# Integration of word and semantic features for theme identification in telephone conversations

Yannick Estève, Mohamed Bouallegue*, Carole Lailler†, Mohamed Morchid, Richard Dufour, Georges Linarès, Driss Matrouf, Renato De Mori

**Abstract** The paper describes a research about the possibility of integrating different types of word and semantic features for automatically identifying themes of real-life telephone conversations in a customer care service. Features are all the words of the application vocabulary, the probabilities obtained with Latent Dirichelet Allocation (LDA) of selected discriminative words and semantic features obtained with a limited human supervision of words and patterns expressing entities and relations of the application ontology. A Deep Neural Network (DNN) is proposed for integrating these features. Experimental results on manual and automatic conversation transcriptions are presented showing the effective contribution of the integration. The results show how to automatically select a large subset of the test corpus with high precision and recall, making it possible to automatically obtain theme mention proportions in different time periods.

**Key words:** Theme identification, human-human spoken conversation, deep neural network

Yannick Estève, Mohamed Bouallegue, and Carole Lailler
LIUM, University of Le Mans, France, e-mail: firstname.lastname@lium.univ-lemans.fr

Mohamed Morchid, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato De Mori
LIA, University of Avignon, France, e-mail: firstname.lastname@univ-avignon.fr

Renato De Mori
McGill University, Montreal, Québec, Canada, e-mail: rdemori@cs.mcgill.ca

* Thanks to the ANR agency for funding through the CHIST-ERA ERA-Net JOKER project.

† Thanks to European Commission for funding through the EUMSSI Project, number 611057, call FP7-ICT-2013-10.

# 1 Introduction

A growing research interest has been observed in the automatic analysis of human/human spoken conversations as reviewed in [1] and [2]. A scientifically interesting and practically important component of this research is topic identification for which an ample review of the state of the art can be found in [3]. In spite of the relevant progress achieved so far, it is difficult to reliably identify multiple topics in real-life telephone conversations between casual speakers in unpredictable acoustic environments. Of particular interest are call-center conversations in which customers discuss problems in specific domains with an advisor. This is the case of the application considered in this paper. The purpose of the application is to collect statistics about the problems discussed in the customer care service (ccs) of the ratp Paris transportation system. Statistics are obtained by analyzing real-world human/human telephone conversations in which an agent attempts to solve a customer problem. Proportions of problem themes are used for monitoring user establishing priorities of problem solving interventions. Application relevant information for the task is described in the application requirements. Themes are the most general entitles of the application ontology outlined in the documentation. Agents follow a pre-defined protocol to propose solutions to user problems about the transportation system and its services. An automatic classification of conversation themes is motivated by the fact that, due to time constraints, agents cannot adequately take note of the discussed themes. A fully automatic system for theme identification must include an automatic speech recognition (asr) module for obtaining automatic transcriptions of the conversations. The acoustic environment on these conversations is unpredictable with a large variety of noise types and intensity. Customers may not be native French speakers and conversations may exhibit frequent disfluencies. The agent may call another service for gathering information. This may cause the introduction of different types of non-speech sounds that have to be identified and discarded. For all these reasons, the word error rate (wer) of the asr system is highly variable and can be very high.

Popular features for topic identification are reviewed in [3]. Concise representations of document contents has been proposed using features obtained with latent semantic analysis (LSA) [4], probabilistic latent semantic analysis (pLSA) (Li) and latent Dirichelet Allocation (LDA) [5]. Among them, LDA features provide rich representations in latent spaces with a limited number of dimensions. Recently, in [6] a detailed analysis has been reported in terms of theme classification accuracy in spoken conversations by varying the word vocabulary size and the number of hidden topics. This suggested performing a detailed analysis of classification accuracy by fine-grained variations of the number of hidden topics [7]) and the value $\alpha$ of the LDA hyperparameter [8]. As a large variation of the classification accuracy was observed, it was proposed to compose each feature set obtained with a specific hidden space size and value of $\alpha$ in a single vector called c-vector. The best improvement over other types of features was observed for a c-vector whose elements are unigram probabilities of a limited set of discriminative words.

