

Detecting and Grounding Important Characters in Visual Stories

Danyang Liu, Frank Keller

Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
danyang.liu@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Characters are essential to the plot of any story. Establishing the characters before writing a story can improve the clarity of the plot and the overall flow of the narrative. However, previous work on visual storytelling tends to focus on detecting objects in images and discovering relationships between them. In this approach, characters are not distinguished from other objects when they are fed into the generation pipeline. The result is a coherent sequence of events rather than a character-centric story. In order to address this limitation, we introduce the VIST-Character dataset, which provides rich character-centric annotations, including visual and textual co-reference chains and importance ratings for characters. Based on this dataset, we propose two new tasks: important character detection and character grounding in visual stories. For both tasks, we develop simple, unsupervised models based on distributional similarity and pre-trained vision-and-language models. Our new dataset, together with these models, can serve as the foundation for subsequent work on analysing and generating stories from a character-centric perspective.

Introduction

Visual storytelling, which aims to generate a fluent and coherent story based on an input image sequence, has recently received increasing attention from both Computer Vision and Natural Language Processing areas (Hsu et al. 2021; Xu et al. 2021; Chen et al. 2021; Hsu et al. 2020; Wang et al. 2020; Huang et al. 2016). Unlike the traditional image captioning task where the output is usually descriptive text, visual storytelling requires more than just knowing the objects in the image sequence and describing them literally. A convincing narrative must be composed by connecting different images and reasoning about what happens in the story. In general, the most important elements of a story are the plot and the characters. A human writer usually identifies the characters in the story from the images first, and determines who the main character is based on the plot they want to describe. Establishing the characters before writing the story will make the plot clearer and the overall narrative more convincing. In recent years, there has been some work published focused on character-centred story generation. For example, Inoue et al. (2022) and Liu et al. (2020) propose to learn

character representations to build a character-centric storytelling model. Brahman et al. (2021) present a new dataset of literary pieces paired with descriptions of characters that appear in them for character-centric narrative understanding. There is also a long line of work which aims to link character mentions in movie or TV scripts with their visual tracks (Rohrbach et al. 2017; Bojanowski et al. 2013; Everingham, Sivic, and Zisserman 2006; Sivic, Everingham, and Zisserman 2009; Tapaswi, Bäuml, and Stiefelhagen 2012).

Existing work on visual storytelling, however, tends to focus on detecting objects in images and discovering relationships between them, while neglecting character information. Characters are treated as equally important as other objects, and fed into a generation pipeline without distinguishing them. For example, a recent approach to visual storytelling (Xu et al. 2021; Chen et al. 2021; Hsu et al. 2020; Yang et al. 2019) is to exploit an external commonsense knowledge graph to enrich the detected objects and thus improve coherence of the generated stories. Despite being successful in describing the sequence of events, such an approach is unable to model the characters in a story. It ends up generating stories with incorrect co-reference and arbitrary characters – characters that don't have a consistent personality or background story. Detailed examples of the output of an existing system can be found in the supplementary material.

Moreover, existing approaches are unable to take into account the importance of characters, which means they cannot distinguish protagonists from side characters in the generated stories. The fact that there are no visual story datasets with character annotation makes it difficult to develop models for character-centric storytelling.

In this paper, we introduce the VIST-Character dataset, which augments the test set of VIST (the Visual Storytelling dataset; Huang et al. 2016) with rich character-related annotation. In VIST-Character, all mentions of the characters in a story are marked in the text and identified by bounding boxes in the image sequence. Mentions of the same character (both textual and visual) are annotated as part of a single co-reference chain. Furthermore, the dataset includes a rating of all characters according to their importance in the story. This makes it possible to distinguish protagonists and side characters. Figure 1 shows an example story from the VIST-Character dataset.

Based on our VIST-Character dataset, we propose two

new tasks: (1) important character detection and (2) character grounding in visual stories. The goal of the first task is to identify and rank characters according to their importance to the story. For this task, we can use the text of the story, or the sequence of images that illustrates the story, or both modalities. Important character detection can be seen as a the first step in a generation pipeline that aims to produce character-centric (rather than object-centric) stories.

The second task, character grounding, requires us to ground the textual mentions of the characters to the relevant bounding boxes in the image sequence. This presents two challenges. Firstly, identifying mentions of the same characters in an image sequence is harder than in videos. This is because we only have only a few discrete and discontinuous images to work with, compared to the many frames typically available for video input. Secondly, character grounding is harder than traditional language grounding. We now need to align a chain of multiple mentions of characters with a chain of multiple visual appearances. Grounding is now many-to-many, compared to traditional one-to-one ground of a single phrase to a single bounding box.

In this paper, we develop simple, unsupervised models for both important character detection and character grounding, using distributional similarity heuristics or large pre-trained vision-and-language models. We believe that our new dataset, together with these models, will serve as the foundation for subsequent work on visual storytelling from a character-centric perspective.

