A Straightforward Approach to Narratologically Grounded Character Identification

Labiba Jahan, Rahul Mittal, W. Victor H. Yarlott, Mark A. Finlayson
School of Computing and Information Sciences
Florida International University, Miami, FL 33199
{ljaha002, rmitt008, wyarl001, markaf}@fiu.edu

Abstract

One of the most fundamental elements of narrative is *character*: if we are to understand a narrative, we must be able to identify the characters of that narrative. Therefore, character identification is a critical task in narrative natural language understanding. Most prior work has lacked a narratologically grounded definition of character, instead relying on simplified or implicit definitions that do not capture essential distinctions between characters and other referents in narratives. In prior work we proposed a preliminary definition of character that was based in clear narratological principles: a character is an animate entity that is important to the plot. Here we flesh out this concept, demonstrate that it can be reliably annotated (0.78 Cohen's κ), and provide annotations of 170 narrative texts, drawn from 3 different corpora, containing 1,347 character co-reference chains and 21,999 non-character chains that include 3,937 animate chains. Furthermore, we have shown that a supervised classifier using a simple set of easily computable features can effectively identify these characters (overall F_1 of 0.90). A detailed error analysis shows that character identification is first and foremost affected by co-reference quality, and further, that the shorter a chain is the harder it is to effectively identify as a character. We release our code and data for the benefit of other researchers[1].

1 Introduction

Characters are some of the most central elements of narratives, and the concept of *character* plays an important role in most definitions of narrative. As an example, Monika Fludernik defines a narrative as "a representation of a possible world ...at whose centre there are *one or several protagonists* of an anthropomorphic nature ...who (mostly) perform goal-directed actions ..." (Fludernik, 2009, p.6; emphasis ours). This definition clearly states that characters are central to stories *per se*. Therefore, it is natural to assume that character identification is an important step in automatic approaches to story understanding.

A number of approaches have been proposed for automatically identifying characters. Some approaches, for example, have sought to solve the character identification task using domain-specific ontologies (Declerck et al., 2012) or reasoning by reference to an existing case base (Valls-Vargas et al., 2014). Others have taken supervised machine learning approaches (Calix et al., 2013; Barros et al., 2019), where a classifier is trained over data annotated by people. Some approaches, e.g., examining characters' social networks (Sack, 2013), take character identification for granted, implementing heuristic-driven approaches over named entities or coreference chains that are not examined for their efficacy. Regardless of approach, all prior work of which we are aware has, unfortunately, had a relatively impoverished concept of *character*, at least from a narratological point of view. In particular, a key aspect of any character is that it *contributes to the plot*—characters are not just any animate entity in the narrative—and all prior work essentially ignores this point. Here we build on a prior proposal of ours

[1]Code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/RB6ZH0.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6089–6100
Barcelona, Spain (Online), December 8-13, 2020

(Jahan and Finlayson, 2019) to incorporate this narratologically grounded definition of character into automatic character identification. We first define and operationalize our concept of character, and use that concept to generate annotated data (170 narrative texts drawn from 3 different corpora) with high inter-annotator agreement. Then we demonstrate a straightforward supervised machine learning model using seven features that performs quite well on these data. Our error analysis reveals several choke points in the performance of the system, most importantly the quality of the co-reference chains.

The paper proceeds as follows. First, we discuss the definition of the character as presented by narratologists, contrasting this concept with those used in prior computational work, and describe an operationalized concept of character that can be annotated with high inter-annotator agreement (§2). We next describe the data to which we applied this concept (§3), following which we discuss the experimental setup, including the features and classification model (§4). We present the results (§5) and analyze the error patterns of the system, discussing various aspects, which leads us to a discussion of future work (§6). Although we have discussed prior work briefly in the introduction, we summarize work related to this study (§7) before we conclude by enumerating our contributions (§8).

2 An Operationalized Concept of Character

2.1 Core Concept of Character

All prior work that tackles the character identification task is unified by it's lack of a clear, operationalized definition of *character*. So far the work that reports the best performance is by (Valls-Vargas et al., 2014), where they give examples of different types of characters such as humans, animals (e.g., a talking mouse), anthropomorphic objects (e.g., a magical oven, a talking river), fantastical creatures (e.g., goblins), and folkloristic characters (e.g., the Russian characters *Morozko* and *Baba Yaga*). Despite this relatively comprehensive list of character examples, they did not provide a procedure for reliably distinguishing characters from other animate entities in a narrative.

Consider the following example. Let's assume we have a story about Mary, a little girl, and her dog named Fido. Mary plays with Fido when she feels lonely. Also, Fido helps Mary in her daily chores and brings letters to Mary from the post office. One day Mary and Fido are walking through town observing the local color. They see a crowd gathered around a fruit vendor; an ugly man crosses the path in front of them; and another dog barks at Fido. Many narratologists and lay people would agree that the story has at least two characters, Mary and Fido. Depending on how the story is told, either Mary or Fido may be the protagonist. But what about the other entities mentioned in the story? What about the unnamed man who crosses their path? Is he a character? What about the faceless crowd? Is the crowd itself a character, or perhaps its constituent people? What about the fruit vendor, who is hawking his wares? And what about the barking dog? Where do we draw the line?

