FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View

♠Sooyoun Han, ♠Sumin Seo, ♠Minji Kang, ♠Jongin Kim
♠Nayoung Choi, ♣Min Song, ♢Jinho D. Choi
♠Department of Digital Analytics, Yonsei University, Seoul, South Korea
♣Department of Library and Information Science, Yonsei University, Seoul, South Korea
♢Department of Computer Science, Emory University, Atlanta, GA, USA
{honeycello, soominn20, codenavy94, jongin.kim}@yonsei.ac.kr
{skdudenql11, min.song}@yonsei.ac.kr, jinho.choi@emory.edu

Abstract

This paper presents a new corpus and annotation guideline for a novel coreference resolution task on fictional texts, and analyzes its unique characteristics. *FantasyCoref* contains 211 stories of *Grimms' Fairy Tales* and 3 other fantasy literature annotated in the omniscient writer's point of view (OWV) to handle distinctive aspects in this genre. This task is more challenging than general coreference resolution in two ways. First, documents in our corpus are 2.5 times longer than the ones in OntoNotes, raising a new layer of difficulty in resolving long-distant referents. Second, annotation of literary styles and concepts raise several issues which are not sufficiently addressed in the existing annotation guidelines. Hence, considerations on such issues and the concept of OWV are necessary to achieve high inter-annotator agreement (IAA) in coreference resolution of fictional texts. We carefully conduct annotation tasks in four stages to ensure the quality of our annotation. As a result, a high IAA score of 87% is achieved using the standard coreference evaluation metric. Finally, state-of-the-art coreference resolution approaches are evaluated on our corpus. After training with our annotated dataset, there was a 2.59% and 3.06% improvement over the model trained on the OntoNotes dataset. Also, we observe that the portion of errors specific to fictional texts declines after the training.

1 Introduction

Coreference resolution is a core NLP task to link all mentions that refer to the same entity together in a document (Pradhan et al., 2011, 2012). A few coreference datasets have been created for several genres (Hovy et al., 2006; Li et al., 2016; Zhou and Choi, 2018); however, coreference resolution on literary texts has been comparatively underexplored. This may be due to the nature of this entertainment genre being not as important. Nonetheless, with the latest advancement in neural-based coreference systems (Lee et al., 2018; Joshi et al., 2020; Wu et al., 2020), the scope of this task is getting broader and likely emerges as a key component to teach children how to read in education applications (Hill et al., 2016) or to conduct engaging conversations in chatbots.

Unlike non-fictional texts, referents in literary texts can be interpreted quite differently, depending on which point of view that the annotator takes (e.g., which character's point of view, which part of the story that the reader is at). This unique property in literature is more frequently noticeable in the fantasy genre.

For instance, in the famous story *Snow White*, the *evil queen* disguises herself as an *old woman* and gives Snow White a poisonous apple. From Snow White's point of view, she does not know that the *old woman* is her *step-mother* at this point, but the writer knows it. Moreover, the three mentions, *evil queen*, *old woman*, and *step-mother*, all refer to the same entity but appear differently, which can be confusing even in the reader's point of view before one finishes reading the entire story.

To accomplish the main purpose of coreference resolution, language understanding, this issue of inconsistent interpretation needs to be addressed with specific guidelines. Distinguished from most of previous work, this study conducts coreference annotation in the omniscient writer's point of view, which can reflect the author's intention of how the text should be understood, and therefore, is proper for reading comprehension.

Also, OntoNotes, the most widely known coreference dataset (Hovy et al., 2006), does not ana-

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 24–35
Punta Cana, Dominican Republic, November 10–11, 2021.

lyze this issue in its guidelines. Moreover, previous studies of coreference resolution on literary texts (Bamman et al., 2020; Roesiger et al., 2018; Yoder et al., 2021) present only limited analyses on dynamic issues that occur in fictional texts (e.g., asymmetry of knowledge, comprehensive physical or status change in entities).

