Dialogue Act Recognition using Visual Information

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

[1] Dept. of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society University of West Bohemia Plzeň, Czech Republic
{jimar,pkral,llenc}@kiv.zcu.cz

Abstract. Automatic dialogue management including dialogue act (DA) recognition is usually focused on dialogues in the audio signal. However, some dialogues are also available in a written form and their automatic analysis is also very important.
The main goal of this paper thus consists in the dialogue act recognition from printed documents. For visual DA recognition, we propose a novel deep model that combines two recurrent neural networks.
The approach is evaluated on a newly created dataset containing printed dialogues from the English VERBMOBIL corpus. We have shown that visual information does not have any positive impact on DA recognition using good quality images where the OCR result is excellent. We have also demonstrated that visual information can significantly improve the DA recognition score on low-quality images with erroneous OCR.
To the best of our knowledge, this is the first attempt focused on DA recognition from visual data.

Keywords: Dialogue Act Recognition · Multi-modal · OCR · RNN · Visual Information

1 Introduction

Dialogue Act (DA) recognition is a task to segment a dialogue into sentences (or their parts) and to assign them appropriate labels depending on their function in the dialogue [1]. These labels are defined by several taxonomies [2] (e.g. questions, commands, backchannels, etc).

The standard input is a speech signal which is usually converted into textual representation using an Automatic Speech Recognition (ASR) system [3]. The combination of the following information sources is often considered for recognition: lexical (words in the sentence), prosodic (sentence intonation), and the dialogue history (sequence of the DAs) [4].

However, dialogues are also available in a written form (books and comics), and their automatic analysis is also beneficial for further text analysis. Hence, the

main goal of this paper is the DA recognition from the documents in a printed form.

Similarly, as in the DA recognition from the audio signal, we first convert the images into a lexical representation using Optical Character Recognition (OCR) methods. We assume that the image form (as the speech signal) can contain some additional information.

Therefore, the main contribution of this paper lies in the usage of visual information for automatic DA recognition from printed dialogues. To the best of our knowledge, there is no prior work that focuses on the DA recognition from printed / handwritten documents.

For evaluation, we create a novel image-based DA recognition dataset from written dialogues. This corpus is based on the dialogues from the `VERBMOBIL` corpus [5] and the scripts for the creation of such a dataset are available online. These scripts represent another contribution of this work.

We further assume that with the decreasing quality of the printed documents, the importance of the visual text representation will play a more important role for DA recognition, since a recognized text contains greater amount of OCR errors. We will also evaluate this hypothesis using four different image quality in the corpus.

For visual DA recognition, we propose a deep neural network model that combines Convolutional Recurrent Neural Network (CRNN) and Recurrent Neural Network (RNN). We utilize the Bidirectional Long Short-term Memory (BiLSTM) as a recurrent layer in both architectures.

2 Related Work

This section first briefly outlines the DA recognition field and presents popular datasets. Then, we describe recent multi-modal methods that use text and image inputs to improve the performance of a particular task.

Usually, the research in the DA recognition field is evaluated on monolingual standard datasets such as Switchboard (SwDA) [6], Meeting Recorder Dialogue Act (MRDA) [7] or DIHANA [8]. Colombo et al. [9] proposed a *seq2seq* deep learning model with the attention and achieved excellent results that are comparable or even better than current state-of-the-art results.

Shang et al. [10] presented experiments with a deep (BiLSTM-CRF) architecture with an additional extra input representing speaker-change information. The evaluation was conducted on SwDA dataset.

The `VERBMOBIL` Dialogue Acts corpus [5, 11] has been used in the past as a representative of the multi-lingual corpus (see e.g. Reithinger and Klesen [12], Samuel et al. [13] or Martínek et al. in [14]).

Recently, experiments on joining DA recognition and some other Natural Language Processing (NLP) tasks have begun to emerge. Cerisara et al. in [15] presented a multi-task hierarchical recurrent network on joint sentiment and dialogue act recognition. A multi-task recognition which is related to the DA recognition is presented by Li et al. [16]. They utilize the *DiaBERT* model for

DA recognition and sentiment classification and evaluate their approach on two benchmark datasets.