We noted these problems in prior work, and proposed a preliminary definition of character grounded in narrative theory that addressed these questions (Jahan and Finlayson, 2019). We began by studying different books and literature reviews on narratology that provided different definitions of character. Helpfully, Seymour Chatman, in his classic book "Story and Discourse: Narrative Structure in Fiction and Film" (1980), collected a number of views on character across multiple narratological traditions. Several of the definitions were complex and would be quite difficult to model computationally. Others were too vague to inform computational approaches. However, one definition provided a reasonable target:

> The view of the Formalists and (some) structuralists resemble Aristotle's in a striking way. They too argue that characters are products of plots, that their status is "functional," that they are, in short, participants or *actants* rather than *personnages*, that it is erroneous to consider them as real beings. Narrative theory, they say, must avoid psychological essences; aspects of character can only be "functions." They wish to analyze only what characters do in a story, not what they are—that is, "are" by some outside psychological or moral measure. Further, they maintain that the "spheres of action" in which a character moves are "comparatively small in number, typical and classable." (Chatman, 1980, p.111)

Here, an *actant* is something that plays (i.e., *acts in*) any of a set of active roles in a narrative, and *plot* denotes the main events of a story. This definition, then, though presented via somewhat obscure narratological terminology, gives a fairly conceptually concise definition of a character: a character is *an animate being that is important to the plot*. By this measure then, we are justified in identifying Mary and Fido as characters, but not the various entities they casually encounter in their stroll through town.

2.2 What Makes an Entity Important?

Our definition considers animate beings who can contribute to the plot as characters. But this definition leads to another problem, namely, how can we measure the importance of the characters? How much of a contribution is enough to be a character? Unfortunately, narratologists' answers are not especially clear, and indeed very few narratologists have attacked this question directly. As Chatman writes, "It is remarkable how little has been said about the theory of character in literary history and criticism." (1980, p.107). According to the famous cultural theorist and narratologist, Mieke Bal, it is difficult to explain the ideas of *character* because a character so closely resembles a human being. She writes "... no satisfying, coherent theory of character is available is due to this anthropomorphic aspect. The character is not a human being, but it resembles one." (Bal and Van Boheemen, 2009, p.113). Despite this, for the purposes of reliable inter-annotator agreement to support training and testing effective computational approaches, it is critical for us to define specific tests by which we can decide if an animate being is a character or not.

Chatman points out the importance of the *functionality* of a character with regard to the plot. We formulated our test by starting with the original theoretical work that led to the development of the theory of functionality and actants, namely Vladimir Propp's *Morphology of the Folktale* (1968). In that theory, Propp describes the concept of a *function*, which is an action or event that drives the plot forward, and is intimately interwoven with the main characters (i.e., the *dramatis personae*). For example, the Villain of a story may cause harm or injury to some member of the Hero's family: Propp names this plot function *Villainy* and assigns it the symbol *A*. He defined 31 such functions. Prior work on annotating Propp's morphology has shown that the main characters can be reliably identified (Yarlott and Finlayson, 2016), and so those characters which are directly and unambiguously involved in the forwarding of the plot are generally not difficult to identify. These main characters have numerous mentions, close involvement in the main events, and highly distinctive character traits. What about, however, edge cases—potential minor characters—such as the examples in the Mary and Fido above? Minor characters have many fewer mentions, little involvement with main events, and often no uniquely distinguishing traits.

To illustrate the difficulty consider the following example from Propp's data, namely the story *Vasilisa the Beautiful*, which is found in one of our corpora, the extended ProppLearner corpus (Finlayson, 2017). In this story, the heroine is Vasilisa, whose mother dies right after giving birth to her. Before dying, the mother gave Vasilisa a doll, and the rest of the story concerns how Vasilisa survives the predations of her stepmother with the help of that doll. There is no doubt that Vasilisa is a main character of this story—she is the Heroine—but there is some question about her birth mother. Does the birth mother count as a character, albeit a minor one? We can apply our test of functionality by asking whether the mother's actions or presence are critical to the progression of plot. In particular, the mother gives Vasilisa a critical magical artifact (the doll, which itself become a major character) without which Vasilisa would have been unable to effect much of the action of the story. Because of the mother's involvement, indirect though it may be, in key events of the plot, we can reasonably consider the birth mother a minor character.

In addition to the extended ProppLearn corpus, we also annotated texts from OntoNotes 5.0 (Weischedel et al., 2013) which presented many interesting edge cases. As an example, OntoNotes contains many short news texts, one consisting only of 13 lines about a day in the life of Bill Clinton just before the U.S. election of 2000. In that article "all Americans" is mentioned: "The day got worse when he urged all Americans to vote on November 2.". It is clear that Bill Clinton is a character of this news article because the whole story is about him, but what about the referent "all Americans"? Do they contribute to the "plot" of the article, such as it is? Do they support the development of the main character? In this case, "all Americans" neither effect any functional action in the plot of the article, nor do they

contribute anything necessary to the progression of the plot. Indeed, if the reference to "all Americans" was struck from the text, the plot would remain essentially unchanged. Based on this judgement, we do not consider "all Americans" to be a character, even a minor one.

