Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 68–85

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Impact of Word Error Rate on theme identification task of highly imperfect human–human conversations[☆]

Mohamed Morchid[*], Richard Dufour, Georges Linarès

Laboratoire Informatique d'Avignon (LIA) – University of Avignon, 339, chemin des Meinajaries, Agroparc,
BP 91228, 84911 Avignon Cedex 9, France

Received 2 February 2015; received in revised form 7 December 2015; accepted 22 December 2015
Available online 11 January 2016

Abstract

A review is proposed of the impact of word representations and classification methods in the task of theme identification of telephone conversation services having highly imperfect automatic transcriptions. We firstly compare two word-based representations using the classical Term Frequency-Inverse Document Frequency with Gini purity criteria (TF-IDF-Gini) method and the latent Dirichlet allocation (LDA) approach. We then introduce a classification method that takes advantage of the LDA topic space representation, highlighted as the best word representation. To do so, two assumptions about topic representation led us to choose a Gaussian Process (GP) based method. Its performance is compared with a classical Support Vector Machine (SVM) classification method. Experiments showed that the GP approach is a better solution to deal with the multiple theme complexity of a dialogue, no matter the conditions studied (manual or automatic transcriptions) (Morchid et al., 2014). In order to better understand results obtained using different word representation methods and classification approaches, we then discuss the impact of discriminative and non-discriminative words extracted by both word representations methods in terms of transcription accuracy (Morchid et al., 2014). Finally, we propose a novel study that evaluates the impact of the Word Error Rate (WER) in the LDA topic space learning process as well as during the theme identification task. This original qualitative study points out that selecting a small subset of words having the lowest WER (instead of using all the words) allows the system to better classify automatic transcriptions with an absolute gain of 0.9 point, in comparison to the best performance achieved on this dialogue classification task (precision of 83.3%).

Keywords: Speech analytics; Human–human dialogue; Latent Dirichlet allocation; Topic representation; Principal component analysis; Classification performance study

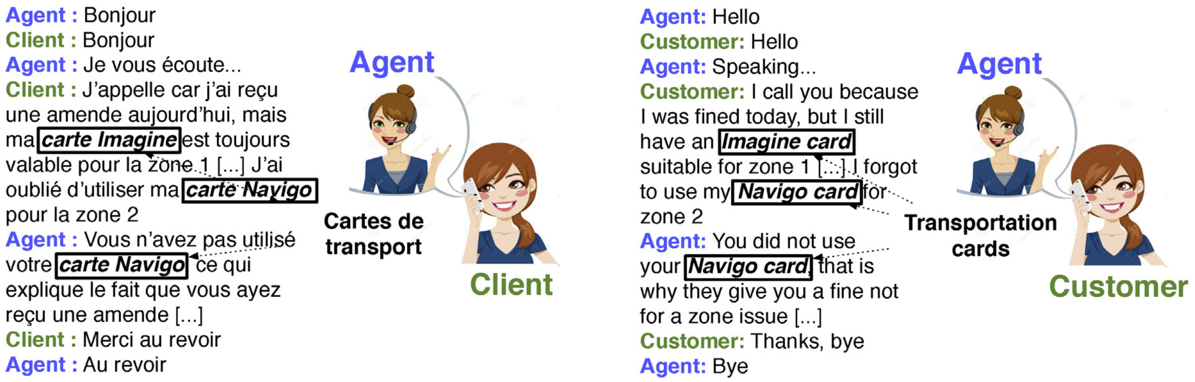
1. Introduction

Automatic Speech Recognition (ASR) systems frequently fail on noisy conditions and high Word Error Rates (WER) make difficult the analysis of the automatic transcriptions. 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

[☆] This paper has been recommended for acceptance by Alexandros Potamianos.

[*] Corresponding author. Tel.: +33 490843577.
E-mail addresses: mohamed.morchid@univ-avignon.fr (M. Morchid), Richard.Dufour@univ-avignon.fr (R. Dufour), Georges.Linares@univ-avignon.fr (G. Linarès).

http://dx.doi.org/10.1016/j.csl.2015.12.001



Agent : Bonjour
Client : Bonjour
Agent : Je vous écoute...
Client : J'appelle car j'ai reçu une amende aujourd'hui, mais ma ***carte Imagine*** est toujours valable pour la zone 1 [...] J'ai oublié d'utiliser ma ***carte Navigo*** pour la zone 2
Agent : Vous n'avez pas utilisé votre ***carte Navigo*** ce qui explique le fait que vous ayez reçu une amende [...]
Client : Merci au revoir
Agent : Au revoir

(a) Original dialogue (in French)

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

(b) Translated dialogue (in English)

Fig. 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*). The original dialogue conversation in French is presented (a) as well as its English translation (b).

paper proposes a global study to improve the robustness of speech analytics by first comparing word representations as well as classification methods and the impact of WER in topic space learning process, on the theme identification task (Bechet et al., 2012) in the application framework of the RATP call centre (Paris Public Transportation Authority).

Telephone conversation is a particular case of human–human interaction whose automatic processing encounters many difficulties, especially due to the speech recognition step required to obtain the transcription of the speech contents. First, the speaker behavior may be unexpected and the training/test mismatch may be very large. Second, speech signal may be strongly impacted by various sources of variability: environment and channel noises, acquisition devices, etc.

Themes are related to the reason why the customer called. Various classes corresponding to the main customer requests are considered (*lost and founds, traffic state, timelines*, etc). In addition to classical problems in such adverse conditions, the topic-identification system should face issues to classes proximity. For example, a *lost and 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. Fig. 1 shows an example of a dialogue manually labeled by the agent as an issue related to an *infraction*. However, words in bold suggest that this conversation could also be related to a *transportation card* issue.

Agents then annotate a conversation with what they consider the major theme of the customer request: as a result, a single theme is associated for each conversation.

In the context of Information Retrieval (IR) tasks, the main feature used is the *term frequency* that allows to obtain a subset of discriminative[1] words for a considered class. This set of discriminative words should permit to compose a vector representation of conversation themes in the semantic space. Its application to automatic transcriptions is more difficult since transcription errors would lead to an incorrect word representation. Thereby, we assume that dialogues have to be considered in an intermediate thematic representation to fully perform this multiple themes' complexity. For this reason, the projection of the automatically transcribed words in a more abstracted space could increase the robustness to the Automatic Speech Recognition (ASR) errors.

Thus, we propose to first explore a term frequency representation, with the TF-IDF-Gini method, and a topic space representation, with a latent Dirichlet allocation (LDA) approach (Blei et al., 2003), coupled with a classification method to automatically identify themes from highly imperfect automatic transcriptions. The other main issue is the choice of the best classification method that does not modify the dialogue topic representation.

In the second part of this paper, the classical SVM method (Yuan et al., 2012), that modifies the word representation with a kernel function, is compared with a Naive Bayesian classifier, that does not modify it. We assume that this study will highlight the fact that these two assumptions are relevant: the Gaussianity of the theme classes and the equality of the class covariances.

[1] The term "discriminative" is associated to a word if it permits to discern a class from the others.