Annotating Roles and Alignments in Folklore

Autumn Sehy
Brandeis University
autumnsehy@brandeis.edu

Ben Lambright
Brandeis University
blambright@brandeis.edu

Chris Tam
Brandeis University
christophertam@brandeis.edu

Abstract

Folklore worldwide often uses flora (plants) and fauna (animals) to portray characters as different archetypes. It is our goal to see if those archetypes are or are not consistent across cultures. We present our efforts to annotate flora and fauna spans in these stories with associated character attributes: roles (protagonist, antagonist) and alignments (good, bad). The result is a small gold dataset that is useful as a starting point for literary and cross-cultural data analysis. Here, we will report our project development annotation efforts, dataset statistics, and machine learning baselines.

1 Introduction

Folklore often uses flora and fauna to convey moral lessons or share ecological information. Our motivation is to find patterns and trends for how cultures use certain animals and how language is used to characterize them to uncover these trends or ecological information. Doing so can allow us to potentially track the movement of humans and these stories as well. Our annotation task was to identify the role (protagonist, antagonist) and alignment (good, bad) of flora and fauna in a curated dataset of manually collected stories. There has been little computationally related work in this field, however this used to be a more popular field pre-computers in the 1990s. There is still a group at Duke researching this as well, but in humanities. We believe using computational methods may help provide a new tool for this research, as it alleviates the task of reading thousands of stories to search for global trends.

Within the field of cross-cultural folklore there is one document known as the go-to in the field, the "Motif Index of Folk Literature" by Stith Thompson from 1955 (Thompson, 1955). It has over 200 archetypes, which we immediately recognized as too many for this task. We dug into the archetypes and distilled them to their most basic to create the roles and alignments framework. There is also another group out there using NLP studies have used a labelling framework called Proppian types that is more complex than what we currently have, but it was out of range for our capabilities in this assignment (Valls-Vargas et al., 2016) (Despontin et al., 2010).

2 Guidelines

The first iteration of guidelines started with a more complex and nuanced taxonomy of character archtypes. This was inspired from Thompson's work on archetypes, and included categories such as "The Trickster", "The Herald", or "Pleiades". We intented for annotators to tag the archetype of flora/fauna and then link the trigger words indicating their archetype to the tag. However, from our internal annotation work, we found this made the task time-consuming and difficult to train annotators for. Furthermore, the classes were sparse. We moved to a second iteration with fewer classes and simpler archetypes. This iteration also stopped linking trigger words. Yet our internal IAA still had poor results. Our main issue was lack of training to define and differentiate these archetypes, and time would not allow us to delve into the date enough to delineate them for the semester length to create clear responses for our annotators to create answers a model can predict.

Therefore we simplified more. In order to maintain some nuance we created, roles and alignments, which reduced our label set to two sets of three - Protagonist, Antagonist, Default (roles) and Good, Evil, Neutral (aligments). Our current guidelines include detailed definitions on these classes, including positive and negative examples from the dataset.

We were concerned our annotators would have issues tagging flora and fauna in the stories. Therefore, throughout the process, we maintained a preprocessor that would tag the animals under a taxon-

omy (de Queiroz, 1997). This is also motivated by the tags used by the naturalists at The Glacier Institute in Glacier National Park - the official education arm of the park. Bear and Wolf biologists inspired this taxonomy, as well as the original naturalists such as Darwin. It is also a taxonomy inspired by the constant communication and debate within naturalist circles - with some choices such as using "Ungulate" instea of the "Afrotheria" clade, which is now based in Africa and therefore not useful for our research purposes. This taxonomy combines animals and plants of similar types in order to simplify the number of flora/fauna classifications for our model. It also standardizes the spans.

This pre-processor had some minor issues, including tagging all exact matches as the flora/fauna it was looking for (i.e. tagging the verb "to bear" as the animal), and sometimes it missed tagging some key animals. Additionally, when every instance of an animal is tagged in a story, we found that this could be overwhelimg for annotators and cause them to miss animals, (i.e. when there are two ungulate animals, a pig and a deer, annotators would often notice one, but not both, due to them sharing the same tag). Even so, by standardizing spans and subcategories of animals, this pre-processor drastically improved annotation efficiency, and likely improved IAA.

3 Dataset

Our result is a gold dataset of 479 annotations across 122 story text files. This comprises our manually curated dataset of stories, sourced from Project Gutenberg (Okazaki, Seneca Myths and Folktales, etc) and cultural websites (e.g., www.native-languages.org), across the Cherokee, Filipino, Cherokee, Korean, Japanese, Seneca, and Maori cultures. Two trained annotators dually annotated each set. There avereaged about 4 flora/fauna to tag per document. Of these, the most common flora/fauna by far in the data were trees (61), then ungulates, crops, canines,and birds. This is shown in the figure below.