Based on these examples, we can propose a rule for assessing the importance of an entity: if an animate entity is mentioned numerous times, has clear and close involvement in the main events of the plot, and has highly distinguishing character traits, then it is almost certainly a main character. For other animate entities that are mentioned less often, have more tangential connection to the plot, and perhaps lack distinguishing traits, the key test is whether that entity critically contributes to the plot either by directly participating in a important plot event, or enabling the participation of other characters in the plot. In our annotation, we observed that the difficulty of distinguishing characters from non-characters depends strongly on the length of a story. The shorter the text, the harder it is to identify the characters, primarily because there is much less opportunity for entities to present distinct characteristics and contribute clearly to the development of the plot. As a case in point, identifying characters in our third corpus, the Corpus of English Novels (Weischedel et al., 2013), where the chapters are quite long, was easier than identifying characters in the Propp's folktales, and substantially easier than in the short OntoNotes news texts.

2.3 Other Aspects of Characters

With a operationalized definition of character now in hand, one might ask whether characters can be further characterized along different dimensions. For example, Ismail Talib (2010) described a number of different possible dimensions of characters: protagonist vs. antagonist, flat vs. round, static vs. developing, and so forth. Propp described seven different types of *dramatis personae*: Hero, Villain, Princess, Helper, Donor, Dispatcher, and False Hero. While these are interesting directions to explore, in this work we did not seek to categorize entities in any way other than character or not.

3 Data and Annotation

We annotated characters on 170 texts across three corpora, one with 46 texts (the extended ProppLearner corpus), the second with 94 texts (a subset of the OntoNotes corpus), and the third with 30 texts (a subset of The Corpus of English Novels). Table 1 shows the counts of various items of interest across the data. We manually annotated these corpora as to whether each coreference chain acted as a character in the story. Gold coreference chains were already marked on the ProppLearner corpus and OntoNotes, while the coreference chains were automatically computed for the Corpus of English Novel. According to the definition mentioned above, we marked a chain as a character if it is animate and is important to the plot of the story. First, we read the story and find the events important to the plot, there was no agreement across the annotators what the events important to the plot are. Then we assessed the animate objects directly or indirectly involved those events to determine if they were characters or not. As our supervised model is highly dependent on the annotation, therefore, if there are more than one plot, or if the plot is highly subjective, then it should be reflected by the annotators.

Text Types	**# Texts**	**# Coref. Chains**	**#Ani. Chains**	**#Inani. Chains**	**#Char. Chains**	**#Non-Char. Chains**
The extended ProppLearner	46	4,950	2,004	2,946	564	4,386
OntoNotes	94	1,145	472	673	347	798
Corpus of English Novels	30	17,251	2,808	14,443	436	16,815
Total	170	23,346	5,284	18,062	1,347	21,999

Table 1: Counts of various text types in the corpus.

The extended ProppLearner (Finlayson, 2017) contains gold-standard annotations for referring expressions, coreference chains, and animacy. It comprises 46 Russian folktales originally collected in Russia in the late 1800s but translated into English within the past 70 years.

The first two authors double annotated this corpus at the coreference chain level for character, achieving an agreement of 0.78 Cohen's kappa (κ). This level of agreement represents substantial overall agreement (Landis and Koch, 1977). The authors discussed any disagreements and corrected them to generate

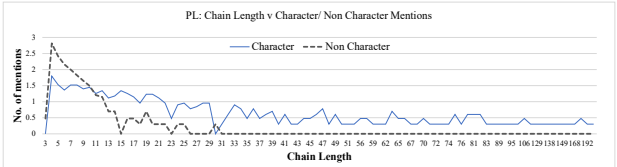

Figure 1: Mentions vs. chain length: PL corpus

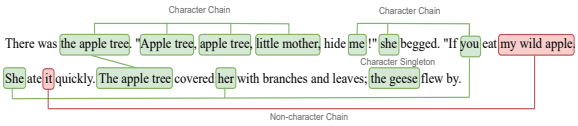

Figure 2: Sample text fragment of PL corpus

a gold-standard annotation. Our high agreement measures are in accordance with prior work that has shown that *dramatis personae* (i.e., main characters) can be annotated with high reliability. In particular, Yarlott and Finlayson (2016) showed that dramatics personae can be annotated with agreements of $F_1 > 0.8$ and $\kappa > 0.6$. Because of the high agreement for this annotation task, we single-annotated the remaining two corpora for the sake of efficiency.

OntoNotes (Weischedel et al., 2013) is a large corpus containing a variety of genres, including news, conversational telephone speech, broadcast news transcripts, talk show transcripts, among others, in English, Chinese, and Arabic. We extracted 94 English broadcast news transcripts that had gold-standard coreference chain annotations. The first author annotated the coreference chains as to character. Despite having clear narrative elements, including characters and events, the news texts have very different goals and textual properties. For example, the plot is only partially represented in a news text, while we have a full plot in many narrative texts.