In the application considered in this paper, the conversions to be analyzed are made available by relatively small sets collected in different time periods. Relevant discriminative words may change in time even if most of them belong to the same semantic category defined in the application ontology. An example is a bus line whose itinerary has been temporary modified. The names of some streets are likely to be frequently mentioned in that time period by customers inquiring about the itinerary. These specific names are unlikely to have been selected as be discriminative words even if their co-presence with other words is very useful to characterize a traffic state. A new approach is proposed in this paper to embed different types of latent features, some of them representing words of the entire vocabulary as proposed in [9] for call routing, some other representing expressions of fragments of the application ontology and some others being LDA features.

## 2 Features used for theme identification

The corpus of the application considered in this paper is annotated with 8 conversation themes: *problems of itinerary*, *lost and found*, *time schedules*, *transportation cards*, *state of the traffic*, *fares*, *infractions*, and *special offers*. Three types of features are considered for theme identification. They are the words of the application vocabulary $V_W$, a set S of labels for application concepts and a conversation summary represented by a c-vector whose elements are unigram probabilities of words belonging to a reduced vocabulary $V_S \subset V$ of theme discriminative words.

### *2.1 Word features*

All the words of the application vocabulary $V_W$ are considered as features for theme identification. For each conversation, a vector W of binary values is built in which each element correspond to a word in $V_W$ and its value is set to 1 only if the corresponding word is present in the conversation

A discriminative word vocabulary $V_D \subset V_W$ is formed as described in [6] with the top 116 words of $V_W$ ranked with the product of Term Frequency (TF), Inverse Document Frequency (IDF), and word purify in the themes. Unigram probabilities of discriminative words are computed in an r-dimensional hidden space using LDA as follows:

$$P_r(w_i|d) = \sum_{n=1}^{N_r} P(w_i|z_n^r)P(z_n^r|d) \tag{1}$$

where $N_r$ is the number of hidden spaces, $P(w_i|z_n^r)$ is the probability of word $w_i$ in the n-th hidden topic of the r-th hidden space and $P(z_n^r|d)$ is the probability of the n-th hidden topic in the d-th conversation.

Let $x^r(d) = P_r(w|d)$ be the vector having probabilities $P_r(w_i|d)$ as elements. All vectors $x^r(d)$ are then integrated into a unique C *vector* obtained with Joint Factor Analysis as described in [7].

## *2.2 Semantic features*

Conversations of the train set labelled with only one theme are selected. Using them, a unigram language model (LM) is obtained for each theme and for the ensemble of all the conversations.

Let $\Im_k$ be the set of conversations of theme $\tau_k$. For each conversation theme, a set of words is selected using the approach described in [10].

Let $P_k(w)$ represent the LM probability distribution estimated with the data in $\Im_k$ and $P_g(w)$ represent the probability distribution estimated with the data in $\Im_g = \cup_{k=1}^{K} \Im_k$, where $K$ is the number of conversation themes.

The two distributions $P_k(w)$ and $P_g(w)$ diverge and a measure of their divergence is the *Kullback-Leibler* divergence (KLD) measure:

$$KLD[P_k(w), P_g(w)] = \sum_{w \in \Im_g} P_k(w) \log \frac{P_k(w)}{P_g(w)} \tag{2}$$

It has been shown in [10] that, when comparing word unigram distributions, the addends that mostly contribute with a positive value to the summation in $KLD[P_k(w), P_g(w)]$ are useful features for performing relevance feed-back in information retrieval. The same approach is applied to the train set for making a list of words for each conversation theme. Another application of an approach of this type can be found in [11]. A human expert analyses the words of each list starting from the top of the list. Words that express facts and other concepts of the application ontology are selected and labelled with concepts of the application ontology. Generalizations are performed by associating the same concept to words not observed in the train set but belonging to the same class record of the application database.