The majority of our tags had a 'neutral' alignment and 'default' role. We expected this, as when reading the stories we found many characters were either tricksters or unimportant to the narratives. For the final dataset tags, 66 characters were good, 40 evil, and 297 were neutral. Of these, 66 tags were also protagonists (though not all protagonists were always good/evil), 35 were antagonists, and

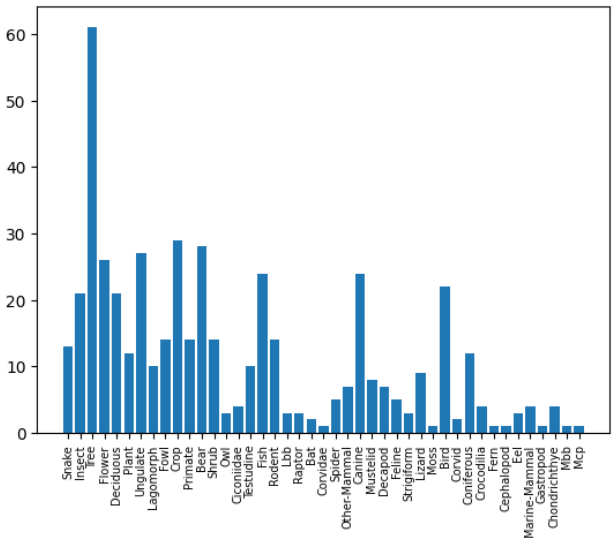

Figure 1: Distribution of labels tagged

302 were default. Importantly, almost all fauna were default and neutral with the exception of crops. Indeed, plants in our data significantly increased the amount of default neutral tags in the story. The following figure shows the distribution of flora/fauna that receieved tags that weren't default or neutral. Even with this, few flora/fauna had any alignment. When comparing this to Figure 1, one can see that flora/fauna get tagged as something other than default, neutral in less than half of all of their instances. Ideally, we would have annotated more data to have more instances of flora/fauna with non-neutral tags.

4 Inter-annotator Agreement and Adjudication

4.1 Adjudication

Flora and fauna entities were preannotated in brat with a string matcher and informed by a dataset of plant/animal names matched with their broad classification in our annotation schema (e.g., trees, fish, foxes, etc.). Annotators annotated spans that were identified with this step, but were not instructed to annotate or correct additional entities themselves.

Our annotators were split up into two pairs, Kasey/Emily and Keer/Joyce, from the emoji sentiment project. Each annotator annotated 61 stories across five cultures to annotate the entire dataset. Each of the cultures were split relatively equally among the annotators, so each annotator would have equal exposure to each culture. On average, each annotator tagged about 250 flora/fauna and each label took about one minute to tag. A figure

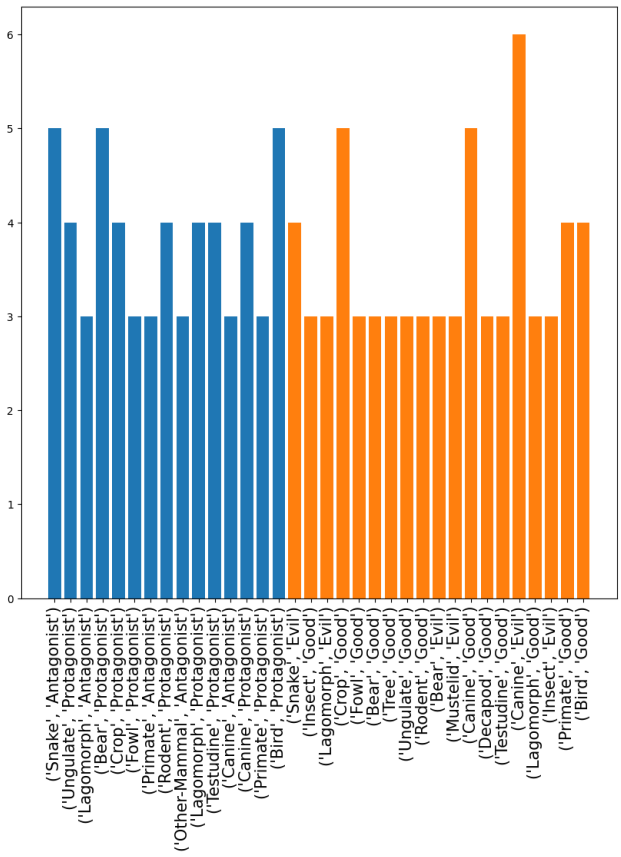

Figure 2: Distribution of labels which recieved at least three tags that weren't default or neutral

of how they were split can be see below. The idea behind this is that no annotator gets used to the common tropes of flora/fauna, so that there will be as little bias as possible when annotating the next culture. We split the annotators in pairs for the purpose of inter-annotator agreement.

