COMICORDA: Dialogue Act Recognition in Comic Books

Jiří Martínek, Josef Baloun, Martin Prantl, Ladislav Lenc, Pavel Král
Department of Computer Science & Engineering, NTIS - New Technologies for the Information Society
University of West Bohemia, Univerzitni 8, Pilsen, Czech Republic
{jimar, balounj, perry, llenc, pkral}@kiv.zcu.cz

Abstract

Dialogue act (DA) recognition is usually realized from a speech signal that is transcribed and segmented into text. However, only a little work in DA recognition from images exists. Therefore, this paper concentrates on this modality and presents a novel DA recognition approach for image documents, namely comic books. To the best of our knowledge, this is the first study investigating dialogue acts from comic books and represents the first steps to building a model for comic book understanding. The proposed method is composed of the following steps: speech balloon segmentation, optical character recognition (OCR), and DA recognition itself. We use YOLOv8 for balloon segmentation, Google Vision for OCR, and Transformer-based models for DA classification. The experiments are performed on a newly created dataset comprising 1,438 annotated comic panels. It contains bounding boxes, transcriptions, and dialogue act annotation. We have achieved nearly 98% average precision for speech balloon segmentation and exceeded the accuracy of 70% for the DA recognition task. We also present an analysis of dialogue structure in the comics domain and compare it with the standard DA datasets, representing another contribution of this paper.

Keywords: Comics Processing, Dialogue Act Recognition, Speech Balloon Segmentation, OCR, YOLOv8

1. Introduction

Dialogue act (DA) recognition is essential for dialogue management and understanding, which is usually utilized in dialogue systems and chatbots. The task consists of determining the type of utterance corresponding to its function in a dialogue. This task is usually targeted to the audio signal (e.g., telephonic or spontaneous dialogues, meetings, etc.). The complete task involves text transcription (manual or automatic from a speech recognition system), dialogue segmentation, and DA recognition.

This work concentrates on DA recognition from an alternative modality, namely image documents. To the best of our knowledge, no prior work explores dialogues in comic books. Even though there are some works dealing with comics and solving tasks such as emotion recognition in comics (Nguyen et al., 2021), detection and segmentation of speech balloons (Dubray and Laubrock, 2019), faces (Qin et al., 2017), characters (Sun et al., 2013; Nguyen et al., 2017), and their association (Rigaud et al., 2015), none of them tackles DA recognition.

Therefore the main contribution of this work consists in proposing a novel DA recognition approach using this modality as an input. In the case of image data, we have to perform tasks that can be considered as counterparts for the DA recognition in the speech area, as mentioned above. The first step is speech balloon segmentation, follows the OCR of the extracted regions and, finally, the DA classification.

There are several comics corpora (e.g., French eBDtheque (Guérin et al., 2013), Japanese Manga109 (Matsui et al., 2017), or BCBID (Dutta et al., 2022)) that contain annotated panels (speech or narration) balloons and of course the transcribed text. Some even include the characters and the faces (including connection to the speech balloons). Nevertheless, none of them have the DA annotation of the transcribed text.

The first motivation for this work is that DAs can be used as a clue for speech balloon order, especially in complicated situations where relying only on positional information is impossible. The correct order of speech balloons is crucial for automatic readers (e.g., for visually impaired people). This task further represents the first steps to building a model for comic book understanding.

We also present a newly created dataset COMICORDA (COMIc CORpus of Dialogue Acts) containing DA annotations. This dataset is freely available for research purposes. We further analyze the structure of dialogues in the domain of comic books and compare it with the structure of standard DA datasets.

To summarize, the main contribution of the paper is as follows:

- Proposing a novel DA recognition approach for dialogues in comic book image documents;
- Creation of the novel COMICORDA dataset, which consists of comic book images annotated with dialogue acts;
- Comparison of the newly created COMICORDA dataset with other standard DA cor-

LREC-COLING 2024, pages 3566–3578
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

pora.

The rest of the paper is organized as follows. Section 2 describes related work in the area of comics analysis and DA recognition. Section 3 presents our approach. Section 4 deals with the created dataset. The results of our experiments are provided in Section 5. The final section concludes the paper and proposes some further research directions.

2. Related Work

2.1. Comic Books Segmentation

A segmentation method designed solely for comic books was presented by Rigaud et al. (2017). The authors developed an approach based on traditional image processing techniques, such as connected component analysis and color features for speech balloon segmentation.

However, nowadays, segmentation problems are usually solved using neural networks. Many of the segmentation algorithms are based on the well-known U-Net (Ronneberger et al., 2015) architecture. A neural network based approach for speech balloon segmentation was proposed by Dubray and Laubrock (2019). The work is based on an architecture similar to U-Net utilizing VGG16 in the encoder part. This approach was later extended to classify captions and panels (Nguyen et al., 2019). Nguyen et al. (2020) presented a solution to segment bubbles based on incomplete ground-truth data. During the training, ground truths are updated based on the results from the loss function.