There are efforts to join the visual information with text and improve to some extent text-based NLP tasks. Zhang et.al [17] investigated Named Entity Recognition (NER) in tweets containing also images. They showed that visual information is valuable in the name entity recognition task because some entity word may refer directly to the image included.

Audebert et al. [18] present a combination of image and text features for document classification. They utilize Tesseract OCR together with *FastText* [19] to create character-based embeddings and, in the sequel, the whole document vector representation. For the extraction of image features, they use the *MobileNetv2* [20]. The final classification approach combines both features.

A very nice approach for multi-modal document image classification has been presented by Jain and Wigington in [21]. Their fusion of visual features and semantic information improved the classification of document images.

3 Model Architectures

We describe gradually three models we use for the DA recognition. First of all, we present the visual model that we use for DA recognition based only on image features. Next, we describe our text model and, finally, the joint model that combines both image and text inputs.

3.1 Visual Model

The key component in this model is the Convolutional Recurrent Neural Network (CRNN) that has been successfully utilized for OCR (e.g. [22, 23]) and also for image classification [24].

For the visual DA recognition, the input is the image of an entire page of a dialogue where each text line represents an utterance. This page is processed by the *Utterance Segmentation* module that produces segmented images of text lines. These images are fed into the CRNN that maps each utterance to the predicted label. The scheme of this approach is depicted in Figure 1.

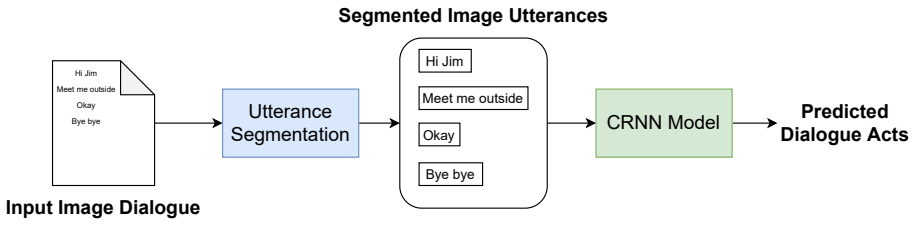


Fig. 1: Visual DA recognition model

Convolutional layers within CRNN create feature maps with relation to the specific receptive fields in the input image. Due to the pooling layers, dimensionality is reduced, and significant image features are extracted, which are further processed by recurrent layers. The recurrent layers are fed by feature sequences (the feature vectors in particular frames in the image).

The CRNN model is depicted in Figure 2 in two forms: the original model proposed by Shi et al. [22] for OCR and our adapted version for DA recognition.

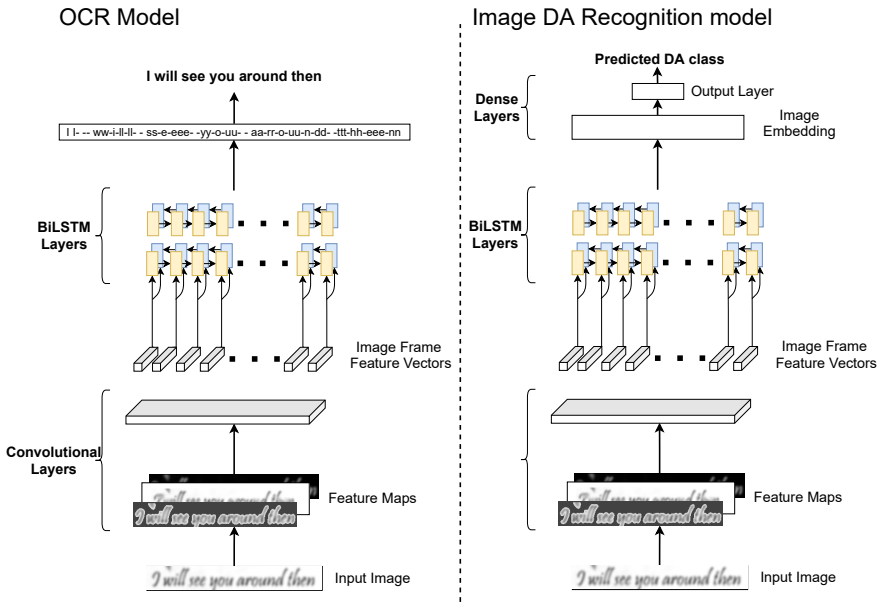


Fig. 2: CRNN models: OCR model proposed by Shi et al. [22] (left); our modified version used for visual DA recognition (right)