The VIST-Character Dataset

The VIST-Character dataset contains 770 visual stories with character-related annotations, including the character co-reference chains in both stories and image sequences, and the importance rating of characters. The visual stories are selected from the VIST test set. We exclude the stories where there is no such character in any of the images (e.g., if the images all show landscape or scenery). Detailed selection criteria are included in the supplementary material. Figure 1 shows an example story from VIST-Character.

Annotation Procedure The two authors conducted an annotation pilot study, based on which they devised the annotation instructions. Overall, the annotation involves three steps: (1) Mark the words that refer to the same characters in the story sentences. (2) Identify the characters in the images. The annotators are asked to draw bounding boxes around the whole body of individual characters. For plural (e.g., *students*) and group (e.g., *the team*) characters, draw bounding boxes for each individual character when it refers to less than five characters, and draw a single bounding box to contain all the characters when it refers to at least five characters. (3) Rate the importance of each character by giving 1–5 stars. We quadruple annotated fifty stories; the rest of the dataset was single annotated. We created an annotation interface based on LabelStudio (<https://labelstud.io/>), an HTML-based tool that allows us to combine text, image, and importance annotation.

Inter-annotator Agreement Four annotators with linguistic and NLP background were trained using our annota-

Name	Precision	Recall
Character Detection	76.4%	71.5%
Co-reference (B-Cubed)	82.0%	79.4%
Co-reference (Exact Match)	63.2%	58.1%
Bounding Boxes	67.8%	58.6%
Importance Ranking	–	73.0%

Table 1: Inter-annotator agreement for the double annotations on fifty examples of the VIST-Character dataset.

Name	Value
Number of stories	770
Number of characters	3,119
Number of plural and group characters	768
Number of bounding boxes	4,979
Average number of bounding boxes per story	6.47
Average number of characters per story	4.05
Average length of textual co-reference chain	2.00
Average length of visual co-reference chain	2.02

Table 2: Statistics of the VIST-Character dataset.

tion instructions. Given the complexity of the task, we computed the inter-annotator agreements between the four annotators in four categories: character detection, co-reference chains, bounding boxes, and importance ranking. Table 1 shows the results. The annotation of one of the authors was used as ground truth to compute the precision and recall of of the other three annotators. We evaluate the co-reference chains using B-Cubed and exact match. B-Cubed is the proportion of correctly predicted mentions in a co-reference chain. Exact match means the predicted co-reference chain as a whole is equivalent to the gold-standard chain. For bounding boxes, we define $IoU = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}$, i.e., the intersection over the union of the bounding boxes B_1 and B_2 . A bounding box is considered correct if its IoU with the gold bounding box is higher than 60%. The results in Table 1 show good agreement given the complexity of the task.

Dataset Statistics Table 2 presents various statistics of the VIST-Character dataset. We can observe that there are four characters on average in each story, and each character appears about twice. There are 768 plural and group characters, accounting for 24.6% of the total. This shows that plural and group characters are frequent and important in stories.

Task Formulation

We propose the task of important character detection and grounding in visual stories. The input can be a story of C sentences $\{S_i\}_{i=1}^C$, a sequence of C images $\{I_i\}_{i=1}^C$, or a sequence of C images $\{I_i\}_{i=1}^C$ and associated sentences $\{S_i\}_{i=1}^C$ ($C = 5$ in the VIST dataset). The task is then:

1. For text-only input, we need to identify the character co-reference chains $\{\mathcal{T}_i\}_{i=1}^{K_t}$ in the story text, where K_t denotes the the number of characters in the story text and \mathcal{T}_i is the co-reference chain that contains all the mentions referring to the i -th textual character.



Figure 1: An example of the VIST-Character dataset. The right-hand panel shows the annotator’s importance rating for each character on a scale of 1–5. The characters are colour-coded across sentences, images, and importance rating.

- For image-only input, we detect all the characters in the image sequence and cluster them into visual co-reference chains $\{\mathcal{V}_i\}_{i=1}^{K_v}$, where K_v denotes the the number of characters in the image sequence and \mathcal{V}_i is the visual co-reference chain that contains all the visual appearances referring to the i -th visual character.
- For multi-modal input, after the above two steps, the textual and visual co-reference chains (i.e., $\{\mathcal{T}_i\}_{i=1}^{K_t}$ and $\{\mathcal{V}_i\}_{i=1}^{K_v}$) are aligned to obtain multi-modal co-reference chains $\{\mathcal{M}_k\}_{k=1}^{K_m}$, where $K_m = \max(K_t, K_v)$ is the total number of characters and \mathcal{M}_i is the multi-modal co-reference chain that contains all textual mentions and visual appearances of the i -th character.

Regardless of the modality of the input, we also need to output an importance score for each character (i.e., for each co-reference chain).

Methodology

In this section, we will introduce our model for detecting and grounding important characters. This model is meant as a baseline for our proposed new dataset and task. For modelling purposes, we decompose the task into four sub-tasks: (1) character detection and co-reference resolution in the sentences, (2) character detection and co-reference in the image sequence, (3) alignment of textual and visual co-reference chains, and (4) ranking of the importance of each character. Figure 2 illustrates the proposed model architecture. Next we will introduce the four parts respectively.