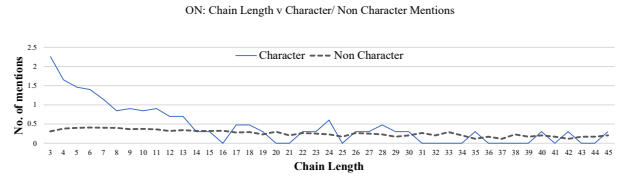

Figure 3: Mentions vs. chain length: ON corpus

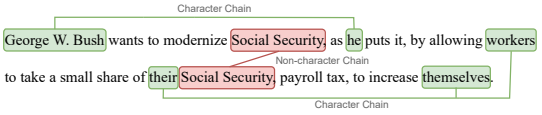

Figure 4: Sample text fragment of ON corpus

The Corpus of English Novels (CEN) (De Smet, 2008) contains 292 English novels written between 1881 and 1922, comprising various genres, including drama, romance, fantasy, adventure, etc. We selected 30 novels and from each extracted a single chapter that contained a significant number of characters. We computed coreference chains using Stanford CoreNLP (Manning et al., 2014), and the first author annotated those chains as to character. We annotated only one chapter per novel due to time constraints: we are aware that in a full novel, the picture might be very different than in a single chapter.

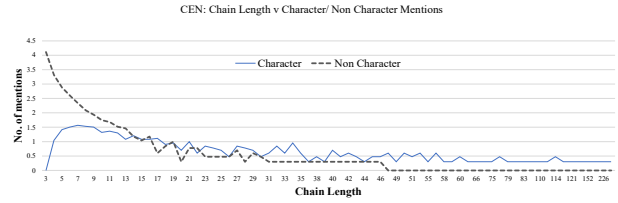

Figure 5: Mentions vs. chain: CEN corpus

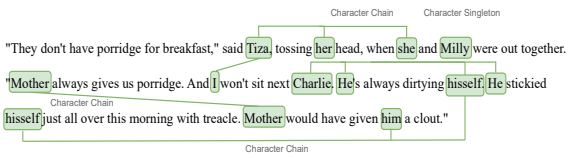

Figure 6: Sample text fragment of CEN corpus

4 Approach

Our character detection model comprises two steps: first, we automatically mark the animacy of coreference chains, and second we apply a supervised machine learning classifier to identify the characters.

4.1 Step 1: Animacy Detection

According to our definition of character, it must be an animate object that is important to the plot. Thus one first step to identifying characters is to identify the animate entities. We used an existing animacy classifier for coreference chains (Jahan et al., 2018), and tried two of their best-performing models, both of which achieved state-of-the-art performance; one is a hybrid model incorporating supervised machine learning and hand-built rules, and the other is a rule-based model consisting of hand-built rules only. As

we have gold standard animacy annotation in the extended ProppLearner corpus that allows training the supervised portion of the hybrid model, we trained and ran the hybrid model on this data. For OntoNotes and the Corpus of English Novels, we ran the rule-based model, which did not require gold-standard animacy markings for training, to detect animacy.

4.2 Step 2: Character Classification

4.2.1 Features

We explored seven different integer and binary features to train the character identification model. As we have mentioned earlier, not all animate entities are characters, but all characters are animate entities. Therefore, we incorporated the animacy features while adding additional features for character, and so most of the features are designed to interrogate whether an animate entity acts as a semantic subject of an event or has person-like characteristics. Some of the features are drawn or inspired by prior work.

1. **Coreference Chain Length (CL)**: We computed the length of a coreference chain and then normalized the numeric length feature by z score $= (x - \mu)/\sigma$, where x is the raw chain length, μ is the chain length mean, and σ is the chain length standard deviation. This feature explicitly captures the tendency of the long chains to be characters, as discussed in prior work (Eisenberg and Finlayson, 2017).

2. **Semantic Subject (SS)**: We also computed whether or not the head of a coreference chain appeared as a semantic subject (ARG0) to a verb, and encoded this as a boolean feature. We used the semantic role labeler associated with the Story Workbench annotation tool (Finlayson, 2008; Finlayson, 2011) to compute semantic roles for all the verbs in the stories. Semantics roles have been previously used for Named Entity Recognition (NER) as seen in (Pang and Fan, 2009)

3. **Named Entity (NE)**: We checked whether or not the head of a coreference chain was a named entity with the category *PERSON*, and encoded this as a boolean feature. The named entities were computed using the standard API of the Stanford dependency parse (Manning et al., 2014, v3.7.0).

4. **WordNet (WN)**: We detected if the head of a coreference chain is a descendant of *Person* in WordNet, and encoded this as a boolean feature.

5. **Dependency Link (DP)**: We computed whether or not the head of a coreference chain appeared as a dependent of `nsubj` dependency link among the enhanced-plus-plus-dependencies of a sentence. The dependencies were extracted using the standard API of the Stanford dependency parse (Manning et al., 2014, v3.7.0) we have used for Named Entity feature. Similar dependencies were used as features elsewhere (Valls-Vargas et al., 2014).

6. **Triple (TP)**: We computed if the head of a coreference chain matches the subject position of any triple and encoded this information as a boolean feature. The triples were extracted from Stanford OpenIE associated with the classic API of the Stanford CoreNLP toolkit (Manning et al., 2014, v3.7.0). (Goh et al., 2012a) used a similar extraction of an S-V-O triplet.