The task can also be viewed as object detection. In this case, we do not detect pixel-perfect bubble areas but only bounding boxes. A very efficient one-stage object detector called YOLO was proposed by Redmon et al. (2016). It started a wave of successors, currently ending with YOLOv8 Reis et al. (2023). Another example of one-stage detectors is RetinaNet Lin et al. (2017). Alternatives to one-stage detectors are two-stage ones that first detect the objects and then classify their category. Representatives of this group are RCNN Girshick et al. (2015) and Faster RCNN Ren et al. (2015). In the comics domain, such object detectors are successfully used in Qin et al. (2017) and Nguyen et al. (2017) to detect faces and characters, respectively.

2.2. OCR in Comic Books

Ponsard et al. (2012) presented one of the first approaches using OCR in the comics domain. The authors presented a comics viewer with a built-in OCR system based on traditional pattern matching on the character level.

State-of-the-art OCR engines are usually based on Convolutional-recurrent neural networks (CRNN) and connectionist temporal classification (CTC) loss (Graves et al., 2006). Rigaud et al. (2016) compare several methods for different types of text (printed or hand-written) that can be found in comics. The results show that printed texts can be fairly accurately recognized with standard Tesseract models (Smith, 2007), but hand-written texts require specially trained systems.

Iyyer et al. (2017b) presented a new dataset COMICS and tested several OCR engines, including Tesseract, Ocular (Berg-Kirkpatrick et al., 2013), and Abbyy FineReader[1]. The conclusion was that these systems could not handle the variability of fonts in the comics. Finally, they utilized Google Vision[2], which achieved much better results than the other OCR engines.

Hartel and Dunst (2021a) proposed an OCR pipeline that combines a U-Net-like architecture for text detection and the Calamari engine, which is trained for the OCR itself. They evaluated the method on the GNC corpus and achieved a 3.47% character error rate (CER) with training on the train part of the corpus. A significant drawback is that this model is not publicly available.

2.3. Dialogue Act Recognition

Dialogue act recognition models are mainly evaluated on several standard corpora containing dialogues in the form of the speech signal with the appropriate text transcription: Switchboard (SwDA) Godfrey et al. (1992), Meeting Recorder Dialogue Act (MRDA) Shriberg et al. (2004), VERBMOBIL Jekat *et al.* (1995) or DailyDialog Li et al. (2017). Previously, the recognition was mainly realized by statistical methods with handcrafted features, as presented in Stolcke et al. (2000b) or Jurafsky et al. (1997). However, state-of-the-art approaches are based on deep neural networks.

Colombo et al. (2020) presented a *seq2seq* model using neural machine translation techniques. The proposed model exploits a hierarchical encoder with an attention mechanism that does not need handcrafted features. They obtained an accuracy of 85% on the SwDA corpus and 91% on the MRDA dataset.

Chapuis et al. (2020) proposed an efficient DA recognition approach using sentence representations with a hierarchical encoder. They pre-trained a transformer model on a large corpus of spoken dialogues containing over 2.3 billion tokens. They obtained competitive results against state-of-the-art methods on several standard corpora and new state-of-the-art results on the MRDA dataset.

Martínek et al. (2019a) proposed multi-lingual dialogue act recognition approaches based on deep neural networks with word2vec embeddings for word representation. The authors used a deep

[1] https://www.abbyy.com/

[2] https://cloud.google.com/vision

convolutional neural network and an LSTM with different setups as DA classifiers. They achieved new state-of-the-art results on the VERBMOBIL corpus with an accuracy of 74.9%.

Cerisara et al. (2018b) created a new corpus from the social network Mastodon, which is annotated by dialogues and sentiment labels. They trained a multi-task hierarchical recurrent network. They experimentally showed a strong correlation between these two tasks (sentiment analysis and DA recognition).

Li et al. (2020) proposed a model for a joint task of DA recognition and sentiment classification by representing the local contexts of sentences. They first used a dynamic convolutional network for encoding to capture the dialogue contexts. Then, they proposed a context-aware dynamic convolutional network to represent the local contexts better. They extended the frameworks into a bi-channel version. They also enhanced the tasks by employing the DiaBERT language model. The authors obtained new state-of-the-art results on two standard corpora mentioned above: Mastodon and DailyDialog. Dialogue act recognition from image documents was studied only in the paper proposed by Martínek et al. (2021). The authors created the image version of the VERBMOBIL dataset and combined visual and OCRed text features to perform DA recognition.

3. Proposed Approach

The proposed method consists of three steps that are depicted in Figure 1:

1. Segmentation;
2. OCR;
3. DA recognition.

Each of them is crucial for the final recognition of DAs. Poorly segmented bounding boxes (or predicted masks) of speech balloons influence the OCR accuracy, and OCR errors have a negative impact on dialogue act recognition.