Character Detection and Co-reference in Text

Given a textual story $\{\mathcal{S}_i\}_{i=1}^C$, we find all the mentions that refer to the same character as follows:

- Use a pre-trained part-of-speech tagger to identify all noun and pronoun words in the sentences.
- Filter these nouns using hypernyms in WordNet (Miller 1995). As the VIST stories contain people, animals and vehicles as characters, we used the following hypernyms: $\{person.n.01, animal.n.01, vehicle.n.01\}$.
- Use pre-trained co-reference resolution tools to group the mentions from step 2 into co-reference chains.

After these steps, we have the character co-reference chains $\{\mathcal{T}_i\}_{i=1}^{K_t}$, where \mathcal{T}_i denotes the co-reference chain of the i -th character in the story and K_t denotes the total number of textual characters detected in the input sentences.

Plural and Group Characters We separately detect the mentions for plural and group character. For plural characters, we consider words with the PoS tag *NNS* or *NNPS*. For group nouns, we create a vocabulary list that contains words whose PoS is not *NNS* or *NNPS*, but refer to a group of people (e.g., team, family). The full list is in the supplementary material. We add special labels to the chains referring to plural and group characters for the subsequent grounding step.

Character Detection and Co-reference in Images

For the input image sequence $\{I_i\}_{i=1}^C$, we identify characters and obtain visual character co-reference chains $\{\mathcal{V}_i\}_{i=1}^{K_v}$.

We experiment with two different approaches to obtaining face regions and features: (1) the pre-trained face detector MTCNN (Xiang and Zhu 2017) in conjunction with Inception Resnet (Szegedy et al. 2017) pretrained on VGGFace2 (Cao et al. 2018) for feature extraction, and (2) MTCCN for face detection combined with the CLIP pre-trained vision-language model (Radford et al. 2021) for feature extraction.

After obtaining the face features, we employ k-means (Lloyd 1982) on the face features to obtain the co-reference chains. k-means requires us to specify K , the number of clusters, before we run the algorithm. We experiment with $2 \leq k \leq 10$, as we assume a maximum of ten characters in a story and use the Calinski and Harabasz metric (Caliński and Harabasz 1974) to evaluate the visual co-reference results, and select the k with the highest score as the final cluster number K_v for the image sequence $\{I_i\}_{i=1}^C$. Note that some clustering algorithms (e.g., mean-shift) don’t require a predefined number of cluster. But in our experiments k-means outperforms all other algorithms we tried (details in the supplementary material).

Character Grounding

We aim to align textual co-reference chains $\{\mathcal{T}_i\}_{i=1}^{K_t}$ and visual co-reference chains $\{\mathcal{V}_i\}_{i=1}^{K_v}$ at this stage. We model character grounding as a bipartite graph matching problem. Specifically, we propose to first generate a similarity matrix

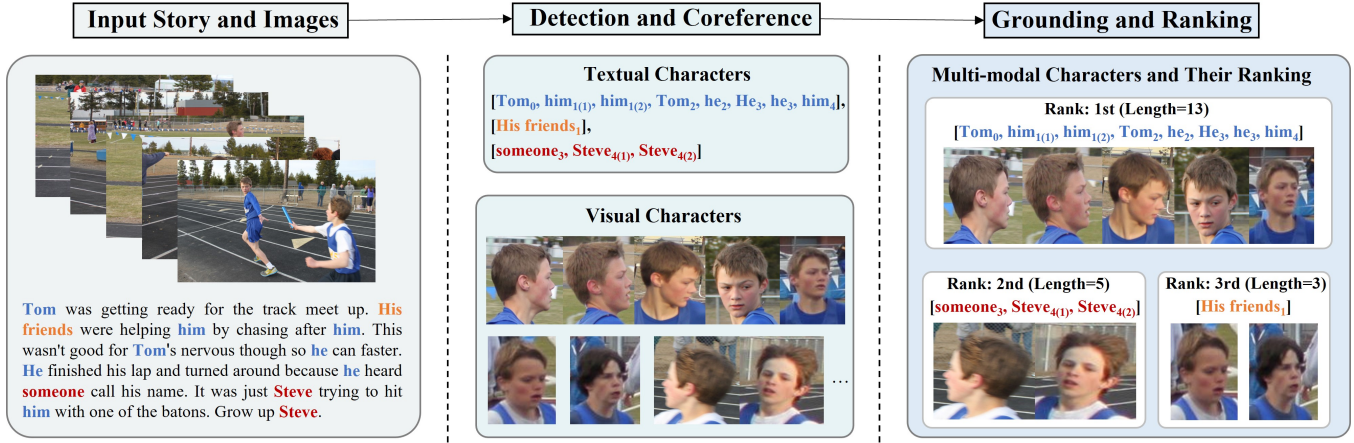


Figure 2: The architecture of the proposed model for important character detection and grounding in visual stories. Textual characters are color-coded. The subscripts of textual mentions represent their position in the story to distinguish identical words. Background characters in the images have been omitted in the figure for simplification.

