Dialogue Act Recognition using Visual Information

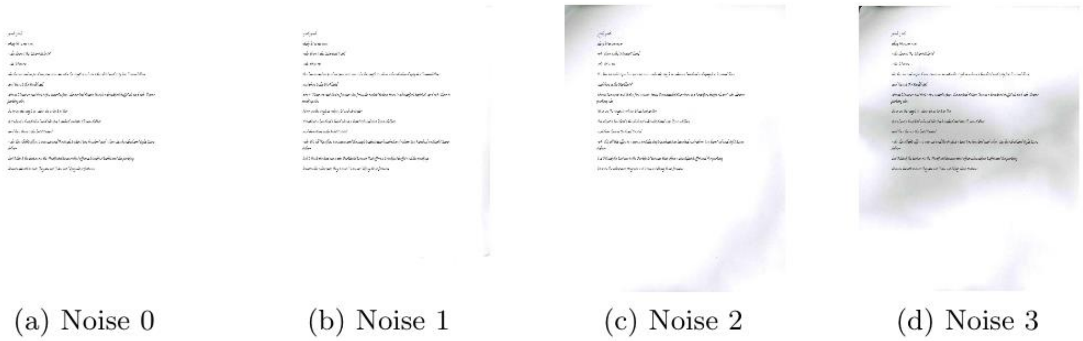
Jiří Martínek [1], Pavel Král[2], Ladislav Lenc[2]

1 Introduction

The goal of dialogue act (DA) recognition is to perform the spoken utterance classification (questions, answers, offers, commits, greetings, etc.). Dialogues and their understanding are most often related to the processing of an audio (speech) signal. However, this task might be performed on printed documents (e.g. comic books) as well. We focus on the DA recognition from raster images. In such a case, transcriptions aren't available and to obtain the text from the image, we utilize optical character recognition (OCR). Errors may occur in the resulting text leading to poor classification performance. To balance this loss of information, we employ a neural network model to perform an utterance classification with both text and visual features. Our experiments have demonstrated that visual information can significantly improve the overall classification score, especially when used on low-quality images with erroneous OCR.

2 Dataset

Our dialogue source is the VERBMOBIL dataset (Alexandersson et al. (1998)). It contains English and German dialogues in text files and there are 16 different classes. We transformed these text files into an image dataset by rendering the dialogues to the blank A4 pages. We used four different backgrounds with varying levels of noise (see **Figure 1**).



(a) Noise 0 (b) Noise 1 (c) Noise 2 (d) Noise 3

Figure 1: Page example from all four noisy backgrounds

Furthermore, we created a copy of all four datasets while conducting random transformations (blurring and tiny rotations) to simulate the real situation where such images are produced by a scanner. So we have created 8 datasets in total labeled as: *0, 1, 2, 3 and 0_trans, 1_trans, 2_trans, 3_trans*. We utilized a simple algorithm to segment individual utterances from the whole dialogue and cropped text-line images.

[1] Ph.D student, University of West Bohemia, Faculty of Applied Sciences, Dept. of Computer Science & Engineering, specialization: Informatics, machine learning and neural networks, e-mail: jimar@kiv.zcu.cz

[2] University of West Bohemia, Faculty of Applied Sciences, Dept. of Computer Science & Engineering, e-mail: pkral@kiv.zcu.cz, llenc@kiv.zcu.cz

3 Models

We proposed three neural network models to address three different scenarios:

1. **Text input only**
2. **Visual input only**
3. **Combined inputs** – text-line images and their predicted transcription by OCR

The main building blocks of the models are recurrent neural network – LSTM (Hochreiter and Schmidhuber (1997)) and convolutional neural network.

4 Experiments

The performance of each model of particular datasets is depicted in **Figure 2.**

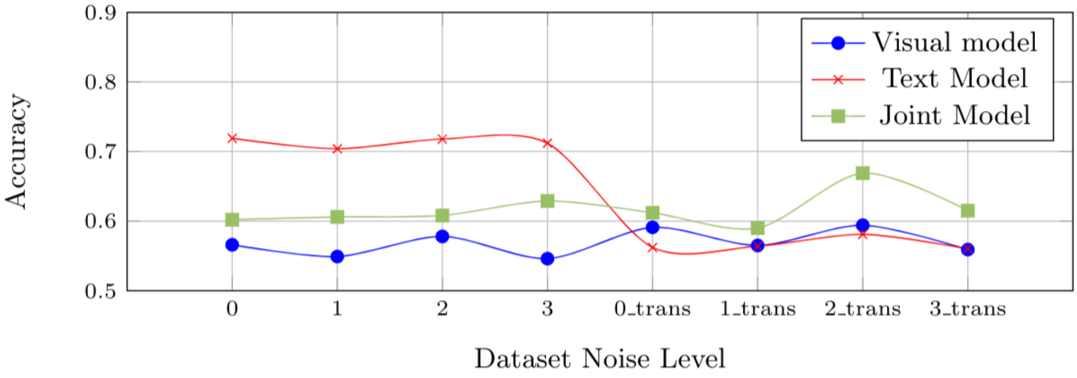


Figure 2: Depicted results and comparison of all models

To compare our results with other approaches, we employed the text model results since there is no prior work based on the visual features in DA recognition. The best-published accuracies based solely on text features are in range: 74% – 75%. On the basis of the observation, we can state that with the increasing noise level the performance of the text model significantly decreases (primarily due to the poor OCR predictions). The joint model performance is relatively stable and its power is demonstrated when used on difficult datasets where the best results have been obtained.

Acknowledgements

This work has been partly supported by Cross-border Cooperation Program Czech Republic - Free State of Bavaria ETS Objective 2014-2020 (project no. 211) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications.

References

Martínek, J, Král, P., Lenc, L., Cerisara, C. (2019) Multi-lingual dialogue act recognition with deep learning methods. arXiv preprint arXiv:1904.05606.

Hochreiter S., Schmidhuber J. (1997), Long short-term memory, Neural computation,vol. 9, no. 8, pp. 1735–1780.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., Siegel, M., (1998), Dialogue acts in Verbmobil 2 . DFKI Saarbrücken