	Keer and Joyce	Emily and Kasey
Cherokee	9	10
Filipino	11	10
Japenese	11	11
Maori	19	19
Seneca	11	11

Figure 3: Assigned documents per culture

It is important to note that it is not uncommon to have all documents be annotated by two annotators for more complicated forms of annotation even though it is more expensive. For example, the Brandeis-Simmons VP ellipsis corpus has every document annotated by two annotators, then has several rounds of adjudication.

We finalized our gold dataset after a round of adjudication across pairwise annotations. We had a series of rules we would follow when doing the adjudication.

1. We choose one of the two annotators to be the one we would copy for the gold data, any additional adjudication would just be modifications of this document (rules 2-3). This would be the annotator who labeled more flora/fauna in total (rather than forgetting to label them, which sometimes happened). The particular flora/fauna tag they chose to attribute also played a role, as we prefer gold data that has labels closer to words that would 'trigger' an attribute decision, such as seeing 'evil' or 'hero' in the context window of the attributed tag. Having a context window with such words is beneficial for training our model. This was of course in our guidelines, but some annotators were more proficient at this than others, with this being said they still often picked the same specific tag to give attributes.

2. Similarly to the IAA, any cases where one annotator did not label the flora/fauna but the other annotator did, we immediately chose the attributes that the one annotator did tag.

3. In any cases where the annotators did otherwise disagree, we would review the tags within the document, review our guidelines, and choose the attributes that best fit the tag ourselves. In almost all cases, this meant that one of the annotators was correct, and we would use that one for the gold data. In rare cases, we would decide that both annotators were wrong and we would choose an alternative, correct answer.

4.2 Inter-annotator Agreement

While developing our annotation schema, we also did several rounds of less formal inter-annotator agreement. However, not only had we not done a significant amount with the final annotation schema, but our annotation schema is still complicated. Determining whether or not flora or fauna is good or evil can be a very difficult thing, especially when reading non-Western stories when the reader has only been exposed to Western stories. Given this, we expected there to always be a fair amount of disagreement between annotators that

would need to be adjudicated, regardless of the quality of our guidelines.

IAA had interesting results, both across roles and alignments. Scores reported are averaged Cohen's kappa across our two pairs of dual annotators.

Surprisingly, calculating Cohen's kappa over the raw dually annotated data gave practically zero agreement for both role and alignment. It appears that most of the disagreement cases occurred where one entity was annotated, but was missed by the other annotator. We could consider this a shortcoming of our guidelines or annotation setup - we intended for annotators to label every flora/fauna. If we make the assumption that annotators would have annotated Default/Neutral rather than not annotate at all, agreement drastically improved to moderate agreement (role = .66, alignment = .58). Inspecting the outcome now now, we found that most agreements occurred when neither annotator labelled with a positive role/alignment - both annotations were Default/Neutral. Further filtering out these unimportant cases where both annotators agreed the character was completely neutral, agreement dropped to only fair (role = .33, alignment = .15). The complications here underscore the importance of making annotation span labels completely clear, as well as the pitfalls of including a "default" case, which creates a highly imbalanced dataset and inflates agreement scores.

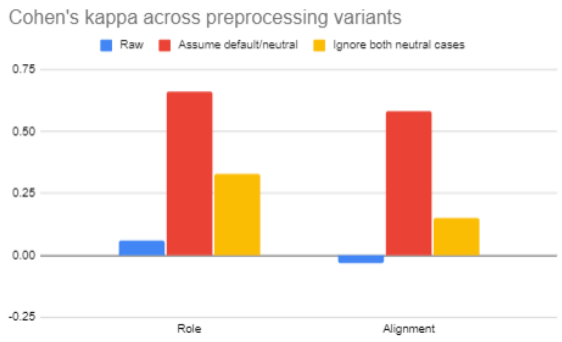

Figure 4: Kappa scores across different preprocessing interpretations

Regardless, we observe an interesting trend - alignments have a lower rate of agreement than roles. This is likely due to a combination of our guidelines (the annotators could argue that the instructions for labeling good/evil were less black and white), and the fact that it was difficult for annotators to determine the morality of flora/fauna in non-Western stories where these flora/fauna acted or had roles that our own Western culture might see differently to the morality of the culture to which they belong.