Algorithm 1: Distributional Similarity-based Alignment

Input: Textual chains $\{\mathcal{T}_i\}_{i=1}^{K_t}$ and visual chains $\{\mathcal{V}_j\}_{j=1}^{K_v}$
Output: Alignment results \mathcal{R} .

- 1: Let $\mathcal{R} = []$ and $A =$ zero matrix with shape (K_t, K_v) .
- 2: **for** \mathcal{T}_i in $\{\mathcal{T}_i\}_{i=1}^{K_t}$ **do**
- 3: $t_i \leftarrow C$ -dim binary vector of the dist. of \mathcal{T}_i
- 4: **for** \mathcal{V}_j in $\{\mathcal{V}_j\}_{j=1}^{K_v}$ **do**
- 5: $v_j \leftarrow C$ -dim binary vector of the dist. of \mathcal{V}_j
- 6: $A_{ij} = \frac{t_i \cdot v_j}{|t_i| \cdot |v_j|} \{|\mathbf{x}|_1 = \sum_{r=1}^N x_r\}$
- 7: **end for**
- 8: **end for**
- 9: $\mathcal{R} = \text{Kuhn-Munkres}(A)$
- 10: **return** Alignment results \mathcal{R} .

A of shape (K_t, K_v) , where each element A_{ij} represents the similarity of textual chain \mathcal{T}_i and visual chain \mathcal{V}_j . Then we apply the Kuhn–Munkres algorithm, also known as the Hungarian algorithm, to the similarity matrix and obtain the matching results. We propose two methods for computing the similarity between \mathcal{T}_i and \mathcal{V}_j , namely using distributional similarity and using pre-trained vision-language model.

Distributional Similarity Algorithm 1 shows the details of the proposed distributional similarity-based alignment algorithm. The assumption is that the textual mentions and visual appearances of the same character should have a similar distribution across the five sentences and images in a story. For example, textual mentions in S_1 and S_2 and visual appearances in the corresponding images I_1 and I_2 tend to refer to the same character.

Mathematically, we use a C -dimensional binary vector (t_i and v_j) to represent the distribution of a textual or visual character. More precisely, $t_i[k] = 1$ when the i -th textual character is in the k -th sentence, and $v_j[k] = 1$ when the i -th visual character appears in the k -th image.

CLIP-based Similarity As described in the previous section, we use CLIP to measure the similarity of image-text pairs. CLIP is designed to handle one-to-one similarity. We take the average of the similarity of each mention in the textual chain $\{\mathcal{T}_i\}_{i=1}^{K_t}$ and each visual appearance in the visual chain $\{\mathcal{V}_j\}_{j=1}^{K_v}$ as the similarity of the two chains:

$$A_{ij} = \frac{1}{|\mathcal{T}_i| \cdot |\mathcal{V}_j|} \sum_{m=1}^{|\mathcal{T}_i|} \sum_{n=1}^{|\mathcal{V}_j|} \text{CLIP}(\mathcal{T}_i^m, \mathcal{V}_j^n) \quad (1)$$

where A_{ij} is the average similarity between $\{\mathcal{T}_i\}_{i=1}^{K_t}$ and $\{\mathcal{V}_j\}_{j=1}^{K_v}$, \mathcal{T}_i^m denotes the m -th mention in the textual coreference chain \mathcal{T}_i , \mathcal{V}_j^n denotes the n -th visual appearance in \mathcal{T}_i , and $|\cdot|$ means the length of a co-reference chain.

Plural and Group Characters We first ground individual characters as described so far. Then we deal with plural and group characters. For each plural/group textual coreference chain, we compute its similarity with each visual coreference chain using the same similarity method used for singular characters. We then select the best match plural/group textual character for each visual character. If the matching score is higher than a threshold, the visual character will be added to the plural/group co-reference chain.

After the above steps, we can obtain multi-modal coreference chains $\{\mathcal{M}_k\}_{k=1}^{K_m}$, where $\mathcal{M}_k = \{\mathcal{T}_i, \mathcal{V}_j\}$ if the textual character \mathcal{T}_i and the visual character \mathcal{V}_j are aligned. Note that the remaining uni-modal characters will also be included in $\{\mathcal{M}_k\}_{k=1}^{K_m}$. Therefore, K_m is equal to the maximum of K_t and K_v .