3.1. Segmentation

The first step is the detection and extraction of regions of interest (RoI) – in our case, speech balloon segmentation. There are usually four types of text within comic books:

1. Dialogues (speech balloons) – direct speech;
2. Narration text (narration balloons) – a story description;
3. Onomatopoeia (comic drawings of words that phonetically imitate, resemble, or suggest the sound that they describe) – for example: “Boom!” or “Bang!”;
4. Other text (e.g., shop names, bus stops, menu in restaurants, text in a newspaper that some character reads, etc.).

A segmentation method must consider this fact. If a scene text detector is used, it is necessary to categorize detected texts into four previously mentioned classes. However, the categorization is difficult if performed only on the text bounding boxes without a broader context.

To utilize the advantages of recently booming object detectors that can do both segmentation and categorization of detected objects, we have decided to employ YOLOv8 and Faster R-CNN and compare their performances. Both architectures can be considered as state of the art (SoTa) in this field, and they also have different characteristics (one-stage vs. two-stage, real-time optimization, etc.), which can bring interesting insights into the applicability of such methods.

3.1.1. YOLOv8

The original YOLO architecture was proposed as a real-time object detector (Redmon et al., 2016). The emphasis was placed on the inference time so that it can be used in time-critical applications. YOLOv8 (Reis et al., 2023) is currently the most-recent incarnation of this successful family of methods. The architecture came with several improvements over the older YOLO models. It brought a new anchor-free detection system. Convolutional blocks were changed and improved, and it also applies mosaic augmentation during the training.

3.1.2. Faster R-CNN

This object detector is a variant of a region-based convolution neural network (RCNN). It usually operates in two steps: a region proposal and classification. Since original versions of RCNN had some limitations (expensive training or slow object detection), the development of better and faster models has emerged, such as Faster R-CNN (Ren et al., 2015) or Mask R-CNN (He et al., 2017) which provides not only a predicted class and bounding box but also an object mask. For our experiments, we use Faster R-CNN with ResNet-50 backbone.

3.2. OCR

As an OCR engine, we have selected the solution from Google – Google Vision OCR. Based on findings from Iyyer et al. (2017b), this engine performs much better on comics than any other system they have used. It was the best out-of-the-box solution that did not require any additional training. There is no general solution to train an OCR system for comics because new comics might have a different font type, and a previously prepared solution may not work. On the other hand, Google Vision is trained on large corpora and gradually improved by Google.

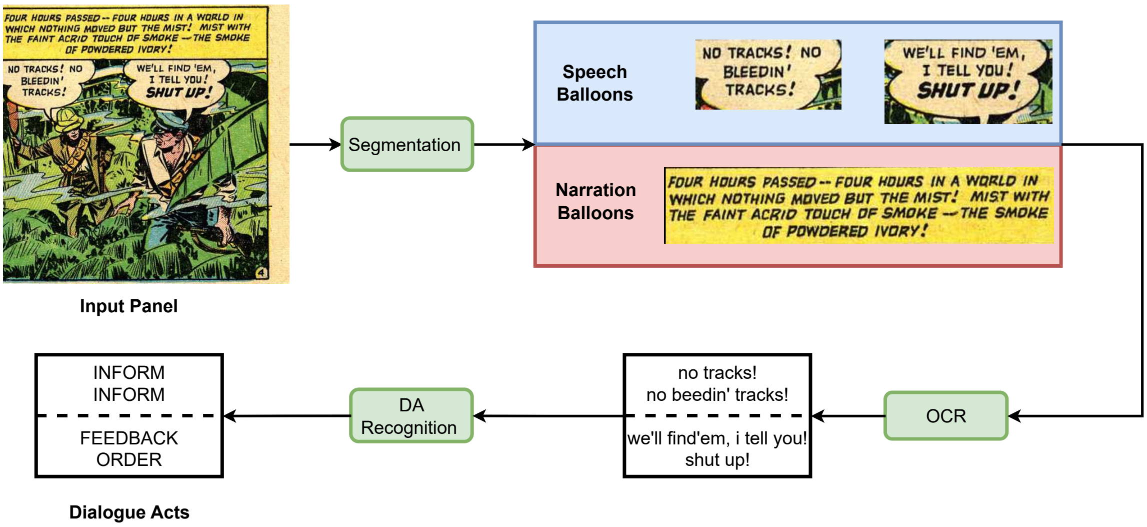

FOUR HOURS PASSED -- FOUR HOURS IN A WORLD IN WHICH NOTHING MOVED BUT THE MIST! MIST WITH THE FAINT ACRID TOUCH OF SMOKE -- THE SMOKE OF POWDERED IVORY!
NO TRACKS! NO BLEEDIN' TRACKS!
WE'LL FIND 'EM, I TELL YOU! SHUT UP!
Input Panel
Segmentation
Speech Balloons
Narration Balloons
OCR
no tracks!
no beedin' tracks!
we'll find'em, i tell you!
shut up!
DA Recognition
INFORM
INFORM
FEEDBACK
ORDER
Dialogue Acts

Figure 1: Pipeline of the proposed method

3.3. Balloon Type Classification

To tackle the possible misclassification of speech and narration balloons (since the segmentation is not perfect), we have created a simple text model that takes the OCR text as an input and classifies it into *narration* and *speech*. For this task, we have evaluated several Transformer-based models similar to the DA recognition task.