Let $V_S$ be the vocabulary of these concept labels. For each conversation, a vector S of binary values is built in which each element correspond to a concept in $V_S$ and its value is set to 1 only if the corresponding concept is present in the conversation. Features embedding the co-presence of word and concept features are automatically learned in the approach proposed in this paper. For example, the co-presence of a location concept with at least two different locations words is expected to be a useful feature for the theme itinerary. Other dependencies expressing the co-presence of words and concepts are expected to be useful expressions of application relevant distant semantic relations. Simple patterns involving words and concepts expressing local semantic relations are also manually derived with a minor human effort. The corresponding meanings are also represented by elements of vector S.

## 3 A Deep Neural Network architecture for theme identification

A Deep Neural Network (DNN) architecture is proposed for integrating word features represented by vector W, semantic features represented by vector S and conversation summaries represented by the C vector.

This architecture should be simple enough to allow effective training of its parameters to obtain features that correspond to requirements inspired by the application ontology. The most important requirement is to capture sufficient concept relations between concepts characterizing each theme. The second requirement is to ensure coherence between specific dependencies expressed by hidden features and a concise global representation of a conversation expressed by LDA features.

In order to capture some dependencies between words, concepts expressed by words and semantic entities expressed by short distance patterns vectors W and S are concatenated and encoded into a vector $X = SW$ by the equation:

$$h_1(X) = f(U \times X + b_1) \tag{3}$$

The elements of matrix U and vector $b_1$ are estimated by a multilayer perceptron having vector $X$ as input, vector $h_1(X)$ computed by the hidden layer and a vector $V_K$ of $K = 8$ output nodes corresponding to the 8 themes of the application.

The vector $h_1(X)$ is then concatenated with an embedding:

$$h_2(C) = f(B \times C + b_2) \tag{4}$$

of vector C to obtain a vector:

$$Q = h_1(X)h_2(C) \tag{5}$$

A vector $h_3(Q)$ is obtained at a third hidden layer with the relation:

$$h_3(Q) = f(G \times Q + b_3) \tag{6}$$

The DNN output values are computed as

$$V_k = f(L \times h_3(Q) + b_4) \tag{7}$$

Eventually only the elements of matrices B and Q are estimated with back propagation while the values of U are kept fixed. The reason is that in this way the number of parameters to be estimated is kept relatively small to avoid over fitting due to the limited size of the train set.

The resulting DNN architecture is depicted in Figure1.

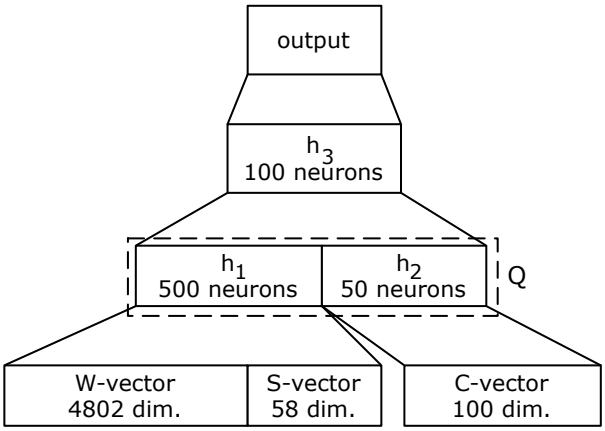


**Fig. 1** DNN architecture for theme identification.

## 4 Experimental set up

The corpus of the DECODA project [12] has been used for the theme identification experiments described in this section. This corpus is composed of 1,067 telephone conversations from the call centre of the public transportation service in Paris. The corpus is split into a train set (740 dialogues) and a test set (327 dialogues). Conversations have been manually transcribed and labeled with one theme label corresponding to the principal concern mentioned by the customer. A portion of the train set (175 dialogues) is also used as a development set for selecting the dimension of the hidden topic spaces. All hidden spaces were obtained with the manual transcriptions of the train set. The number of turns in a conversation and the number of words in a turn are highly variable. The majority of the conversations have more than ten turns. The turns of the customer tend to be longer ($>$ 20 words) than those of the agent and are more likely to contain out of vocabulary words that are often irrelevant for the task.