Importance Ranking

Intuitively, the more important a character is, the more crucial it is to the plot, and the more often it will be mentioned in the story. In Experiments section we analyse the relationship between character frequency and importance and find that the two factors are well correlated. We therefore use

Modality	Precision	Recall
Textual Story	74.4%	90.3%
Image Sequence	40.5%	69.1%

Table 3: Results for character detection in textual story and image sequence on VIST-Character.

count-based importance ranking, which means the importance of each character depends on the length of its multi-modal co-reference chain.

$$Importance(\mathcal{M}_k) = |\mathcal{M}_k| = |\mathcal{T}_i| + |\mathcal{V}_j| \quad (2)$$

In addition, we’re also interested in the importance ranking task in a single modality setting, i.e., when the input is the story text or the image sequence only. For text-only input, the importance of each character is $|\mathcal{T}_i|$ and for image-only input, it is $|\mathcal{V}_i|$.

Experiments

Character Detection

We use the spaCy PoS tagger (<https://spacy.io/api/tagger>) to obtain the nouns in text and then identify the characters using WordNet (Miller 1995). For images, we obtain the faces using MTCNN (Zhang et al. 2016) and resize them to 160×160. Table 3 shows the results of character detection.

Evaluation Metric We use precision and recall to evaluate textual and visual character detection. For textual character detection, a predicted character phrase is considered as a correct character detection when the head word of the noun phrase is the same as the gold-standard one. For visual character detection, a predicted face region is considered as a correct character detection when the bounding box of the face is entirely inside of an annotated body bounding box.

Analysis Overall, recall scores are higher than precision scores for both modalities. The precision of image-based character detection is only 40.5% and the gap with recall is around 28.6%. The reason is that there are a lot of background characters unrelated to the story in the images. Comparatively, text is less noisy, and we can observe better character detection performance than in images. However, we found that using WordNet and the hypernym method to filter character words can cause false positives. For example, *white* is usually used to describe color, but its primary sense in WordNet is a person name. Also, some character words like *great-grandmother* are not included in WordNet.

Character Co-reference

For textual stories, we experiment with two different approaches to obtain co-reference chains: (1) NeuralCoref (Clark and Manning 2016) and (2) SpanBERT (Joshi et al. 2020). For image sequences, we use pre-trained MTCNN for face detection and compare two methods for face feature extraction: (1) Inception ResNet pre-trained on VGGFace2 and (2) the pre-trained CLIP model. The face instances are then clustered using k-means to obtain visual co-reference chains. Table 4 presents the character co-reference results for both modalities.

Model	B-Cubed		Exact Match	
	Precision	Recall	Precision	Recall
NeuralCoref	70.2%	66.6%	29.2%	32.4%
SpanBERT	66.6%	70.0%	30.0%	33.2%
ResNet	57.0%	60.5%	10.7%	17.5%
CLIP	51.8%	60.8%	25.8%	23.8%

Table 4: Results for character co-reference. For the textual models, the input is the character mentions detected by PoS + WordNet. For the visual models, face instances are obtained by MTCNN and k-means is used for clustering.

Input	Model	Recall	Precision
Predictions	Distribution-based	27.3%	27.5%
	CLIP-based	21.7%	28.5%
Gold-standard	Distribution-based	77.5%	-
	CLIP-based	58.5%	-

Table 5: Performance of distribution similarity-based and CLIP-based alignment with and without the gold standard annotations as input. Best results are highlighted in bold.

Evaluation Metric Each character is represented by a co-reference chain. We evaluate the co-reference chains from two perspectives. B-Cubed recall and precision (following Cai and Strube 2010) indicate the average percentage of the correctly detected mentions in a chain. In contrast, exact match precision and recall require that the two chains are exactly the same.

Analysis In general, B-Cubed scores are higher than exact match, which means that co-referred mentions can be successfully detected, but it is difficult to identify the whole chain correctly. The reason is twofold. First, the fact that we are working on multi-sentence stories increases the difficulty of textual co-reference resolution. Second, the image sequences in the VIST dataset usually contain different views of the same character, which makes it hard to form accurate clusters, see Figure 2 for examples.

Comparing modalities, we observe that co-reference resolution in text performs better than in images across the board. We attribute this to the redundant characters detected in images, i.e., characters that are depicted but do not play a role in the story (e.g., people in the background).

For text-based models, SpanBERT outperforms NeuralCoref in terms of exact match, while for visual models, CLIP outperforms ResNet. In further experiments, we will use SpanBERT and CLIP, respectively.

Character Grounding

We evaluate distributional similarity-based and CLIP-based alignment applied on the output of the SpanBERT and MTCNN + CLIP pipelines. In addition, we report the result of the alignment algorithm on gold standard input to obtain an estimate of its performance without accumulated error. The results are shown in Table 5.

Evaluation Metric We use recall and precision to evaluate the character grounding task. Note that recall is preferred

Story Type	Pearson’s Correlation
Textual Stories	0.61
Visual Stories	0.55
Multi-modal Stories	0.62

Table 6: Pearson’s correlation between the number of occurrences and the importance ranking in different story modalities in the gold-standard VIST-Character dataset.

to precision in this task, as visual character detection returns redundant characters (see above), which depresses the precision score.

Analysis We can observe that the performance is much better with gold-standard input than when detection-and-co-reference is used. This indicates that the alignment algorithm works well, but is very sensitive to errors in the input. The errors of previous components accumulate and have a negative impact on the alignment performance.