There were other cases where either the guidelines needed to be clearer or perhaps the guidelines should have been emphasized in different ways, such as the case of labeling multiple flora/fauna as protagonists. With the understanding that this annotation schema is complex and our guidelines had many rules, the guidelines clearly stated that in most cases there should only be one protagonist in the story, yet there was often annotator disagreement when one of the annotators opted to label multple flora/fauna as protagonists. On a similar note for common disagreements, there were many cases where sometimes annotators would either label flora/fauna as default neutral or not label the flora/fauna at all. In almost all cases where the flora/fauna were not labeled, their attributes should have been default neutral and our guidelines stated to label the tags as thus, but it is likely that the annotator either missed the tag or assumed that the flora/fauna played such an insignificant role that they did not have to be tagged, even though this is contrary to the guidelines.

5 Machine learning baseline

Due to the somewhat small size and low agreement of our dataset, as well as the fact that the majority of our data was tagged 'default' and 'neutral', indicating apparent difficulty, we opted to use LLMs to establish a baseline for our gold dataset as opposed to data-dependent models like bag-of-words logistic regression or RNNs. LLMs are known for their strengths in low-resource domains as they are pretrained, but it must be acknowledged that biases may exist in that pretrained data (Brown et al., 2020). We used OpenAI's GPT 3.5-turbo and set its temperature to 0 to ensure deterministic responses. We prompted the LLM with 50 tokens around the annotated spans as context, and asked it to classify a certain animal's role and alignment. Using this method, role acccuracy was 45%, alignment accuracy was 68%, and the F1 of the alignment predictions was 34% and the F1 of role predictions was 39%, performing decently above chance.

For each culture we tested the alignment accuracy, as this was our highest accuracy. They all averaged between 65% - 68% alignment with no outliers.

We tried some other ways to manipulate the

model. After increasing the context window to 500 characters on each side of the word, accuracy did not improve: accuracy of aligment: 67%, accuracy of roles: 40%, but lower F1 scores: 28% for alignments and only 11% for roles. Although recall was high (68-82%), precision was very low (14-29%). That said, while increasing the context window has the potential of making the model more accurate, we are really only using it to test the quality of our data, given that we used all of our gold data as test data for the model. If we assume that the model always gives perfect restults, this could also be a test of the quality of our gold data. In such a case, our data still performed decently above chance and so was better than no data at all for this model.

6 Future directions

Though our work resulted in a well-crafted gold dataset of culturally accurate stories, opening exciting new directions for folklore and narrative research, it suffers from two main weakness: its small dataset size and low inter-annotator agreement scores. With this in mind, our work could benefit from further collection of stories around the world, as well as more iterations of the MAMA cycle (model-annotate, model-annotate) to improve guidelines. However, there was high accuracy with the protagonist/antagonist depictions. Therefore, it could be a promising lead to follow - if protagonists and antagonists are clear in folklore we can see if there are global trends in this regard. We could also try to apply the trained program to social media or medical papers to see if it can pull antagonists/protagonists from these texts using training from folkloric literature alone.

However, to achieve this we'd require better guidelines to differentiate between antagonist/protagonist. We'd also need many more annotated stories.

Further improvements to our guidelines based on this first round likely include clearer examples and nuance to distinguish between edge cases, perhaps taking more inspiration from Proppian archetypes. Finally, we could improve the automatic flora/fauna entity labeler with context-dependent understanding. A few creatures were missing, and the incorrect annotating of 'bear' and 'rose' as nouns when they were often verbs made annotating annoying.

We also plan on running more analyses on metadata, like culture of origin, but due to time constraints, we are not adding those to this final assignment.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin de Queiroz. 1997. The linnaean hierarchy and the evolutionization of taxonomy, with emphasis on the problem of nomenclature. *Aliso: A Journal of Systematic and Floristic Botany*, 15(2):125–144.

Marco Despontin, Licia Sbattella, and Roberto Tedesco. 2010. Natural language processing for storytelling and role playing: a training system based on the propp model. In *ICERI2010 Proceedings*, pages 5036–5045. IATED.

Stith Thompson. 1955. *Motif-index of Folk-literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-books, and Local Legends, Volume 1*, volume 1. Indiana University Press, 1955.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2016. Predicting proppian narrative functions from stories in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 12, pages 107–113.