The ASR system used for the experiment is the LIA-speeral system [13] with 230000 Gaussians in the triphone acoustic models. Model parameters were estimated with maximum a-posteriori probability (MAP) adaptation of 150 hours of speech in telephone bandwidth with the data of the train set. The vocabulary contains 5782 words. A 3-gram language model (LM) was obtained by adapting with the transcriptions of the train set a basic LM. An initial set of experiments were performed with this system resulting with an overall WER on the test set of 58% (53% for agents and 63% for users). These high error rates are mainly due to speech disfluencies and to adverse acoustic environments for some dialogues when, for example, users are calling from train stations or noisy streets with mobile phones. Furthermore, the signal of some sentences is saturated or of low intensity due to the distance between speakers and phones.

Experiments were performed with different types of inputs and components of the network whose scheme is shown in Figure 1. For the sake of comparison, a set of experiments was performed with the manual transcriptions (TRS) of the conversations using simple multi-layer perceptrons (MLP) with one hidden layer, an output layer with K nodes and fed by different input vectors. The results for different architectures are reported in Table 1 for the development (DEV) and the test (TEST) sets. The confidence interval is +- 3.69.

| Input | Train/test corpus | DEV | TEST |
|---|---|---|---|
| W | TRS/TRS | 86.9 | 83.2 |
| S | TRS/TRS | 82.9 | 78.9 |
| W+S | TRS/TRS | 89.7 | 85.9 |

**Table 1** Percent accuracies for MLP architectures, for the development (DEV) and the test (TEST) sets. The data of the train and test sets are manual transcriptions (TRS)

The results of Table 1 show a superiority by using as input the word vector W rather than the vector S of semantic labels and a further improvement by concatenating vectors W and S at the input.

The same type of experiment was performed using ASR transcriptions (ASR). The results are reported in Table 2. An improvement is again observed by concatenating W with S at the input of an MLP network. Nonetheless, the improvement is inside the confidence interval, suggesting to consider additional input features in the architecture with the structure shown in Figure 1, indicated as DNN in Table 2.

The results are promising, but more interesting is the Precision  Recall relation shown in Figure 2. It has been obtained by selecting conversations based on the posterior probability $P_{k1}(d)$ of the theme $t_1$ ranked first for conversation $d$. Posterior probabilities of theme hypotheses for conversation $d$ are computed with the *softmax* function applied to the outputs of DNN fed by features of $d$. The curve shows that a precision of 90% can be achieved with 84% recall, making it possible to obtain practically useful conversation survey proportions with a small rejection of samples that could be manually annotated with a limited effort. These results compare favourably with the best results obtained so far with the same corpus ([14]) where the 90% precision was obtained with 78% recall.

| Input | Train/test corpus | Architecture | TEST |
|---|---|---|---|
| W | ASR/ASR | MLP | 79.5 |
| W+S | ASR/ASR | MLP | 82.3 |
| W+S+C | ASR/ASR | DNN | 82.9 |

**Table 2** Percent accuracies for different architectures for the test (TEST) set. The data of the train and test sets are ASR transcriptions (ASR).