3.4. DA Recognition

We have chosen the following Transformer-based models for DA recognition and for comparison of their performance:

1. BERT model;
   (a) bert-base-uncased;
   (b) SwDA fine-tuned bert-base-uncased;
2. RoBERTa model;
3. Microsoft/DialoGPT model;
   (a) DialoGPT-small;
   (b) DialoGPT-medium;
   (c) DialoGPT-large.

We have fine-tuned all the aforementioned Transformer models for the DA recognition task. In addition to the well-known BERT model (Devlin et al., 2018), our experiment included the RoBERTa model (Liu et al., 2019) and Microsoft/DialoGPT, which represents an autoregressive model in the dialogue field. This model was trained based on the GPT-2 (Radford et al., 2019) using a causal language modeling objective on conversational data (Zhang et al., 2019).

Since the BERT model demonstrated the best performance in our preliminary experiments, we decided to use a second BERT model fine-tuned on the SwDA dataset to transfer knowledge from a related set. In this approach, we first conducted 50 training epochs on the SwDA corpus, saved the model state, and then fine-tuned it on the COMICORDA dataset. The amount of 50 epochs is based on the observation of a high amount of runs. We used the [CLS] token in the BERT and RoBERTa models to predict Dialogue Acts. For the unidirectional DialoGPT, our initial approach was to use the last token (where the aggregated information should be) during fine-tuning. However, our validation experiments indicated better results (3-5% on average) when we used all output tokens that were averaged.

We also analyzed the performance of the hierarchical-dialog-bert model from (Chalkidis et al., 2022). However, the validation results obtained were, on average, 5% lower than those of the original vanilla BERT. Therefore, this model is not included in our experiments.

4. COMICORDA Dataset

The source of comic images for our corpus is the COMICS public dataset[3] containing more than 1.2M extracted panels (Iyyer et al., 2017a).

From this database, we downloaded 800 annotated panels (speech and narration bounding boxes together with automatic text transcriptions by Google Vision) following the authors of the EMORECOM ICDAR competition Nguyen et al. (2021). The images were manually verified to correct the errors in the text recognized by the OCR engine. This step was necessary for the evaluation of the Google Vision OCR performance.

Furthermore, the trained annotators with an excellent knowledge of English were assigned dialogue

[3] https://obj.umiacs.umd.edu/comics/index.html

act labels for dialogue act recognition. We also established bounding boxes for faces and connected them to speech balloons. This annotation might be helpful for some tasks, for instance, emotion recognition.

Moreover, to have a broader spectrum of various comics, we picked another set of 638 comic panels from a different source that has been annotated in the same way.

For our experiments, we divided our dataset into three subsets:

1. Panels that contain dialogues (at least one speech balloon);

2. Panels that contain only narration balloons (no dialogue in the panel);

3. Panels that contain neither dialogues nor narration (no balloons).

All the above-mentioned groups are explicitly labeled so we can easily filter out a particular subgroup for desired experiments. In this paper, we work only with panels that contain dialogues.

In total, we have more than 1,400 annotated image panels with following attributes and properties:

- Speech balloon bounding boxes;
- Narration balloon bounding boxes;
- Text of dialogues (speech balloons) and narration (narration balloons);
- Faces with a relation to the speech balloons;
- Dialogue acts for each sentence (utterance) in a dialogue.

The exact numbers of panels, speech balloons, and utterances are summarized in Table 1. The annotations for the COMICORDA dataset are freely available for research purposes[4]. The `img_id` tags correspond to the IDs of image panels from the above-mentioned public dataset (Iyyer et al., 2017a).

Dataset item	**Counts**
Panel	1438
Speech balloon (dialogue)	2196
Utterance (DA)	2282

Table 1: Main dataset information

4.1. Annotation Scheme

Our annotations and instructions for annotators are based on the annotation guide (Alexandersson et al., 1997), where the DAs are defined, explained, and accompanied by examples. We also used the `Not_Classifiable` tag, which is used when dialogue segments are too fragmentary or uncomprehensible (e.g., "...but you." or "is ... is"). The group of annotators was composed of three trained persons: A, B & C, with good knowledge about the DA tagging. If the annotations differ, the resulting DA tag is chosen based on the most frequent label assigned by the annotators (majority voting). In cases of total disagreement, the final DA tag is determined through a consensus among the annotators.

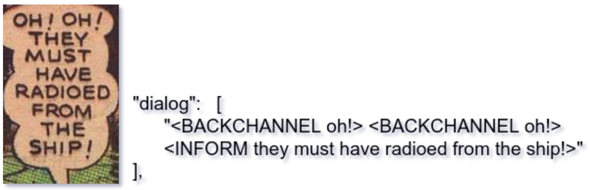
To assess the quality of annotations and inter-annotator agreement, we calculated both Cohen's kappa and Fleiss's kappa values (see Table 2).