7. **ConceptNet Feature (CN)**: We checked if the head of a coreference chain has any edge that related to *Person* in the ConceptNet semantic network (Speer et al., 2017) and encoded this information as a boolean feature. Features extracted from ConceptNet have also been used as features elsewhere (Calix et al., 2013; Valls-Vargas et al., 2014).

4.2.2 Model

Our character classification model is a simple supervised machine learning classifier with the hand-built features identified above. We used the extended ProppLearner corpus to explore different combinations of features and their importance to model performance. The best-performing model uses all seven features. We then trained and tested this model to the OntoNotes and Corpus of English Novels corpora to see how our model works on different kinds of data sets. The implementation of our model is done by using an SVM (Chang and Lin, 2011) with a Radial Basis Function Kernel[2]. We have demonstrated the results on different corpora in Table 2. We trained each model using ten-fold cross-validation, and report macro-averages across the performance on the test folds.

[2]SVM parameters were set at $\gamma = 1$, $C = 0.5$ and $p = 1$.

Corpus	Feature Set	Acc.	Non Character				Character			
			κ	Prec.	Rec.	F_1	κ	Prec.	Rec.	F_1
Propp-Learner	Baseline MFC	70%	0.0	0.70	0.1	0.82	0.0	0.0	0.0	0.0
	CL, WN	86%	0.67	0.89	0.92	0.91	0.67	0.81	0.73	0.77
	CL, SS, WN, NE	89%	0.74	0.90	0.97	0.93	0.74	0.91	0.73	0.81
	CL, SS, WN, NE, DP, CN	90%	0.76	0.91	0.96	0.93	0.76	0.90	0.77	0.81
	CL, SS, WN, NE, DP, TP, CN	89%	0.74	0.91	0.95	0.93	0.74	0.74	0.86	0.81
	Over Sampling	84%	0.68	0.82	0.88	0.85	0.68	0.86	0.82	0.84
	Over and Under Sampling	88%	0.76	0.87	0.90	0.88	0.76	0.90	0.86	**0.88**
OntoNotes	Baseline MFC	60%	0.0	0.60	0.1	0.74	0.0	0.0	0.0	0.0
	CL, SS, WN, NE, DP, TP, CN	49%	0.0	0.0	0.0	0.0	0.0	0.50	1.0	0.66
	*Evaluation by Random Sampling**	70%	0.0*	0.0*	0.0*	0.0*	0.0*	0.50	1.0*	0.82
	Over and Under Sampling	24%	0.0	0.18	1.0	0.29	0.0	0.83	1.0	0.91
	Over Sampling	87%	0.50	0.76	0.50	0.58	0.50	0.90	0.95	**0.92**
CEN	Baseline MFC	95%	0.0	0.95	0.1	0.97	0.0	0.0	0.0	0.0
	CL, SS, WN, NE, DP, TP, CN	97%	0.71	0.97	0.99	0.98	0.71	0.87	0.63	0.73
	Over Sampling	94%	0.80	0.95	0.98	0.97	0.80	0.91	0.78	0.83
	Over and Under Sampling	90%	0.80	0.88	0.92	0.90	0.80	0.91	0.88	**0.90**
	*Evaluation by Random Sampling**	96%	1.0*	0.96	1.0*	0.98	1.0*	1.0*	1.0*	1.0*
Weighted Average (by # of Coref Chains)		89%	0.92	0.93	0.95	0.94	0.92	0.91	0.88	**0.90**

Table 2: Performance of different features sets for identifying characters. MFC = most frequent class. κ = Cohen's kappa (Cohen, 1960). *Because the co-reference chains on the OntoNotes and CEN corpora are automatically computed, they are noisy. We performed an evaluation by randomly sampling and correcting coreference chains, and running the classifer on those, thus estimating performance of the classifier on the dataset if it had clean chains. Details of the sampling are given in footnotes 3 and 4.

5 Results & Discussion

The extended ProppLearner We performed some preprocessing on this corpus, primarily involved in correcting minor errors in the coreference chain annotation. This included removing duplicate coreference chains generated by Stanford CoreNLP, merging coreference chains with the same chain heads, and merging pronouns with the correct chain heads. As expected, we obtained good results using this corpus as the coreference chains are of high quality (i.e., we started with gold standard chains and corrected the small number of errors we found). Table 2 shows the full set of experiments with the model on each corpus. For the ProppLearner corpus we experimented with different combinations of features as shown. Using all seven features our model achieved an F_1 of 0.81 on this corpus. As mentioned above, we used ten-fold cross-validation. Furthermore, to evaluate the effect on the performance due to the character class imbalances in the animate chains from the animacy classifier, we experimented with two types of class balancing approaches: (1) oversampling the minority class only, and (2) oversampling the minority class and undersampling the majority class. In case 1, the performance improved marginally to an F_1 of 0.84, and in case 2, the performance improved to an F_1 of 0.88.

