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

Laboratoire Informatique d'Avignon
University of Avignon
France

Mohamed Morchid, Mohamed Bouallegue, Richard Dufour,
Georges Linarès, Driss Matrouf and Renato De Mori, Fellow, IEEE
firstname.lastname@univ-avignon.fr

SUPPORTED BY ANR
Research supported under contract ANR-10-CORD-007 and ANR-09-CORD-005

Context

Conversations agent/customer customer care service of the Paris transportation system

Agent: Hello
Customer: Hello
Agent: Speaking ...
Customer: I call you because I was fined today, but I still have an imagine card suitable for zone 1 [...] I forgot to use my navigo card for zone 2
Agent: You did not use your navigo card, that is why they give you a fine not for a zone issue [...]
Customer: Thanks, bye
Agent: bye

Difficulties

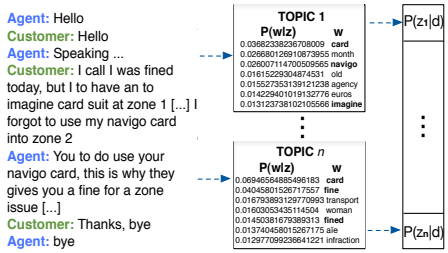
- A dialogue may contain **a major and other semantically related themes** (e.g: transportation card, lost and found, infractions)
- **Transcriptions** obtained from an Automatic Speech Recognition (**ASR**) system are **error prone** ⇨ **Use astract features** (latent Dirichlet allocation)
- Abstract representation involves **selecting** the **right number** of **classes** composing the topic space

Solution

- 2 **drawbacks** by using a **compact** representation from **multiple** topic spaces
- But, multi-view representation introduces:
  - **relevant** variability needed to represent different contexts of the document
  - **noisy** variability related to topic space mapping
- **Reduction** of the **noisy variability** with **factor analysis** technique
- **Extraction** of a **compact** representation containing **useful** information named *i*-vector

Theme Identification Methods

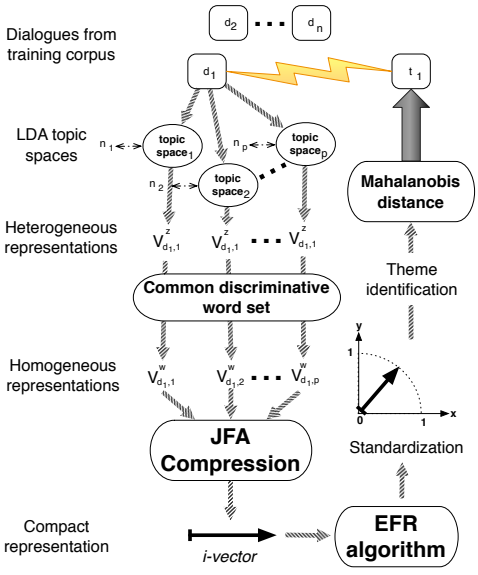
Semantic representation with a Latent Dirichlet Allocation (LDA) model



Gaussian based Bayes classification

Based on two simple assumptions :
- distributions of theme classes are Gaussian
- the covariances of these classes are equal

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



Experiments and Results

Experimental protocol

- **DECODA project corpus** of conversations:
  - Train = 740 / Dev=175 / Test = 327
- Automatic Speech Recognition (**ASR**) system: **Speeral**
  - Word Error Rate (**WER**) with stop-list of 126 words:
    - train = **33.8%** / dev = **45.2%** / test = **49.5%**
- **8 conversation themes**
  TRS: manual transcriptions - **ASR:** automatic transcriptions
- **3,000 hidden topic spaces** with a **different topic number** was built using the train corpus

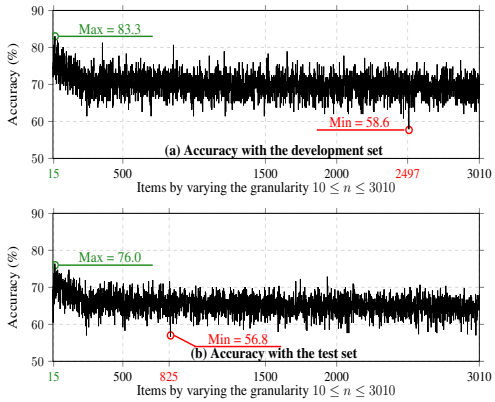
Theme hypothesization accuracies using different c-vectors and GMM-UBM sizes

c-vector size	DEV				TEST			
	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

→ Classification **performance** is **stable** (**5.9 points** difference for dev)

→ Using **comparable training and testing** configurations allows to achieve the best classification performance

Theme Identification using various topic-based representations



→ Best reached **classification** accuracy: 83% dev and 76% test

→ Classification **performance** is **unstable** (Difference of **25 points** on the dev.)

Conclusion

- In spite of very high WER, possible to classify effectively documents with the proposed compact representation (*c*-vector) with an **accuracy of 85%** + allows us to both solve the difficult choice of the right number of topics and the theme proximity
- **Future work** will seek to find the best combination of LDA hyper-parameters and evaluate effectiveness in other NLP tasks