	A&B	**A&C**	**B&C**
Cohen's kappa	0.748	0.788	0.855
Fleiss kappa	0.783		

Table 2: Inter-annotator agreement for DA annotations.

Since sentences expressing an order/command appear quite often, we have dedicated a special DA class to it `Order`, despite the fact that from the lexical point of view, it is a statement. In VERBMOBIL, all questions are grouped into one general DA `Request`. The most common question types, though, might be easily distinguished by even an amateur annotator, so annotators had an instruction to label all questions as: `Yes_No_Question`, `Wh_Question` or `Or_Question` that are the most common variants of questions.

In our case, one dialogue segment is the text within a single speech balloon. This segment is composed of individual sentences that we consider utterances. Each utterance is annotated with a DA label (see Figure 2). Table 3 shows the DA distribution in the dataset.



```
"dialog": [
    "<BACKCHANNEL oh!> <BACKCHANNEL oh!>
    <INFORM they must have radioed from the ship!>"
],
```

Figure 2: Example of DA annotation

We can see that more than half of the utterances are labeled as *Inform*. This is mainly due to the fact that in comics, the story must be pushed forward, and there is also a lack of space. Contrary to standard spoken dialogues, where sentences are often much longer and more complex in terms of word choice, events in comic books are described in both text and images. The author of a comic book is forced to create relatively short information statements with fewer words because speech balloons cannot be so big as to overlap the image behind and possibly "corrupt the image story".

Besides speech balloons, there are also "narration balloons" which contain text that represents narra-

[4] `https://corpora.kiv.zcu.cz/comicorda/`

Dialogue act	Counts	%
Inform	1280	56.1
Backchannel	230	10.1
Order	192	8.4
Feedback	155	6.8
Wh_Question	107	4.7
Greet	73	3.2
Not_Classifiable	69	3.0
Yes_No_Question	64	2.8
Politeness_Formula	63	2.8
Commit	20	0.8
Thank	10	<0.5
Offer	9	<0.5
Or_Question	5	<0.5
Close	4	<0.5
Bye	1	<0.5

Table 3: DA distribution and frequencies in our dataset

tions (plot within a story). These areas do not contain dialogues and are often created as rectangle-shaped balloons. Last but not least, the annotators had instructions not to annotate the *Onomatopoeia* since these texts are very different from narrations or dialogues and are out of the scope of this paper.

4.2. Comparison with Other DA Corpora

This section compares the created dataset with other standard DA recognition corpora. We analyze utterance lengths, label types, and their distributions. Lastly, we present state-of-the-art results across all corpora to demonstrate the performance of our proposed dialogue act recognition approach compared to the existing state-of-the-art methods, focusing solely on text data.

4.2.1. Utterance Length Analysis

The average utterance length is six words. Generally, the speech balloon utterances are shorter than usual utterances from sound recordings. The summary is provided in Table 4. In COMICORDA, the vast majority of utterances have from 1 to 18 words, while in the other three corpora, the upper limit is 25. Furthermore, the longest utterance has 40 words which is significantly fewer than in the other corpora. The authors of comics obviously try to keep the utterances and dialogues as short as possible due to the limited size of balloons.

Dataset	MAX	AVG	MED
SwDA	79	7.82	6
MRDA	79	6.88	4
VERBMOBIL	113	7.54	5
COMICORDA (our)	40	6.09	5

Table 4: Average number of words per utterance in the different DA corpora

4.2.2. Analysis of the DA Label Distribution

This analysis studies the distribution of the DA labels across the corpora. However, it is difficult to compare the label distributions directly because DA recognition corpora use different sets with a different number of labels (see Figure 3) defined by various annotation schemes and also depending on the target domain. Therefore, we compare the most frequent ones in each corpus (see Table 5).

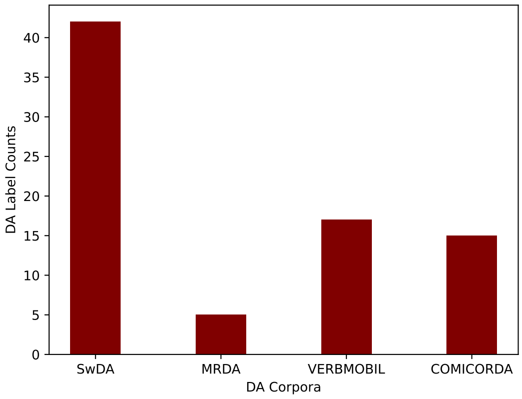


Figure 3: Counts of DA labels in various DA corpora

Based on this table, we can state the following:

- Statements (information utterances) are predominant in all corpora;
- Questions have similar distributions among all corpora (in VERBMOBIL annotated as “Request”);
- MRDA and SwDA datasets have relatively big amounts of “disruption” or “abandoned” tags which means interrupted dialogues/utterances. Interruptions are hard to express in the form of an image and are very rare in comic books;
- Distribution of DA classes between VERBMOBIL and COMICORDA corpora is different even though they use a very similar annotation guide. This is caused due to the different target domains (spontaneous speech vs. comic books).