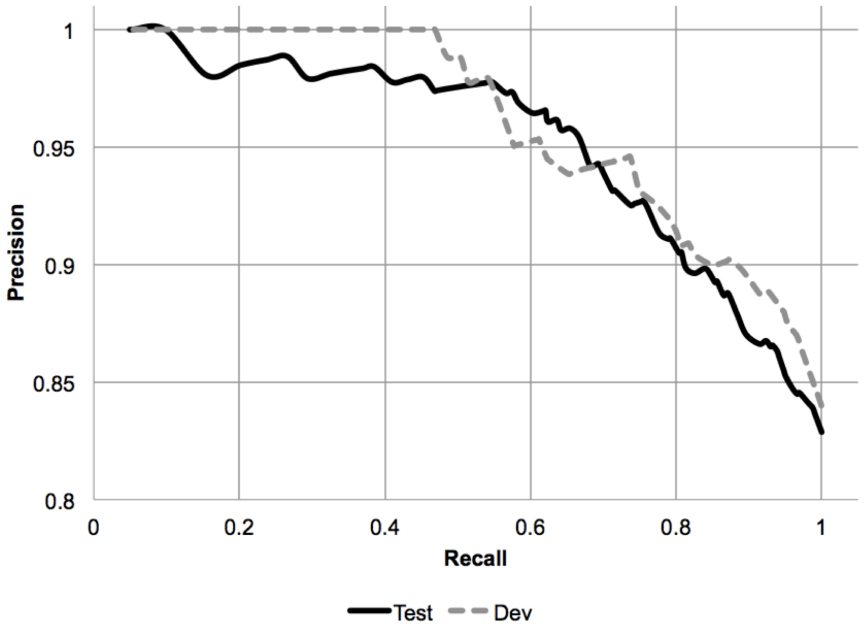


**Fig. 2** Precision - Recall results for the test set.

## 5 Conclusion

A DNN architecture has been proposed for theme identification in human/human conversations by integrating different word and semantic features. With the proposed network high precisions can be obtained by rejecting a small proportion of conversations classified with high equivocation. As some of the semantic features used for themes identification are descriptions of basic facts characterizing a theme, it will be possible in future work to automatically verify the mentions of basic facts for an automatically identified theme. An absence of fact mention is a clue for selecting the conversation as an informative example to be manually analysed for discovering and generalizing possible new mentions of basic facts for one or more themes discussed in the selected conversation.

## References

1. Gokhan Tur and Renato De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.

2. Gokhan Tur and Dilek Hakkani-Tür, "Human/human conversation understanding," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 225–255, 2011.
3. Timothy J Hazen, "MCE training techniques for topic identification of spoken audio documents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2451–2460, 2011.
4. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.
5. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
6. Mohamed Morchid, Richard Dufour, Pierre-Michel Bousquet, Mohamed Bouallegue, Georges Linarès, and Renato De Mori, "Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule," in *ICASSP*, 2014.
7. Mohamed Morchid, Mohamed Bouallegue, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato De Mori, "An i-vector based approach to compact multi-granularity topic spaces representation of textual documents," in *the Conference of Empirical Methods on Natural Lnguage Processing (EMNLP) 2014*. SIGDAT, 2014.
8. Mohamed Morchid, Mohamed Bouallegue, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato De Mori, "I-vector based representation of highly imperfect automatic transcriptions," in *Conference of the International Speech Communication Association (INTERSPEECH) 2014*. ISCA, 2014.
9. Ruhi Sarikaya, Geoffrey E. Hinton, and Anoop Deoras, "Application of deep belief networks for natural language understanding," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 22, no. 4, pp. 778–784, 2014.
10. Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi, "An information-theoretic approach to automatic query expansion," *ACM Trans. Inf. Syst.*, vol. 19, no. 1, pp. 1–27, 2001.
11. M.S. Wu, H.S. Lee, and H.M. Wang, "Exploiting semantic associative information in topic modeling," in *SLT Workshop, 2010*. IEEE, 2010, pp. 384–388.
12. Frederic Béchet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Bèze, Renato De Mori, and Eric Arbillot, "Decoda: a call-centre human-human spoken conversation corpus," in *LREC'12*, 2012.
13. Georges Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The LIA speech recognition system: from 10xRT to 1xRT," in *Proceedings of the 10th international conference on Text, speech and dialogue*. Springer-Verlag, 2007, pp. 302–308.
14. Mohamed Morchid, Richard Dufour, Mohamed Bouallegue, Georges Linarès, and Renato De Mori, "Theme identification in human-human conversations with features from specific speaker type hidden spaces," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.