The inputs of both models are segmented images of utterances. The activation function of convolutional and recurrent layers is ReLU and we employed the *Adamax* optimizer.

The crucial part of the OCR model is the connectionist temporal classification (CTC) loss function which has been presented by Graves et al. [25]. The CTC is designed to create an alignment between the labels and specific image frames. It allows to use a simple form of annotation, for example, image and annotation text without the necessity of providing the precise character positions in the image. The output of the BiLSTM is given to the output which represents a probability distribution of characters per image frame.

The right part of Figure 2 visualizes our modified version for the image-based DA recognition. It doesn't utilize the CTC loss function but we use the *categorical cross-entropy* since the output is a vector of probabilities indicating the

membership in the particular DA class. The size of the output layer corresponds to the number of recognized DA categories.

3.2 Text Model

The centerpiece of this model is the Bidirectional Long Short-Term memory [26] (BiLSTM). The input utterance is aligned to 15 words, so the utterances with less than 15 words are padded with a special token while the longer ones are shortened. We chose *Word2Vec* [27] embeddings as a representation of the input text. The word vectors (with the dimension equal to 300) are fed into the BiLSTMlayer and the final states of both LSTMs are connected to a dense layer with size 400. Then a DA label is predicted through the softmaxed output layer. The model is depicted in Figure 3.

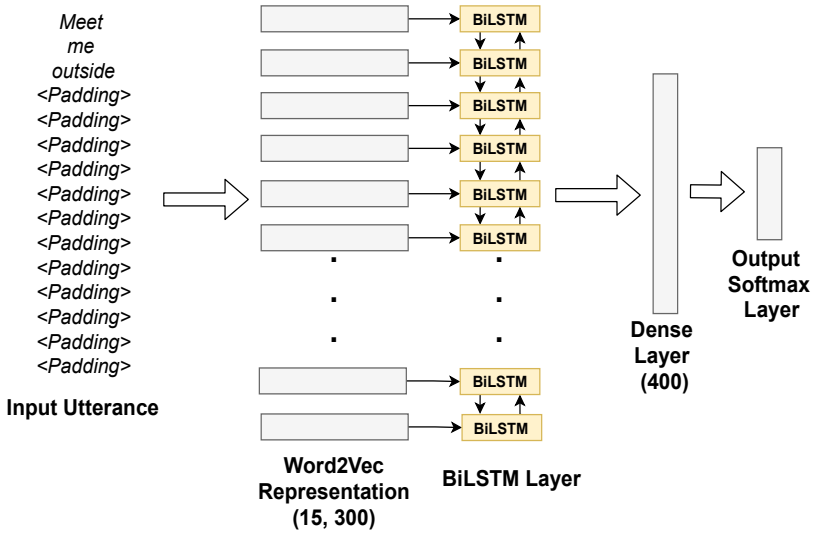


Fig. 3: Text DA recognition model

3.3 Joint Model

The second employment of the CRNN is in the combination with the text model presented in the previous section. The objective is to create a joint model that takes multi-modal input (segmented utterance image and simultaneously the text of an utterance). Figure 4 shows the Joint model with both inputs.

Since the input text doesn't have to be well-recognized, some words which are out of vocabulary might appear resulting in a worse performance of the text model. In such a case, the Image Embedding input should help to balance this loss of text information.

