation. The classification of conversations requires a multi-class classifier. The *one-against-one* method is chosen with a linear kernel. This method gives a better testing accuracy than the *one-against-rest* method (Yuan et al., 2012). For this multi-theme problem, T denotes the number of themes and $t_i, i = 1, \ldots, T$ denotes the T themes. A binary classifier is used with a linear kernel for every pair of distinct theme. As a result, all together binary classifiers $T(T-1)/2$ are constructed. The binary classifier $C_{i,j}$ is trained from example data where t_i is a positive class and t_j a negative class ($i \neq j$).

For a new vector representation (semantic eq. 2 or topic eq. 3) of a dialogue d from the test corpus, if $C_{i,j}$ means that d is in the theme t_i, then the vote for the class t_i is added by one. Otherwise, the vote for the theme t_j is increased by one. The dialogue d is finally assigned with the theme having the highest number of votes.

4. Experiments

The next sections describe the experimental protocol and evaluate both dialogue representations and classification methods. Furthermore, a short study gives some interesting perspectives for WER consideration and determination knowing a task.

4.1. Experimental protocol

In order to perform experiments on the conversation theme identification, the corpus of the DECODA project was used (Bechet et al., 2012). This corpus is composed of 1,067 telephone conversations split into a train set (740 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*.

The train set is used to compose a subset of discriminative words (section 3.1.). This set allows to elaborate a semantic space for each conversation of the test corpus with the basic TF-IDF-Gini method. In the experiments, the number of discriminative words has been varied from 800 to the total number of words contained in the train corpus (7,920 words). The test corpus contains 3,806 words (70.8% occur in the train corpus).

In the same way, a topic vector is calculated by mapping each dialogue of the test corpus with each topic space. A set of 25 topic spaces with a different topic number ($\{5, \ldots, 600\}$) is elaborated by using a LDA model in the train corpus (example: test = TRS $\rightarrow$ LDA train corpus = TRS) . The topic spaces are made with the Mallet Java implementation (McCallum, 2002) of LDA.

Then, for both configurations (semantic or topic vector), a SVM classifier is learned with the LIBSVM library (Chang and Lin, 2011). SVM parameters are optimized by cross validation on train corpus.

The LIA-Speeral ASR system is used for the experiments (Linarès et al., 2007). This system results in an overall Word Error Rate (WER) of 45.8% (train set) and of 58.0% (test set). These high error rates are mainly due to speech disfluencies and to adverse acoustic environments. A "stop list" of 126 words[1] was used to remove unnecessary words which results in a WER of 33.8% (train set) and of 49.5% (test set).

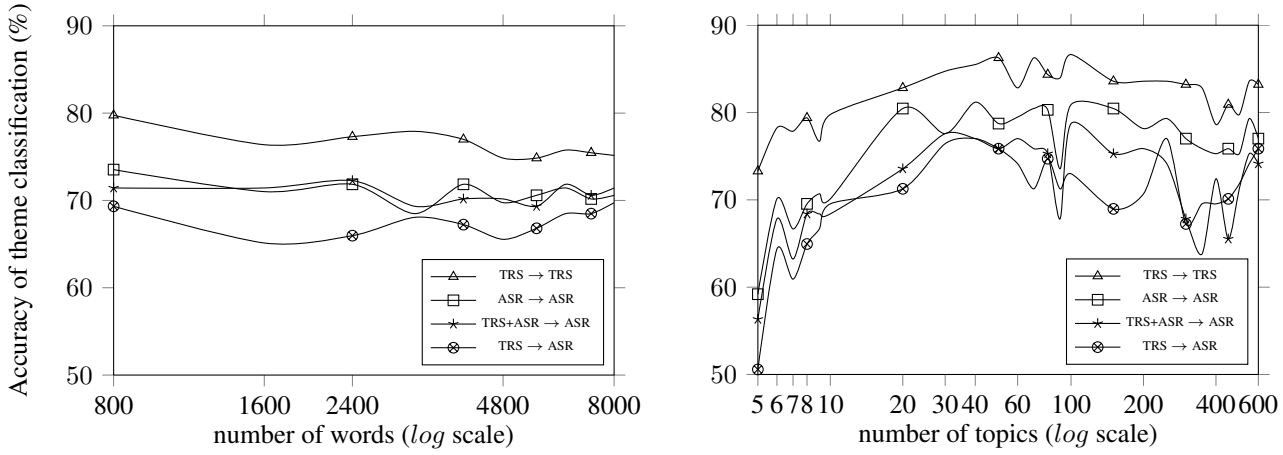
Experiments are conducted with the two unsupervised methods (TF-IDF-Gini / LDA) on the manual (TRS) and the automatic transcriptions only (ASR). We also propose to study the combination of both manual and automatic transcriptions (TRS+ASR) in order to see if ASR errors can be supplied by the correct reference words.

4.2. Theme classification performance

Figure 2 presents the theme classification accuracy obtained by the TF-IDF-Gini and the LDA approaches on the test corpus for all transcription configurations (TRS/ASR) when varying the word extraction conditions (number of discriminative words and number of topics). We can see that the LDA-based method outperforms the best theme classification accuracies obtained by the TF-IDF-Gini approach (see table 1).