OntoNotes We evaluated our model in three ways on OntoNotes data. First, we trained and tested the character model on the complete OntoNotes data with all features, achieving an F_1 of 0.66. Because the OntoNotes coreference chains are not completely clean (containing some duplicates and incorrect chains), we used direct sampling (Saunders et al., 2009) to select a subset of the chains[3] and manually corrected them, and trained and tested the full model over this subset. This achieved an improved F_1 of 0.82, suggesting that the classes are imbalanced because the model voted for majority class only. Finally, as the classes are imbalanced (there are many fewer character chains with non-character chains), we performed over- and under-sampling in the same fashion as for the ProppLearner data. When oversampling only, we achieved an improved performance of 0.91 F_1. When over- and under-sampling simultaneously, we achieved a performance of 0.92 F_1.

Corpus of English Novels (CEN) We evaluated our model on CEN in exactly the same way as on OntoNotes. First, we ran our character model on the whole CEN data and achieved an F_1 of 0.73. We

[3]Confidence Level = 95% , Confidence Interval = 4, Population = 1,145 and Sample Size = 394

Corpus	Sampling	Settings
The extended ProppLearner	Over	Duplicated 532 character chains
	Over & Under	Duplicated 532 character chains, removed 266 non-character chains
OntoNotes	Over	Duplicated 347 character chains
	Over & Under	Duplicated 225 character chains, removed 225 non-character chains
Corpus of English Novels	Over	Duplicated 2,927 character chains
	Over & Under	Duplicated 6,104 character chains, removed 10,275 non-character chains

Table 3: Different settings for Over & Under Sampling.

	Test Corpus									Micro-Avg
	ProppLearner			OntoNotes			CEN			
Train Corpus	**Acc.**	κ	$\mathbf{F_1}$	**Acc.**	κ	$\mathbf{F_1}$	**Acc.**	κ	$\mathbf{F_1}$	$\mathbf{F_1}$
CEN	84%	0.68	0.84	48%	0.37	0.60	94%	0.85	0.96	**0.93**
OntoNotes + CEN	85%	0.72	0.85	50%	0.47	0.62	91%	0.83	0.92	0.89
ProppLearner + CEN	86%	0.72	0.86	47%	0.33	0.59	91%	0.83	0.91	0.89
All	85%	0.72	0.85	54%	0.07	0.66	91%	0.83	0.92	0.89
ProppLearner + OntoNotes	87%	0.74	0.87	81%	0.46	0.88	87%	0.75	0.88	0.88
ProppLearner	89%	0.79	0.89	69%	0.27	0.80	87%	0.75	0.88	0.85
OntoNotes	71%	0.43	0.76	88%	0.47	0.93	71%	0.42	0.76	0.77

Table 4: Performance of **character** model for different training and testing setups. κ = Cohen's kappa (Cohen, 1960).

used direct sampling[4] to select and correct coreference chains, and the model achieved an F_1 of 1.0 over this corrected data, suggesting that coreference chain quality was a significantly larger factor in performance over this data. Finally, tried oversampling alone to achieve a significantly improved F_1 of 0.83. We also tried simultaneous over- and under-sampling to achieve an improved result of 0.90 F_1.

5.1 Generalizability Experiments

We evaluated the generalizability of our model by experimenting with different corpora in training and testing. Table 4 shows that the model trained on ProppLearner performed best on every test corpus, and the model trained on OntoNotes performed poorly on others. The overall performance for these experiments is not as high as the experiments keeping the training and testing corpus the same. As we have discussed before, the three corpora are different in size, type, and structure. The ProppLearner is a well-structured corpus including Russian folktales between 647 and 5,699 words; OntoNotes is a corpus full of short broadcast news texts (<1,028 words) that are loosely-structured story-wise; while the CEN corpus includes large chapters from English novels (1,402 - 7,060 words each) where the plot and characters are well developed. As a result, when we run the experiments on different training and testing corpus, the model sometimes finds it challenging to identify the right pattern for another type of corpus.

6 Error Analysis

A detailed error analysis of the results revealed problems for the character identification model that depend mainly on the external tools we have used and the quality of the data.

First, the character model uses the output of the animacy detector and so if a character was not marked animate, the character model also missed it. Conversely, sometimes inanimate chains are incorrectly marked animate, providing an additional opportunity for the character model to err. Thus, the character model's performance is bounded by that of the animacy model. This dependency is shown in Figure 7, where the character model performed better when we used the human-annotated animacy labels.

Second, the quality of coreference chains is critical for the character model. We can see from Table 2 that in the initial experiments, our model achieved excellent results for the extended ProppLearner (F_1 of 0.81) data because of its clean and hand-corrected coreference chains. On the other hand, the character

[4]Confidence Level = 95% , Confidence Interval = 4, Population = 17,251 and Sample Size = 580

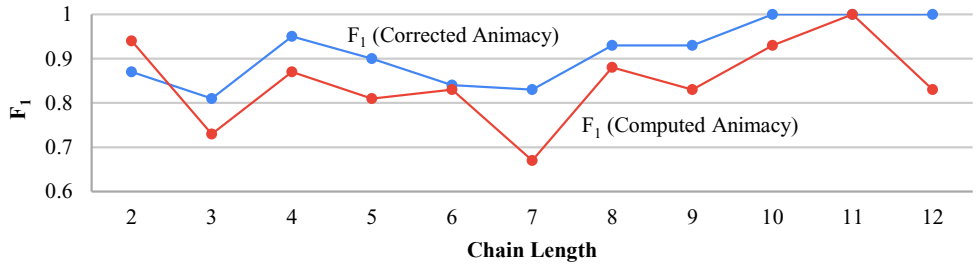

Figure 7: F_1 vs. chain length of the character identification model on OntoNotes for both manually corrected and automatically computer animacy markings.