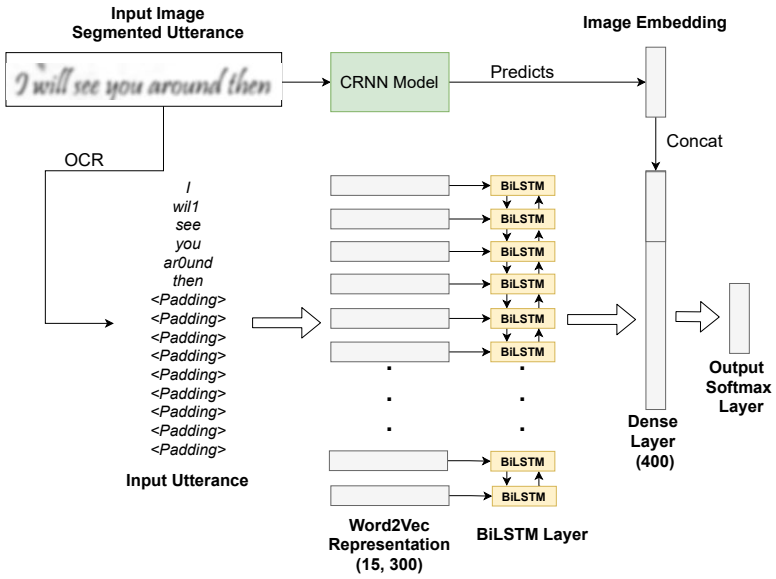


Fig. 4: Joint DA recognition model

4 Dataset

The multi-lingual VERBMOBIL dataset [5, 11] contains English and German dialogues, but we limit ourselves only to the English part. The dataset is very unbalanced. The most frequent labels are FEEDBACK (34%) and INFORM (24%) while the eight least frequent labels occur in only 1% of utterances or less.

The VERBMOBIL data are already split into training and testing parts and stored in CONLL format. We created a validation data by taking the last 468 dialogues from the training part. To summarize, we have 8921 utterances in the training part, 667 utterances as validation data and, finally, 1420 utterances serve as our test dataset.

4.1 Image Dataset Acquisition

For each dialogue, we have created four pages with image backgrounds of different noise level and programmatically rendered the utterances.

The first background (*noise_0*) contains no noise (perfectly scanned blank piece of paper) while the fourth level (*noise_3*) contains significant amount of noise[3].

[3] The noise is not artificial (i.e. we didn't perform any image transformation), but we have created the noise by real usage of the scanner. We put a blank piece of paper in the scanner and we changed the scanning quality by different scanning options and the amount of light.

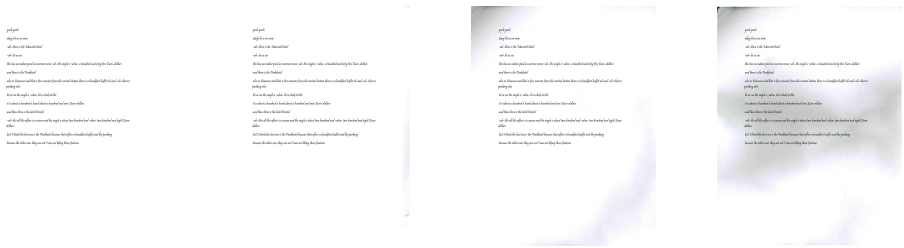
Each rendered utterance is considered as a paragraph. We must take into account, though, the utterances that are too long to fit the page width. In such a case, it continues on the next line and we would struggle with the situation where the beginning of the next utterance and continuing of the current utterance would be indistinguishable. Therefore, we increased the vertical space between paragraphs and we employed the intending of the first line of paragraphs. These two precautions together solve the above-mentioned potential problem and make the segmentation easier.

Another parameter that can be used to adjust the dataset difficulty is the font. We chose the *Pristina Font* which is a hybrid between printed and handwritten font.

Summing up, four steps of the acquisition of the image dataset are as follows.

1. Split original `VERBMOBIL CONLL` files to the individual dialogues;
2. Create the realistic scanned noisy background;
3. Choose a font;
4. Render the dialogues according to the above-mentioned scenario.

Figure 5 shows the examples of each dataset.