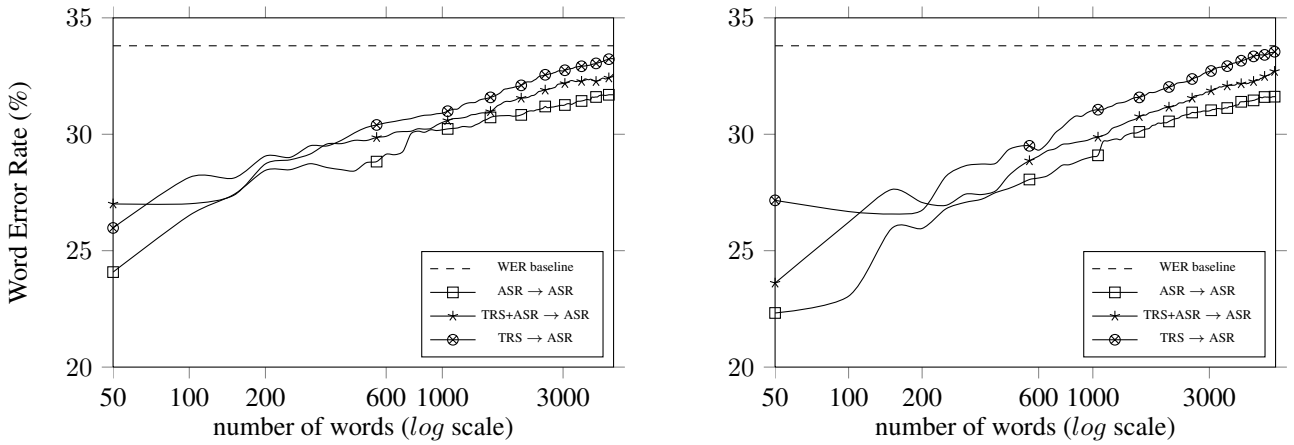
As expected, the TRS train / TRS test configuration (TRS $\rightarrow$ TRS) gives the best classification results with a

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



(a) **TF-IDF-Gini** (b) **LDA**

Figure 2: Theme classification performance by varying the number of discriminative words (a) and the number of topic spaces (b).



(a) **TF-IDF-Gini** (b) **LDA**

Figure 3: Word Error Rate of the n most discriminative words using TF-IDF-Gini (a) and LDA (b) approaches.

DATA		*BEST ACCURACY (%)*			
Train	**Test**	**#words**	**TF-IDF-Gini**	**#topics**	**LDA**
TRS	TRS	*800*	79.7	*100*	**86.6**
TRS	ASR	*8000*	69.7	*40*	**77.0**
ASR	ASR	*800*	73.5	*60*	**81.4**
ASR+TRS	ASR	*2400*	72.2	*100*	**78.7**

Table 1: Theme classification accuracy (*Confidence interval of* $\pm 3.69\%$ *for the LDA system*)

gain of 6.9 points with the LDA method. When comparing the training corpus types, we can also note that best performance on the ASR test is obtained with the ASR training data. A gain of 10.9 points is noted with the LDA method compared to the TF-IDF-Gini approach on the automatic transcriptions of conversations. It seems clear that using comparable training and testing configurations allows to achieve the best classification performance, whether it be on manual or on automatic transcriptions.

We can finally note that the LDA approach performance has a tendency to fluctuate when varying the number of topics. This could be explained by the high Word Error Rate (WER) of the targeted corpus: indeed, the words chosen as discriminative in particular topic number conditions could be wrongly transcribed in a high proportion. We can support this assumption by analyzing results obtained using 90 topics on the figure 2. An important performance drop is observed for the ASR training conditions (ASR $\rightarrow$ ASR and ASR $\rightarrow$ TRS) while a smaller performance lost is seen when including the reference transcriptions during the training process (ASR+TRS $\rightarrow$ ASR and TRS $\rightarrow$ TRS).

4.3. Transcription accuracy of discriminative words

While the performance with the TF-IDF-Gini approach is clearly better on manual transcriptions (table 1), the performance is almost identical on manual and on automatic transcriptions with the LDA method (respectively 86.6% and 81.4% of classification accuracy). We think that the LDA-based approach can better manage the errors contained in the automatic transcriptions by choosing discriminative words depending on their transcription accuracy. Figure 3 compares the Word Error Rates (WER) of the n

most discriminative words using TF-IDF-Gini and LDA approaches on all the configurations (TRS/ASR). The score $s(w)$ used to find the most relevant words for the LDA approach is computed with:

$$\begin{aligned} s(w) = P(w|m) &= \int_z P(w|z)P(z|m)\,\mathrm{dz} \\ &= \sum_{z\in m} P(w|z)P(z|m) \\ &= \sum_{z\in m} V_z^w \times V_m^z \\ &= \left\langle \overrightarrow{V^w}, \overrightarrow{V^m} \right\rangle \end{aligned}$$

where $\overrightarrow{V^w}$ is the vector representation of a word w in all topics z of the topic space m, $\overrightarrow{V^m}$ is the vector representation of all the topics z in m and $\langle\cdot,\cdot\rangle$ is the inner product. The WER is then classically computed on the n most discriminative words (weight of 1 for each word).