model achieved a notably lower performance (F_1 of 0.66 and F_1 of 0.73) on the Ontonotes and CEN corpus, primarily because we have used the automatically generated conference chains for CEN corpus produced by Stanford CoreNLP. This was demonstrated by a random sampling evaluation to manually correct sample of CEN data, after which the model achieved a significantly improved F_1 of 1.0. We need better systems for automatically generating coreference chains to solve this problem.

Third, identifying the character information of short chains is a challenging task because chain length is one of the most effective features of our character model. In the case of short chains, the model only depends on the chain heads, and if a chain head does not carry much meaningful information, then the model can classify that chain incorrectly. We can see the performance improvement of the character model with increasing chain length from Figure 7. Solving this problem is critical, but adding more features that carry semantic information of a chain could be helpful.

Fourth, the data should be balanced to obtain good performance, which is a common requirement for any machine learning model. The performance improvement is shown in Table 2 for the three datasets after applying under- and over-sampling. The character model achieved the best performance when we applied over- and under-sampling together to ProppLearner (F_1 of 0.88). For OntoNotes, our model reached the best performance when oversampling is applied (F_1 of 0.92). Similarly, our model's performance significantly improved when over and under sampling are applied together to CEN (F_1 of 0.90).

Finally, one minor source of error for our model is limited foreign words, which is a data specific problem. The extended ProppLearner data contains numerous Russian character names (e.g., Parakha, Gornya, Shabarsha, etc.) that are not commonly found in English training data for NER systems or linguistic resources (WordNet, ConceptNet). As a result, our system was sometimes not able to identify these chains as a person, and that affects the model's performance. To address this problem, we could, for example, improve coverage of the NER gazetteers.

7 Related Work

Prior work on automatic character identification has relied heavily on statistical techniques and linguistic grammar-based techniques. Our work is mainly inspired by Calix et al. (2013) who used a Support Vector Machine (SVM) classifier to detect sentient actors in spoken stories. The model compares four different ML classifiers with 83 features (including knowledge features extracted from ConceptNet) and reports an F_1 of 0.86. It was found that certain speech features enhanced the results for non-named entities. However, the model focuses on animacy detection rather than character identification.

A similar line of work by Valls-Vargas et al. (2014) implemented a case-based approach using the *Voz* system. Apart from linguistic features, the most important features were extracted from WordNet and ConceptNet. Although they reported a 93.49% accuracy for a subset of the Proppian Folktales, it does not give a concrete definition of a character. They also proposed a similarity measure (*Continuous Jaccard*) that compares the entities from the text and case-base of the *Voz* system. Valls-Vargas (2015) further incorporated a feedback loop into *Voz*; this iterative approach improves co-reference grouping, but there isn't an improvement in character identification.

The most recent work on character identification took a supervised ML approach to classify nouns as characters using 24 different linguistic features, including capitalization and possession-based on

Freeling and JavaRAP (Barros et al., 2019). Out of the different classifiers, *ClassificationViaRegression,* achieved an F_1 of 0.84; however, it only worked for nouns and ignored pronouns.

Other approaches have used NER systems and domain-specific gazetteers in addition to other techniques such as graphs and verb analysis. Vala et al. (2015) proposed an eight-stage pipeline for identifying characters by building a graph where each name is represented as a node, and the nodes representing the same character are connected with edges. NER and co-reference resolution are used to populate the graph and connect nodes co-occurring in a chain, respectively. The main heuristics used distinguish between distinct characteristics compares genders (by looking at honorifics) and names. The model achieves an average F_1 of 0.58 on two datasets; however, it is limited to a corpus with characters that can be easily recognized by NER. Goh et al. (2012a) proposed a NER-based approach to identify the protagonists in fairy tales using WordNet and verb features. They used the Stanford parser to extract NE candidates, which is then filtered by verb analysis. They reported an F_1 of 0.67. In further work, Goh et al. (2013) identified the dominant character in fables using the VAHA (Verbs Associated with Human Activity) architecture (Goh et al., 2012b) and taking into account quoted speech, achieving an F_1 of 0.76. The same architecture, when applied to news articles, achieves an F_1 of 0.88 (Goh et al., 2015). Vani and Antonucci (2019) has described a modular tool called NOVEL2GRAPH, which generates visual summaries of narrative text. As part of the first module, characters are detected using Stanford's NER, which are further filtered using part-of-speech tagging. Character aliases are grouped using the DBSCAN clustering algorithm and stored in a dictionary. They did not report the performance of their approach.

Lastly, Declerck et al. (2012) demonstrated an ontology-based approach for automated character identification in folktales. They compared indefinite noun phrases with ontology labels, and used the matches to propose potential characters. Finally, they applied inference rules, and all occurrences of a particular ontology label were marked as references to the same character. The study reports an F_1 of 0.80. Although this approach has the closest implicit definition of a character to ours, the ontology is domain-based and is unlikely to generalize well to other domains.