Interestingly, CLIP does not outperform the distribution-based model. The reason is that CLIP is pre-trained on image-caption pairs whose the caption is usually specific and detailed. However, the VIST stories contain many generic words (e.g., *he*, *boy*) that appear frequently in the textual co-reference chains, making it difficult for CLIP to predict the alignment relation. See the supplementary material for more details on the CLIP-based grounding results.

Importance Ranking

Intuitively, there is a positive relationship between the importance of a character and its frequency: characters that are central to the plot are mentioned more often than side characters that don’t contribute much to the story.

To verify the assumption that character importance is related to frequency, we compute the Pearson’s correlation between the number of occurrences of a character in a story (i.e., the length of gold-standard co-reference chain), and the importance rating assigned to that character by the annotators. Specifically, we investigate the correlation between a character’s importance rating and (1) the number of occurrences in story text only, (2) the number of occurrences in image sequence only, and (3) the number of occurrences in both the image sequence and associated story text. Note that the correlation coefficient is uncomputable when there is only one character or all characters appear equally often in a story, so these instances were removed. The result is shown in Table 6 and confirms our hypothesis that character importance is correlated with character frequency.

We now turn to the most important character in a story, the protagonist. Being able to identify the protagonist is particularly important for story understanding or generation, as the plot revolves around this character. In our setting, the protagonist is expected to be the character with the highest frequency. We can therefore predict the protagonist as the character with the longest co-reference chain. We again compare text-only, image-only and multi-modal settings. We use SpanBERT for the text-only setting and use MTCNN + CLIP for the last two. The results are given in Table 7.

Model	P@1	P@3	P@5
Text-only	53.6%	79.2%	87.4%
Image-only	30.0%	30.8%	31.3%
Multi-modal	29.4%	34.7%	34.3%

Table 7: Performance for identifying the most important characters of a story. P@k means Precision@k. P@1 corresponds to identifying the protagonist. Multi-modal means that both the text and the images of a story are used.

Evaluation Metric We use the metric Precision@k to measure performance in predicting the most important characters. Precision@1 corresponds to predicting the protagonist correctly, while Precision@3 or Precision@5 indicate how many of the three or five most important characters have been identified correctly, compared to the human ratings of character importance in our VIST-Character dataset. Note that for the multi-modal model, the precision score is calculated based on the match of multi-modal co-reference chains. Two multi-modal co-reference match if the head words (first textual mention) are the same.

Analysis The text-based model achieves the best performance among the three settings with Precision@5 reaching 87.4%. The reason is that the story text is treated as the evidence based on which annotators have rated character importance in our dataset. Moreover, the noise in the story text is less than in the image sequence, which makes the text-based importance ranking less error-prone than ranking based on images.

The results of the image-only model are significantly worse, with precision scores of only around 30%. The reasons are two-fold. Firstly, the performance of clustering for co-reference resolution is not sufficiently high and thus adds a lot of noise to the downstream ranking task. Secondly, the current evaluation method only considers the head image of the visual character co-reference chain. Therefore, better co-reference resolution methods and more reasonable evaluation methods are needed to improve the image-only setting.

We can observe that the multi-modal setting performs slightly better than visual setting in terms of Precision@3 and Precision@5. This means that the injection of textual evidence helps improve the performance of the image-only model somewhat. From another perspective, the injection of visual evidence degrades the performance of the text-only model as the redundant characters in images bring a lot of noise to the length of the multi-modal co-reference chains if aligned with the wrong textual characters. But we don’t think the current method realizes the full potential of the multi-modal model given the large gap between the performance of text-only and multi-modal settings.

Plural and Group Characters

Table 8 shows the character grounding and importance ranking performance for multi-modal input with and without plural and group characters. For character grounding, the precision of plural and group characters reaches 47.5% while the recall is only 20.9%. The reason is our grounding algorithm

Input	Grounding		Ranking	
	Precision	Recall	P@1	P@5
Singular	34.5%	21.5%	33.2%	32.4%
Plural	20.9%	47.5%	-	-
Both	27.3%	27.5%	27.9%	37.9%

Table 8: The character grounding and importance ranking performance of multi-modal input with and without plural and group characters. P@k means Precision@k.

sets a high threshold (i.e., 0.6) for linking a visual character to a plural or group textual character. This threshold prevents the model from overpredicting plural characters, given that these are relatively rare in the VIST stories.

For important ranking, we observe that Precision@1 drops by five points with plural and group characters, while Precision@5 increases by about five points. The reason is that the protagonist is not a plural or group character in most cases. Therefore, including plural and group nouns makes the most important character prediction more error prone. However, the top five characters usually contain plural or group characters, thus Precision@5 will be higher when we consider plural and group characters.