If we firstly compare the different configurations (TRS/ASR), we can note that the higher the theme classification accuracy is (table 1), the lower the WER is. This can be observed on both methods. More, we can see that the WER obtained with the LDA approach is slightly lower than the one obtained with the TF-IDF-Gini method, no matter the configuration considered. This means that a better transcription accuracy is associated to the discriminative words extracted with the LDA approach in comparison to the one obtained with the TF-IDF-Gini method, which could explain the higher classification performance reached by the LDA-based configuration.

5. Conclusions

In this paper, we presented an architecture to identify conversation themes from highly imperfect transcriptions using two different conversation representations coupled with a SVM classification step. We shown that the proposed topic representation using a LDA-based method outperforms the classification results obtained by the classical TF-IDF-Gini approach. The classification accuracy reaches 86.6% on manual transcriptions and 81.4% on automatic transcriptions with a respective gain of 6.9 and 10.9 points.

We also discussed the possible link between classification performance and transcription accuracy. The proposed analysis showed that the best classification results are obtained on configurations which extract the discriminative words having a lower Word Error Rate. The promising observations will lead to a more detailed qualitative study in a future work. Indeed, this preliminary study could be greatly extended with new analysis by taking into account, for example, the discriminative word weights in the transcription accuracy evaluation. A general perspective would be to propose a solution to estimate the classification performance depending on the transcription accuracy. In the context of evaluation metrics, it would also be interesting to find another way to estimate the accuracy of automatic transcriptions in the context of a specific task since the classical WER is not a good indicator of transcription quality in an applicative context.

6. Acknowledgements

This work was funded by the SUMACC and DECODA projects supported by the French National Research Agency (ANR) under contract ANR-10-CORD-007 and ANR-09-CORD-005.

7. References

R. Arun, R. Saradha, V. Suresh, M. Narasimha Murty, and C. E. Veni Madhavan. 2009. Stopwords and Stylometry : A Latent Dirichlet Allocation Approach. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada.

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

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

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

B.E. Boser, I.M. Guyon, and V.N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *5th annual workshop on Computational learning theory*, pages 144–152.

C.-C. Chang and C.-J. Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

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

T. Dong, W. Shang, and H. Zhu. 2011. An improved algorithm of bayesian text categorization. *Journal of Software*, 6(9):1837–1843.

M. Georgescul, A. Clark, and S. Armstrong. 2006. Word distributions for thematic segmentation in a support vector machine approach. In *Conference on Computational Natural Language Learning*, pages 101–108.

T. Griffiths and M. Steyvers. 2002. A probabilistic approach to semantic representation. In *24th annual conference of the cognitive science society*, pages 381–386. Citeseer.

T.J. Hazen. 2011. Topic identification. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 319–356.

T. 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.

Jian hua Yeh and Chun hsing Chen. 2010. Protein remote homology detection based on latent topic vector model. In *International Conference on Networking and Information Technology (ICNIT)*, pages 456–460.

T. Joachims. 1999. Transductive inference for text classification using support vector machines. In *Machine*

learning-international workshop then conference, pages 200–209. Morgan Kaufmann Publishers, Inc.

S. Kim, S. Narayanan, and S. Sundaram. 2009. Acoustic topic model for audio information retrieval. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 37–40.

M. Lan, C.-L. Tan, H.-B. Low, and S.-Y. Sung. 2005. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *International Conference on World Wide Web*, pages 1032–1033.

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

A.K. McCallum. 2002. Mallet: A machine learning for language toolkit.

I.D. Melamed and M. Gilbert. 2011. Speech analytics. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 397–416.

M. Morchid, R. Dufour, and G. Linarès. 2013a. Event detection from image hosting services by slightly-supervised multi-span context models. In *CBMI*. IEEE.

M. Morchid, G. Linarès, M. El-Beze, and R. De Mori. 2013b. Theme identification in telephone service conversations using quaternions of speech features. In *INTERSPEECH*. ISCA.

K. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. 1997. Predicting time series with support vector machines. *ICANN'97*, pages 999–1004.

M. Purver. 2011. Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317.

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

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

J. I. Sheeba and K. Vivekanandan. 2012. Article: Improved keyword and keyphrase extraction from meeting transcripts. *International Journal of Computer Applications*, 52(13):11–15.

S. Tang, J. Li, Y. Zhang, C. Xie, M. Li, Y. Liu, X. Hua, Y.-T. Zheng, J. Tang, and T.-S. Chua. 2009. Pornprobe: an lda-svm based pornography detection system. In *International Conference on Multimedia*, pages 1003–1004.

G. Tur and R. De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. Wiley.

V. Vapnik. 1963. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

G-X Yuan, C-H Ho, and C-J Lin. 2012. Recent advances of large-scale linear classification. 100(9):2584–2603.

M. Zrigui, R. Ayadi, M. Mars, and M. Maraoui. 2012. Arabic text classification framework based on latent dirichlet allocation. *CIT*, 20(2):125–140.