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

Fig. 5: Page examples from all four datasets

We have also created a second version of each of the four datasets. We have artificially applied random image transformations (rotation, blurring, and scaling). These transformations significantly increase the difficulty of our task because the segmentation and OCR will become harder to perform. We call this version the *transformed dataset* and in the following text, it will be labeled as follows: (*noise_0_trans*, *noise_1_trans*, *noise_2_trans*, *noise_3_trans*).

So in total, we have eight datasets of different noise levels and difficulties. Scripts for the dataset creation are available online[4].

[4] https://github.com/martinekj/image-da-recognition

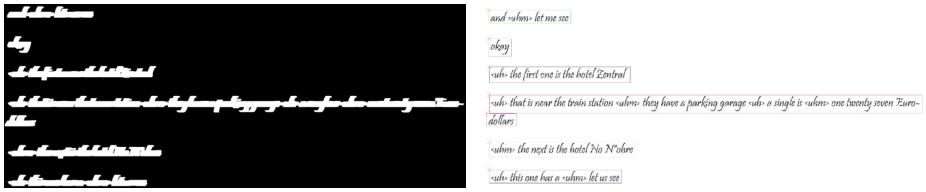
4.2 Utterance Segmentation

This section describes the algorithm we used for segmentation of the entire page into individual text line images – utterances. We utilized a simple segmentation algorithm based on the analysis of connected components.

We first employed the Sauvola thresholding [28] to binarize the input image that is a necessary step to perform the connected components analysis. Before getting to that, though, we carry out the morphological dilation to merge small neigbouring components that represent fragments of words or individual characters (see Figure 6). The ideal case is if one text line is one connected component.

Fig. 6: Example of the morphological dilation with kernel $(2, 10)$

Thereafter, the analysis of the connected components is conducted. Figure 7 shows the output of this algorithm. The left part of the image shows the binarized image after morphological dilation while the right part depicts the bounding boxes detected by the analysis of connected components.



(a) Dilated binary image (b) Detected bounding boxes

Fig. 7: Utterance segmentation

Once the bounding boxes are obtained, we crop these regions from the image and resize them to the common shape (1475×50). To maintain the image quality of narrow images, we perform the image expansion to the desired width by padding with a white background.

5 Experiments

Within this section, we first present the comparison with state-of-the-art (SoTA) results and then we quantify the difficulties of our datasets by measuring the OCR performance. The next experiment presents results with various sizes of

Image Embedding within the visual model and their influences on the overall success rate.

We split the remaining experiments into three scopes to investigate the impact of the visual information in the DA recognition task. The first scope is "image-only" and its goal is to verify the performance of the visual model presented in Section 3.1. The second scope is called "text-only" and similarly as the first scope, the goal is to evaluate our text model (see Section 3.2). The purpose of the final scope is to find the best joint model which is robust enough to be able to respond to the deteriorating quality of text input. The joint model was presented in Section 3.3.

For all experiments, we employed the *Early Stopping* that checks the value of the validation loss to avoid over-fitting. We ran every experiment 5 times and we present average Accuracy, Macro F1-Score, and also Standard Deviation of each run evaluated on the testing part of each dataset.

5.1 Comparison with SoTA

Table 1 compares the results of our text model with state-of-the-art approaches on the testing part of the English `VERBMOBIL` dataset. This table shows that our results are comparable, but we need to take into account that some approaches in the table utilize the information about the label of the previous utterance. In this work, we did not use this information, since the utterance segmentation from the image is not perfect. Some utterances may be skipped or merged that results in jeopardizing the continuity of the dialogue.

Table 1: Comparison with the state of the art [accuracy in %].

Method	Accuracy
n-grams + complex features [12]	74.7
TBL + complex features [13]	71.2
LSTM + Word2vec features [4]	74.0
CNN + Word2vec features [14]	74.5
Bi-LSTM + Word2vec features [14]	**74.9**
Text model (proposed)	73.9

5.2 OCR Experiment

We use Tesseract as the OCR engine within this work. We measured the OCR performance by calculating the Word Error Rate (WER) and Character Error Rate (CER) against ground truth text in `CONLL` files.