8 Contributions

This paper makes three contributions. First, we proposed a more appropriate definition of *character*, contrasting with prior computational works which did not provide a theoretically grounded definition. Additionally, we reported our findings of a review of the literature that is helpful to delineate and define the concept of character. Second, we double annotated 46 Russian folktales and singly annotated 94 OntoNotes news texts and 30 English novels (one chapter per novel) for character, generating data that will be useful for the community[5]. Finally, we have demonstrated a straightforward supervised machine learning classifier for identifying characters, achieving weighted average of 0.90 F_1, establishing a new standard for this task.

Acknowledgements

This work was supported in part by NSF CAREER Award IIS-1749917 and by DARPA Award FA8650-19-6017. We would also like to thank the members of the FIU Cognac Lab for their discussions and assistance.

References

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press, Toronto.

Cristina Barros, Marta Vicente, and Elena Lloret. 2019. Tackling the challenge of computational identification of characters in fictional narratives. In *2019 IEEE International Conference on Cognitive Computing (ICCC)*, pages 122–129, Milan, Italy.

[5]Code and data for this work may be found at https://doi.org/10.34703/gzx1-9v95/RB6ZH0

Ricardo A Calix, Leili Javadpout, Mehdi Khazaeli, and Gerald M Knapp. 2013. Automatic detection of nominal entities in speech for enriched content search. In *Proceeedings of the 26th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 190–195, St. Pete Beach, FL.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Seymour Chatman. 1980. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, NY.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Hendrik De Smet. 2008. Corpus of English novels. `https://perswww.kuleuven.be/~u0044428/`.

Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. 2012. Ontology-based incremental annotation of characters in folktales. In *the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 30–34, Avignon, France.

Joshua Eisenberg and Mark Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2708–2715, Copenhagen, Denmark.

Mark A. Finlayson. 2008. Collecting semantics in the wild: The story workbench. In *the AAAI Fall Symposium on Naturally Inspired Artificial Intelligence*, pages 46–53, Arlington, VA.

Mark A. Finlayson. 2011. The Story Workbench: An extensible semi-automatic text annotation tool. In *the 4th Workshop on Intelligent Narrative Technologies (INT4)*, pages 21–24, Stanford, CA.

Mark A. Finlayson. 2017. ProppLearner: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory. *Digital Scholarship in the Humanities*, 32(2):284–300.

Monika Fludernik. 2009. *An Introduction to Narratology*. Routledge, New York.

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. 2012a. Automatic identification of protagonist in fairy tales using verb. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 395–406, Kuala Lumpur, Malaysia.

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. 2012b. Vaha: Verbs associate with human activity–a study on fairy tales. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 313–322, Dalian, China.

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. 2013. Automatic dominant character identification in fables based on verb analysis – empirical study on the impact of anaphora resolution. *Knowledge-Based Systems*, 54:147 – 162.

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. 2015. Automatic discovery of person-related named-entity in news articles based on verb analysis. *Multimedia Tools and Applications*, 74(8):2587–2610.

Labiba Jahan and Mark Finlayson. 2019. Character identification refined: A proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, MN.

Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. 2018. A new approach to animacy detection. In *the 27th International Conference on Computational Linguistics (COLING)*, pages 1–12, Santa Fe, NM.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60, Baltimore, MD.

Wenbo Pang and Xiaozhong Fan. 2009. Chinese nominal entity recognition with semantic role labeling. In *2009 International Conference on Wireless Networks and Information Systems*, pages 263–266, Milan, Italy.

Vladimir Propp. 1968. *The Morphology of the Folktale (2nd ed.)*. University of Texas Press, Austin, TX.

Graham Alexander Sack. 2013. Character Networks for Narrative Generation: Structural Balance Theory and the Emergence of Proto-Narratives. In Mark A Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister, editors, *the 4th Workshop on Computational Models of Narrative (CMN'13)*, pages 183–197, Hamburg, Germany.

M. Saunders, P. Lewis, and A. Thornhill. 2009. *Research Methods for Business Students*. Always learning. Prentice Hall.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451, San Francisco, CA.

Ismail S Talib. 2010. Narrative theory: A brief introduction. Retrieved from `https://courses.nus.edu.sg/course/ellibst/NarrativeTheory/`.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 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*, pages 769–774, Lisbon, Portugal.

Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. In *the 7th Intelligent Narrative Technologies Workshop (INT7)*, pages 38–44, Milwaukee, WI.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2015. Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops. In *Proceedings of the 24th International Conference on Artificial Intelligence*, page 2517–2523, Buenos Aires, Argentina.

K Vani and Alessandro Antonucci. 2019. Novel2graph: Visual summaries of narrative text enhanced by machine learning. In *Text2Story@ ECIR*, pages 29–37, Cologne, Germany.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0. LDC Catalog No. LDC2013T19, `https://catalog.ldc.upenn.edu/LDC2013T19`.

W Victor H Yarlott and Mark A Finlayson. 2016. ProppML: A complete annotation scheme for proppian morphologies. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.