4.2.3. State-of-the-art Results

Table 6 shows the best-obtained accuracies on all datasets. The highest accuracy in the case of the MRDA dataset is explained by the usage of the basic label set containing only five classes.

The closest label distribution to our proposed dataset COMICORDA is in the VERBMOBIL dataset. Moreover, the BERT model was consistently the best in our validation experiments. Therefore, we used VERBMOBIL corpus and BERT model for this experiment.

We have reached the accuracy of 76.2% and have outperformed the SoTa results from Martínek et al. (2019a).

SwDA		MRDA		VERBMOBIL		COMICORDA	
Statement-non-opinion	36%	Statement	59%	Feedback	26%	Inform	56%
Acknowledgement	19%	Backchannel	14%	Inform	20%	Backchannel	10%
Statement-opinion	13%	Disruption	13%	Suggest	20%	Order	8%
Agreement	5%	Floorgrabber	7%	Request	8%	Feedback	7%
Abandoned	5%	Question	6%	Bye	4%	Wh_Question	5%
Yes-No-Question	2%	-	-	Greet	4%	Greet	3%

Table 5: The summary of the most frequent DA in a particular dataset

Dataset	Accuracy
SwDA	85.0% (Colombo et al., 2020)
MRDA	92.4% (Chapuis et al., 2020)
VERBMOBIL	74.9% (Martínek et al., 2019a) 76.2% SwDA BERT model (ours)

Table 6: State-of-the-art classification results on particular corpora

5. Experiments

Similarly, as Nguyen et al. (2021), we assume a single comic panel as an input instead of a whole comic page. First, we present dialogue segmentation experiments (speech/narration balloon detection). Then, we describe the text classification experiments: 1) narration vs dialogue; 2) dialogue act recognition. For all our experiments, we use 5-fold cross-validation. For all text classification experiments, we use a maximum input length of 50 tokens, a batch size 64, and the AdamW optimizer with a learning rate of 1e-4.

Moreover, all text experiments were split into two scenarios:

1. Text from ground-truths (without OCR errors);
2. Text from OCR.

Last but not least, we employed an SVM (Support Vector Machine) classifier as a baseline for all text experiments.

5.1. Speech Balloon Segmentation

The first step in the whole processing pipeline is the segmentation of dialogues (i.e., speech balloons). To achieve this task, we have evaluated two object detection models – YOLOv8 and Faster R-CNN.

We have tested all acceptance thresholds between 0.05 and 0.95 (with the step of 0.05). The best result was obtained for YOLOv8 with the threshold set to 0.25. With this setting, we obtained the best recall while having high enough precision. In the case of Faster R-CNN, the confidence threshold of 0.85 brought the best combination of high recall and high precision.

The segmentation results are shown in Tables 7, 8 and 9.

Model	Speech Balloons	
	AP$^{iou=0.5}$	AP$^{iou=0.75}$
YOLOv8	0.954	**0.823**
Faster R-CNN	**0.978**	0.816

Table 7: Average precision results: speech balloons

Model	Narration Balloons	
	AP$^{iou=0.5}$	AP$^{iou=0.75}$
YOLOv8	**0.838**	**0.655**
Faster R-CNN	0.658	0.645

Table 8: Average precision results: narration balloons

Model	mAP$^{iou=0.5}$	mAP$^{iou=0.75}$
YOLOv8	**0.896**	**0.739**
Faster R-CNN	0.818	0.731

Table 9: Mean average precision results

The YOLOv8 model achieved the best mean Average Precision (mAP) results.

5.2. Narration vs Dialogue Text Classification

A situation when a balloon is misinterpreted (a speech balloon mistakenly replaced by a narration

Model	GT text Accuracy	OCR Text Accuracy
SVM (baseline)	86.7	71.7
bert-base-uncased	**93.4**	**93.7**
roberta-base	93.1	**93.7**
DialoGPT-small	92.1	92.0
DialoGPT-medium	**93.4**	93.2
DialoGPT-large	92.3	93.3

Table 10: Narration vs dialogue text classification results [in %]

Model	GT text		OCR text	
	Accuracy	Macro-F1	Accuracy	Macro-F1
SVM (baseline)	67.1	43.7	57.2	35.5
bert-base-uncased	77.7	**55.5**	67.5	48.3
SwDA bert-base-uncased	**78.5**	52.3	**70.0**	**49.9**
RoBERTa-base	75.9	51.9	65.5	43.0
DialoGPT-small	72.6	46.1	65.3	44.6
DialoGPT-medium	73.0	48.7	66.8	44.0
DialoGPT-large	77.1	54.1	67.7	45.2

Table 11: DA recognition results [in%]

and vice versa) might happen despite the relatively high mAP score for the YOLOv8 model. To achieve the lowest possible error level in the entire system, we have implemented a binary text classifier that distinguishes between narrations and dialogues in the text. This model is similar to the DA recognition model, but it differs in the number of output neurons. This way, we obtain a score of how likely the text is an utterance (part of a dialogue) to deal with the situation when the DA recognition model wrongly processes a narration text.