Related Work

Visual Storytelling Visual storytelling was first proposed by Huang et al. (2016) and has generated a lot of interest since, starting with early work (Gonzalez-Rico and Fuentes-Pineda 2018; Kim et al. 2018) that employed a simple encoder-decoder structure, with a CNN to extract visual features and an RNN to generate text. More recent visual storytelling approaches (Xu et al. 2021; Chen et al. 2021; Hsu et al. 2020; Yang et al. 2019) introduce external knowledge to give models the necessary commonsense to reason. Sometimes, scene graphs are used to model the relations between objects (Lu et al. 2016; Hong et al. 2020; Wang et al. 2020). However, none of the existing approaches explicitly consider character information – characters are just treated like other objects. This results in a coherent sequence of events rather than a story with a clear plot driven by strong characters. Our work makes it possible to address this limitation by training character-centric models.

Detecting Important People in Images There is some prior work on detecting important people in single still images. Li, Li, and Zheng (2018) contribute the multi-scene dataset and an NCAA Basketball dataset for important people detection in single images. Li, Hong, and Zheng (2019) propose a deep importance relation network that learns the relations between the characters in an image to infer the most important person. Our work deals with image sequences, making important person detection more difficult as it requires re-identification and co-reference resolution.

Character-grounded Video Description There is a long line of work which aims to link character mentions in movie or TV scripts with their visual tracks (Yu et al. 2020; Rohrbach et al. 2017; Bojanowski et al. 2013; Tapaswi,

Bäumel, and Stiefelhagen 2012; Everingham, Sivic, and Zisserman 2006; Sivic, Everingham, and Zisserman 2009). Important datasets for this task include ActivityNet Captions (Caba Heilbron et al. 2015) augmented with bounding boxed and noun phrase grounding (Zhou et al. 2019). Our plural and group character annotation is inspired by their annotation guidelines. The M-VAD Names dataset (Pini et al. 2019) extends the Montreal Video Annotation Dataset (Torabi et al. 2015) with character names and corresponding face tracks, but does not include pronouns or plural nouns on the language side. Rohrbach et al. (2017) propose the MPII-MD Co-ref+Gender dataset, which contains the co-reference and grounding information for singular characters. Compared with this work, our dataset includes plural/group character annotations and richer visual information in terms of the number of bounding boxes (4,989 vs. 2,649). Also, none of the existing datasets considers the importance of characters, making it difficult to distinguish protagonists from side characters. In modeling terms, Rohrbach et al. (2017) handle video description jointly with character grounding and co-reference and Yu et al. (2020) learn visual track embeddings and textualized character embedding to achieve character grounding and re-identification.

Character-Centred Story Research This line of work aims to understand how characters drive the development of a story, resulting in models of character detection and persona learning (Bamman, O’Connor, and Smith 2013; Bamman, Underwood, and Smith 2014; Vala et al. 2015; Flekova and Gurevych 2015) and model of the interactions between characters (Iyyer et al. 2016; Srivastava, Chaturvedi, and Mitchell 2016; Chaturvedi, Iyyer, and Daume III 2017; Kim and Klinger 2019). Surikuchi and Laaksonen (2019) extract and analyse the character mentions from the VIST dataset. Our approach to character mention detection is inspired by their work. Other work has modeled the emotional trajectory of the protagonist (Brahman and Chaturvedi 2020). The character-centric storytelling model of Liu et al. (2020) uses character-embeddings to guide the generation process. Brahman et al. (2021) present a dataset of literary pieces paired with descriptions of the characters in them. In contrast to our work, all these approaches work on text only.

Conclusion

We introduced the VIST-Character dataset, which extends the test set of VIST with co-reference chains for textual and visual character mentions and their importance rating. Utilizing this dataset, we proposed the important character detection and grounding task, which requires character detection, co-reference, visual grounding and ranking. We proposed two simple, unsupervised models for this task: one using distributional similarity and one based on the large pre-trained vision-language model CLIP. In the future, we will exploit the proposed models to build a character-centred visual storytelling model, and extend it to longer and complicated narratives like comic books or movies. We will release the dataset and codebase of the project, and hope this will benefit work in character-centric story understanding and generation.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology Language Sciences.