Tesseract was employed on the testing part (1420 utterances) of each 8 datasets. The results are presented in Table 2 and depicted in Figure 8. We

can conclude that with the increasing difficulty, the WER and CER values are increasing as expected.

Table 2: OCR experiment – Word Error Rate (WER) and Character Error Rate (CER) over all datasets

	Dataset Noise Level							
	0	**1**	**2**	**3**	**0_trans**	**1_trans**	**2_trans**	**3_trans**
WER	0.132	0.149	0.132	0.143	0.319	0.322	0.306	0.325
CER	0.049	0.053	0.049	0.053	0.131	0.128	0.128	0.168

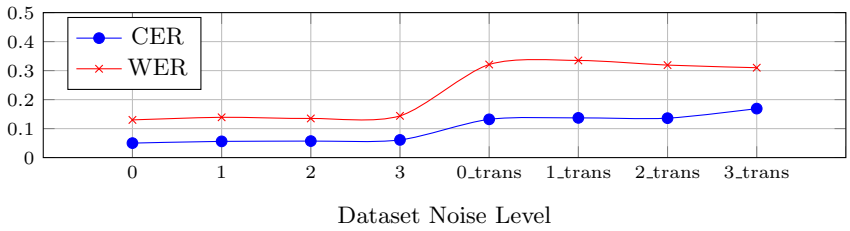


Fig. 8: Average Word Error Rate (WER) and Character Error Rate (CER) of all datasets

5.3 Image Embedding Dimension

The goal of this experiment is to find the optimal dimension of the Image Embedding (a.k.a the size of the the penultimate dense layer in the Visual model).

For this purpose, we limited ourselves only on the dataset with the poorest quality (*noise_3_trans*). We started at dimension equal to 100 and this value was gradually increased by 100. Within each run, a new model with particular embedding size was trained and evaluated. Figure 9 shows the results. We present Accuracy as the evaluation metric. The number of epochs that are needed for training was in the range 8 – 16 depending on the Early Stopping.

We have here an interesting observation that the amount of information is not increasing with the higher dimension. The best results were obtained with values 400 and 500. So for the next set of experiments, we chose the value of the Image Embedding dimension equal to 500.

5.4 Visual Model Experiment

Table 3 shows the performance of the Visual model. This table illustrates that the results are relatively consistent for all given datasets.

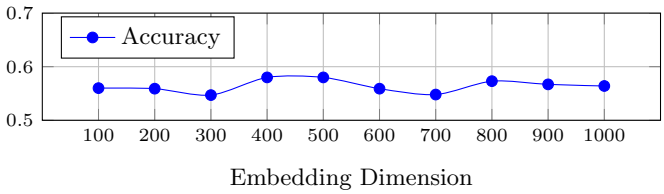


Fig. 9: Experiment to determine the optimal Image Embedding dimension. The standard deviation did not exceed the 0.006 for all runs.

Table 3: Visual model DA recognition results [in %]

	Dataset Noise Level							
	0	**1**	**2**	**3**	**0_trans**	**1_trans**	**2_trans**	**3_trans**
Macro F1	46.8	54.5	47.2	43.7	47.3	43.7	43.2	40.2
Accuracy	56.6	54.9	57.8	54.6	59.1	56.5	59.4	55.9
Std. Dev.	1.1	0.2	0.6	1.0	0.5	0.2	1.1	1.5

5.5 Text Model Experiment

Within a training phase, the model is fed with a text from the `VERBMOBIL` dataset while in the evaluation (prediction) phase the input utterances are provided by the OCR. Our intention is to create a real situation where only images with rendered text will be available and the only way to acquire the text itself is to use OCR methods. Table 4 shows Accuracies and Macro F1-scores for all datasets.