Table 10 shows the classification result. We used the standard Accuracy (ACC) metric for evaluation. All tracked models except the SVM show stable performance regardless of being fed by the OCRed text or the ground truth.

5.3. Dialogue Act Recognition

This experiment evaluates the final module of our pipeline – the DA recognition model (see Table 11). We remind that this experiment has two parts: 1) to test the quality of DA recognition models on ground truth text and 2) to study the impact of OCR errors when the OCR is used.

This table indicates significant differences in results between GT text and text obtained via OCR. BERT-based models achieved the best DA recognition results in both scenarios. Notably, the Macro-F1 values for OCR text almost reached 50%.

We measured the character error rate (CER) of the utilized Google Vision engine and obtained a value of 10.3%. The results from both text classification experiments (Tables 10 and 11) show that the impact of such an amount of errors in the OCR output is considerable.

6. Conclusions

In this paper, we explored dialogue act recognition in comic books. To the best of our knowledge, this is the first attempt to perform DA recognition from this type of documents. To achieve this, we created a novel dataset for DA recognition in comic books. We proposed an approach that includes speech balloon segmentation, OCR, and DA recognition. The task of DA recognition in comics can be beneficial for automated comics processing. Specifically, the detected DAs can be used as a hint for speech balloon ordering in cases where it is complicated based solely on the positional information. This work further represents the first steps to build a model for comic book understanding.

We evaluated and compared the performance of two efficient neural models for balloon segmentation, namely YOLOv8 and Faster R-CNN. We showed that both models work well on this task, obtaining AP above 95% at the 0.5 IoU level for balloon segmentation. However, the YOLOv8 model performed significantly better for narration balloons, which renders it more robust and more suitable for our task. As an auxiliary check, we employed a text classifier to differentiate the two balloon classes (narration vs. speech balloons) with a success rate above 93%.

We used the Google vision API for the OCR task based on relevant studies. The obtained character error rate is slightly above 10%, which is still too high for acceptable DA recognition results. The impact of OCR errors in terms of DA recognition accuracy is −8.5% for our best DA recognition model.

We evaluated and compared several Transformer-based models for the DA recognition task and even employed transfer learning from the Switchboard dataset for the BERT model, achieving the best results with accuracies of 78.5% using manual transcripts and 70.0% for OCRed text. Additionally, we evaluated the final BERT model on the VERBMOBIL dataset and achieved a new state-of-the-art result, surpassing the previous best system by 1.3%.

7. Acknowledgements

This work has been partly supported by the Grant No. SGS-2022-016 Advanced methods of data processing and analysis. We would also like to thank Mr. Matěj Zeman for his help in data annotating.

8. Limitations and Ethical Considerations

This work is the first attempt for DA recognition from comic books. Therefore, no previous work in this research area exists and is not possible to compare DA recognition performance in this area directly with other works. However, we compare the DA recognition step using text input.

We created a new dataset containing DA annotation from comic books and proposed one approach (composed of three steps) for DA recognition obtaining encouraging results. However, these positive results are not guaranteed on other comic corpora from different domains. All other datasets are correctly cited and links are provided.

The last limitation is that the proposed method is mono-lingual and can recognize DAs in English only. However, the possible extension is relatively easy and lies in replacing the DA recognition step.

9. Bibliographical References

Ehsan Ahmadi, Zohreh Azimifar, Maryam Shams, Mahmoud Famouri, and Mohammad Javad Shafiee. 2015. Document image binarization using a discriminative structural classifier. *Pattern recognition letters*, 63:36–42.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. Dialogue acts in verbmobil-2.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. 2019. Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 427–443. Springer.

José-Miguel Benedı, Eduardo Lleida, Amparo Varona, Marıa-José Castro, Isabel Galiano, Raquel Justo, I López, and Antonio Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

S.V. Burtsev and Ye.P. Kuzmin. 1993. An efficient flood-filling algorithm. *Computers & Graphics*, 17(5):549 – 561.

C. Cerisara, P. Král, and L. Lenc. 2018a. On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech and Language*, 47:175–193.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018b. Multitask dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013*.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David Dubray and Jochen Laubrock. 2019. Deep cnn-based speech balloon detection and segmentation for comic books. In *2019 International*

Conference on Document Analysis and Recognition (ICDAR), pages 1237–1243. IEEE.

Arpita Dutta, Samit Biswas, and Amit Kumar Das. 2022. Bcbid: first bangla comic dataset and its applications. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–15.

Joe Frankel and Simon King. 2001. Asr-articulatory speech recognition. In *Seventh European Conference on Speech Communication and Technology*.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2015. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517âĂŞ520, USA. IEEE Computer Society.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. 2013. ebdtheque: a representative database of comics. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1145–1149. IEEE.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Lichy Han and Maulik R Kamdar. 2017. Mri to mgmt: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks.