References

- Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 352–361.
- Bamman, D.; Underwood, T.; and Smith, N. A. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 370–379.
- Bojanowski, P.; Bach, F.; Laptev, I.; Ponce, J.; Schmid, C.; and Sivic, J. 2013. Finding actors and actions in movies. In *Proceedings of the IEEE international conference on computer vision*, 2280–2287.
- Brahman, F.; and Chaturvedi, S. 2020. Modeling Protagonist Emotions for Emotion-Aware Storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5277–5294. Online: Association for Computational Linguistics.
- Brahman, F.; Huang, M.; Tafjord, O.; Zhao, C.; Sachan, M.; and Chaturvedi, S. 2021. “Let Your Characters Tell Their Story”: A Dataset for Character-Centric Narrative Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1734–1752. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Caba Heilbron, F.; Escorcía, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–970.
- Cai, J.; and Strube, M. 2010. Evaluation Metrics For End-to-End Coreference Resolution Systems. In *Proceedings of the SIGDIAL 2010 Conference*, 28–36. Tokyo, Japan: Association for Computational Linguistics.
- Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1): 1–27.
- Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.
- Chaturvedi, S.; Iyyer, M.; and Daume III, H. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Chen, H.; Huang, Y.; Takamura, H.; and Nakayama, H. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 999–1008.
- Clark, K.; and Manning, C. D. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.
- Everingham, M.; Sivic, J.; and Zisserman, A. 2006. Hello! My name is... Buffy”—Automatic Naming of Characters in TV Video. In *BMVC*, volume 2, 6.
- Flekova, L.; and Gurevych, I. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1805–1816.
- Gonzalez-Rico, D.; and Fuentes-Pineda, G. 2018. Contextualize, show and tell: A neural visual storyteller. *arXiv preprint arXiv:1806.00738*.
- Hong, X.; Shetty, R.; Sayeed, A.; Mehra, K.; Demberg, V.; and Schiele, B. 2020. Diverse and Relevant Visual Storytelling with Scene Graph Embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 420–430.
- Hsu, C.-C.; Chen, Z.-Y.; Hsu, C.-Y.; Li, C.-C.; Lin, T.-Y.; Huang, T.-H.; and Ku, L.-W. 2020. Knowledge-enriched visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7952–7960.
- Hsu, C.-y.; Chu, Y.-W.; Huang, T.-H.; and Ku, L.-W. 2021. Plot and Rework: Modeling Storylines for Visual Storytelling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4443–4453.
- Huang, T.-H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1233–1239.
- Inoue, N.; Pethe, C.; Kim, A.; and Skiena, S. 2022. Learning and Evaluating Character Representations in Novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, 1008–1019.
- Iyyer, M.; Guha, A.; Chaturvedi, S.; Boyd-Graber, J.; and Daumé III, H. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1534–1544.
- Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8: 64–77.
- Kim, E.; and Klinger, R. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. *arXiv preprint arXiv:1903.12453*.
- Kim, T.; Heo, M.-O.; Son, S.; Park, K.-W.; and Zhang, B.-T. 2018. Glac net: Glocal attention cascading networks

- for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*.
- Li, W.-H.; Hong, F.-T.; and Zheng, W.-S. 2019. Learning to learn relation for important people detection in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5003–5011.
- Li, W.-H.; Li, B.; and Zheng, W.-S. 2018. Personrank: Detecting important people in images. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 234–241. IEEE.
- Liu, D.; Li, J.; Yu, M.-H.; Huang, Z.; Liu, G.; Zhao, D.; and Yan, R. 2020. A character-centric neural model for automated story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1725–1732.
- Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137.
- Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *Euro-pean conference on computer vision*, 852–869. Springer.
- Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.
- Pini, S.; Cornia, M.; Bolelli, F.; Baraldi, L.; and Cucchiara, R. 2019. M-VAD names: a dataset for video captioning with naming. *Multimedia Tools and Applications*, 78(10): 14007–14027.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Rohrbach, A.; Rohrbach, M.; Tang, S.; Joon Oh, S.; and Schiele, B. 2017. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4979–4989.
- Sivic, J.; Everingham, M.; and Zisserman, A. 2009. “Who are you?”-Learning person specific classifiers from video. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1145–1152. IEEE.
- Srivastava, S.; Chaturvedi, S.; and Mitchell, T. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Surikuchi, A.; and Laaksonen, J. 2019. Character-Centric Storytelling. *arXiv preprint arXiv:1909.07863*.
- Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- Tapaswi, M.; Bäuml, M.; and Stiefelwagen, R. 2012. “Knock! Knock! Who is it?” probabilistic person identification in TV-series. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2658–2665. IEEE.
- Torabi, A.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*.
- Vala, H.; Jurgens, D.; Piper, A.; and Ruths, D. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 769–774.
- Wang, R.; Wei, Z.; Li, P.; Zhang, Q.; and Huang, X. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9185–9192.
- Xiang, J.; and Zhu, G. 2017. Joint face detection and facial expression recognition with MTCNN. In *2017 4th international conference on information science and control engineering (ICISCE)*, 424–427. IEEE.
- Xu, C.; Yang, M.; Li, C.; Shen, Y.; Ao, X.; and Xu, R. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3022–3029.
- Yang, P.; Luo, F.; Chen, P.; Li, L.; Yin, Z.; He, X.; and Sun, X. 2019. Knowledgeable Storyteller: A Commonsense-Driven Generative Model for Visual Storytelling. In *IJCAI*, 5356–5362.
- Yu, Y.; Kim, J.; Yun, H.; Chung, J.; and Kim, G. 2020. Character Grounding and Re-identification in Story of Videos and Text Descriptions. In *European Conference on Computer Vision*, 543–559. Springer.
- Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.
- Zhou, L.; Kalantidis, Y.; Chen, X.; Corso, J. J.; and Rohrbach, M. 2019. Grounded video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6578–6587.