Table 4: Text Model DA recognition results [in %]

	Dataset Noise Level								
	0	**1**	**2**	**3**	**0_trans**	**1_trans**	**2_trans**	**3_trans**	**GT**
Macro F1	59.2	59.9	54.8	59.6	50.9	54.2	56.6	52.2	61.6
Accuracy	71.9	70.4	71.8	71.2	56.2	56.4	58.1	56.0	73.9
Std. Dev.	0.5	0.4	0.9	1.1	0.2	1.3	1.3	0.6	0.5

The left part of the table presents the results on not transformed datasets (*Noise_0* – *Noise_3*). For these datasets, the OCR results turned out well (see Section 5.2 – the average CER value around 0.05), which corresponds to Accuracy exceeding 0.7.

The results on transformed datasets (*Noise_0_trans* – *Noise_3_trans*) are presented in the right part of the table. The OCR performed significantly worse (average CER in range 0.12 – 0.16). Hence, the results are worse as well.

For completeness and comparison the rightmost column of Table 4 shows the results when the perfect ground truth text (from the `CONNL VERBMOBIL` files) is

used instead of the recognized text. This Accuracy is used for the comparison with state of the art (Section 5.1).

Last but not least, for transformed datasets, in terms of Accuracy, the Image and Text model performed similarly. For datasets without transformation, the Text model was significantly better, primarily due to the less amount of recognition errors.

5.6 Joint Model Experiment

The fact that it is possible to successfully train a Visual DA recognition model based solely on images with reasonable results brought us to the idea to use learned image features in combination with text to create the joint model. We assume that it might have better adaption to recognized text with a significant amount of errors.

Similar to the text model, to simulate the real situation, the ground truth text from the `VERBMOBIL` dataset is used to train the model while a recognized text from the OCR is used to test the model to verify its generalization. As long as the very same text is used in both text and joint model, it is very easy to verify and measure the positive impact and the contribution of the visual information.

Our final experiment shows, among other things, the impact of the information which was embedded into a single image feature vector (Image Embedding) by the CRNN model. Based on the preliminary experiment, we chose the dimension of embedding equal to 500.

We have eight stored CRNN models that have been trained separately on particular datasets. We remind that the training of the joint model was carried out in the same way as the training of the text model. The only difference from the previous Text Model experiment is the usage of an auxiliary image input which is predicted by the CRNN model as depicted in Figure 4. We present the results in Table 5.

Table 5: DA recognition results with Joint model [in %]

	Dataset Noise Level							
	0	**1**	**2**	**3**	**0_trans**	**1_trans**	**2_trans**	**3_trans**
Macro F1	49.6	56.8	50.3	51.9	48.3	46.8	54.2	49.3
Accuracy	60.2	60.6	61.1	63.3	61.2	59.9	66.4	60.1
Std. Dev.	0.5	0.2	0.4	0.2	0.4	0.6	0.6	3.1

As you can notice, the results no longer oscillate so much across all datasets. Another important observation is that some transformed dataset results outperformed results based on the not transformed datasets (compare *Noise_0* and *Noise_2* with their transformed versions). The help of auxiliary image input has a bigger impact on transformed datasets where the amount of noise is massive and vice versa.

Figure 10 shows the visual comparison of all models we used in our experiments. The blue curve shows Visual Model results, the red line represents Text Model results and green line depicts the performance of the Joint model (text and image input).

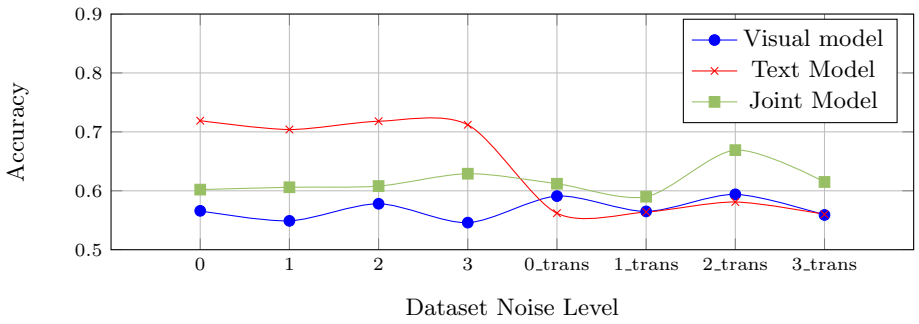


Fig. 10: Depicted results and comparison of all models

As expected, in the case of a better quality of recognized text (*Noise_0* – *Noise_3*), the performance of the text model is the best. However, if the quality of the recognized text is low (*Noise_0_trans* – *Noise_3_trans*), Accuracy and Macro F1-score decrease.