Rita Hartel and Alexander Dunst. 2021a. An ocr pipeline and semantic text analysis for comics. In *International Conference on Pattern Recognition*, pages 213–222. Springer.

Rita Hartel and Alexander Dunst. 2021b. An ocr pipeline and semantic text analysis for comics. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 213–222, Cham. Springer International Publishing.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017a. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry Davis. 2017b. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Rajiv Jain and Curtis Wigington. 2019. Multimodal document image classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 71–77. IEEE.

Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J Joachim Quantz. 1995. Dialogue acts in verbmobil.

S. Jekat *et al.* 1995. Dialogue Acts in VERBMOBIL. In *Verbmobil Report 65*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykem̃a. 1997. Automatic Detection of Discourse Structure for Speech Recognition and Understanding. In *IEEE Workshop on Speech Recognition and Understanding*, pages 88–95, Santa Barbara.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jingye Li, Hao Fei, and Donghong Ji. 2020. 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*, pages 616–626.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiří Martínek, Pavel Král, and Ladislav Lenc. 2021. Dialogue act recognition using visual information. In *International Conference on Document Analysis and Recognition*, pages 793–807. Springer.

Jiří Martínek, Pavel Král, Ladislav Lenc, and Christophe Cerisara. 2019a. Multi-Lingual Dialogue Act Recognition with Deep Learning Methods. In *Proc. Interspeech 2019*, pages 1463–1467.

Jiří Martínek, Ladislav Lenc, Pavel Král, Anguelos Nicolaou, and Vincent Christlein. 2019b. Hybrid training data for historical text ocr. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 565–570. IEEE.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. 2017. Comic characters detection using deep learning. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 3, pages 41–46. IEEE.

Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. 2019. Multi-task model for comic book image analysis. In *International Conference on Multimedia Modeling*, pages 637–649. Springer.

Nhu-Van Nguyen, Christophe Rigaud, Arnaud Revel, and Jean-Christophe Burie. 2020. A learning approach with incomplete pixel-level labels for deep neural networks. *Neural Networks*, 130:111–125.

Nhu-Van Nguyen, Xuan-Son Vu, Christophe Rigaud, Lili Jiang, and Jean-Christophe Burie. 2021. Icdar 2021 competition on multimodal emotion recognition on comics scenes. In *International Conference on Document Analysis and Recognition*, pages 767–782. Springer.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Christophe Ponsard, Ravi Ramdoyal, and Daniel Dziamski. 2012. An ocr-enabled digital comic books viewer. In *International Conference on Computers for Handicapped Persons*, pages 471–478. Springer.

Xiaoran Qin, Yafeng Zhou, Zheqi He, Yongtao Wang, and Zhi Tang. 2017. A faster r-cnn based method for comic characters face detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1074–1080. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. 2023. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*.

Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *Fifth European Conference on Speech Communication and Technology*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. 2017. Text-independent speech balloon segmentation for comics and manga. In *International Workshop on Graphics Recognition*, pages 133–147. Springer.

Christophe Rigaud, Nam Le Thanh, J-C Burie, J-M Ogier, Motoi Iwata, Eiki Imazu, and Koichi Kise. 2015. Speech balloon and speaker association for comics and manga understanding. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 351–355. IEEE.

Christophe Rigaud, Srikanta Pal, Jean-Christophe Burie, and Jean-Marc Ogier. 2016. Toward speech text recognition for comic books. In *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding*, MANPU '16, New York, NY, USA. Association for Computing Machinery.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Ken Samuel, Sandra Carberry, and K Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. *arXiv preprint cmp-lg/9806006*.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Jaakko Sauvola and Matti Pietikäinen. 2000. Adaptive document image binarization. *Pattern recognition*, 33(2):225–236.

Guokan Shang, Antoine Jean-Pierre Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020. Speaker-change aware crf for dialogue act classification. *arXiv preprint arXiv:2004.02913*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. 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*, 39(11):2298–2304.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, International Computer Science Inst Berkely CA.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000a. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meṫeer. 2000b. Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In *Computational Linguistics*, volume 26, pages 339–373.

Weihan Sun, Jean-Christophe Burie, Jean-Marc Ogier, and Koichi Kise. 2013. Specific comic character detection using local feature matching. In *2013 12th International Conference on Document Analysis and Recognition*, pages 275–279. IEEE.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Jerod Weinman, Ziwen Chen, Ben Gafford, Nathan Gifford, Abyaya Lamsal, and Liam Niehus-Staab. 2019. Deep neural networks for text detection and recognition in historical maps. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 902–909. IEEE.

Christoph Wick, Christian Reul, and Frank Puppe. 2020. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly*, 14(1).

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt:

Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuan Zhuang and Ellen Riloff. 2020. Exploring the role of context to distinguish rhetorical and information-seeking questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 306–312.