6 Conclusions

The paper has dealt with the task of dialogue act recognition in the written form using a model with multi-modal inputs. The goal of this paper has been twofold.

First, we have successfully employed the CRNN model as the visual model for image-based DA recognition. We have shown that despite employing only visual features it is possible to obtain reasonable results in the task that is dominantly text-based.

Second, we have carried out a set of experiments where we have used the same CRNN model as an image feature extractor and we have combined it with BiLSTM text model for handling both text input (obtained by OCR) and image input. We have successfully extracted the hidden layer representation of the CRNN model (Image Embedding) and together with the text model we have created the joint model. For poor-quality datasets, where the OCR success rate is low, we have outperformed the text model that uses solely text input.

Hereby, we have shown that the visual information is beneficial and the loss of the text information is partially compensated. The impact of such image features results in improving Accuracy (4% – 10%) depending on the noise level in the particular dataset.

Acknowledgements

This work has been partly supported from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications. We would like to thank also Mr. Matěj Zeman for some implementation work.

References

1. H. Bunt, "Context and dialogue control," *Think Quarterly*, vol. 3, no. 1, pp. 19–31, 1994. 1
2. A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000. 1
3. J. Frankel and S. King, "Asr-articulatory speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001. 1
4. C. Cerisara, P. Král, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech and Language*, vol. 47, pp. 175–193, 2018. 1, 9
5. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. J. Quantz, "Dialogue acts in verbmobil," 1995. 2, 6
6. J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517520. 2
7. E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The icsi meeting recorder dialog act (mrda) corpus," International Computer Science Inst Berkely CA, Tech. Rep., 2004. 2
8. J.-M. Benedı, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López, and A. Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana," in *Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 1636–1639. 2
9. P. Colombo, E. Chapuis, M. Manica, E. Vignon, G. Varni, and C. Clavel, "Guiding attention in sequence-to-sequence models for dialogue act prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7594–7601. 2
10. G. Shang, A. J.-P. Tixier, M. Vazirgiannis, and J.-P. Lorré, "Speaker-change aware crf for dialogue act classification," *arXiv preprint arXiv:2004.02913*, 2020. 2
11. J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel, *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken, 1998. 2, 6
12. N. Reithinger and M. Klesen, "Dialogue act classification using language models," in *Fifth European Conference on Speech Communication and Technology*, 1997. 2, 9
13. K. Samuel, S. Carberry, and K. Vijay-Shanker, "Dialogue act tagging with transformation-based learning," *arXiv preprint cmp-lg/9806006*, 1998. 2, 9

14. J. Martínek, P. Král, L. Lenc, and C. Cerisara, "Multi-lingual dialogue act recognition with deep learning methods," *arXiv preprint arXiv:1904.05606*, 2019. 2, 9
15. C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. Le, "Multi-task dialog act and sentiment recognition on mastodon," *arXiv preprint arXiv:1807.05013*, 2018. 2
16. J. Li, H. Fei, and D. Ji, "Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 616–626. 2
17. Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. 3
18. N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal deep networks for text and image-based document classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 427–443. 3
19. A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016. 3
20. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 3
21. R. Jain and C. Wigington, "Multimodal document image classification," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 71–77. 3
22. B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017. 3, 4
23. J. Martínek, L. Lenc, P. Král, A. Nicolaou, and V. Christlein, "Hybrid training data for historical text ocr," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 565–570. 3
24. L. Han and M. R. Kamdar, "Mri to mgmt: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks," 2017. 3
25. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376. 4
26. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 5
27. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. 5
28. J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000. 8