The lack of a high-quality dataset in terms of size and consistency is the major hindrance in this task. This inspires us to create a new coreference corpus on literature and evaluate a state-of-the-art coreference system on this genre to confirm the feasibility of this research. Contributions of this work are as follows:

- We analyze a full range of coreference annotation related properties observed in fictional texts (especially, fantasy texts) and present the corresponding guidelines for this genre (Section 2).
- We create the corpus called *FantasyCoref*, comprising of 211 stories in *Grimms' Fairy Tales*, two stories from *The Arabian Nights*, and *Alice's Adventures in Wonderland* with coreference annotation that shows high inter-annotator agreement (Section 3).
- We evaluate a state-of-the-art coreference system on *FantasyCoref* (Section 4), and give error analysis specific to fictional texts, depicting limitations of the current system on this genre. We also show how the training the system on our dataset brings change to the error distributions (Section 5).

To the best of our knowledge, this is the first publicly available corpus that provides full coreference annotation on fantasy literature. We believe that this work will lead a new perspective of coreference resolution on this unexplored domain.[1]

2 Annotation

The source materials used for our coreference annotation are *Household Tales by Brothers Grimm by Jacob Grimm and Wilhelm Grimm* (*Grimms' Fairy Tales*; henceforth, GFT), which consists of 211 stories[2]. Three additional fantasy texts (henceforth, AFT) have also been annotated to be used as a separate test set. AFT includes *The Story of Aladdin* and *The Story of Ali Baba and the Forty Thieves* from *The Arabian Nights: Their Best-known Tales by Nora Archibald Smith and Kate Douglas*[3] with *Alice's Adventures in Wonderland by Lewis Carroll*[4]. All sources are available on the Project Gutenberg. The materials consist solely of fantasy texts, from which literary-specific characteristics such as metamorphoses of a character, metaphorical expressions, and asymmetry of knowledge are abundantly found. Meanwhile, these features hardly appear in non-fictional contexts, and thus, the existing corpora (which mostly choose non-fictions as their materials) have paid less attention to these literary features.

2.1 Annotation Process

A group of three linguists is formed for this project, who use an open source tool called CorefAnnotator (Reiter, 2018) to create our *FantasyCoref* corpus. Our annotation guidelines are largely based on the OntoNotes Coreference Guidelines 7.0 (Hovy et al., 2006), while referring to other studies on literary texts (Bamman et al., 2020; Roesiger et al., 2018) to consider the characteristics of the genre. The referents are annotated with the omniscient writer's point of view. Furthermore, while defining coreference relations, we adopt the entity-cluster view, which is also adopted by the CoNLL shared tasks (Pradhan et al., 2011, 2012). This view indicates that two or more mentions are non-hierarchically grouped into the same entity cluster (Cranenburgh, van, 2019).

2.2 Annotation Guidelines

Several annotation issues specific to fictional texts, which have not yet been addressed or developed into concrete guidelines in the previous studies, are outlined under the following four categories. Examples from GFT and the percentage of corresponding stories are shown in Table 1.

The table not only shows the issue percentage of GFT stories, but also that of 100 documents sampled from the OntoNotes corpus. The samples are chosen with a stratified sampling method to have homogeneous distribution of the entire genres included in the dataset (i.e., news, conversational telephone speech, weblogs, usenet newsgroups,

[1] All our resources including the FantasyCoref corpus are publicly available through our open source project: https://github.com/emorynlp/FantasyCoref

[2] https://www.gutenberg.org/ebooks/5314

[3] https://www.gutenberg.org/ebooks/20916

[4] https://www.gutenberg.org/ebooks/11

Issue	Annotation Examples	Issue % (FantasyCoref)	Issue % (OntoNotes)
Asymmetry of Knowledge	[*Cinderella*] They never once thought of [Cinderella]$_x$, and believed that [she]$_x$ was sitting at home in the dirt, picking lentils out of (...) He waited until [her]$_x$ father came, and said to him, "[The stranger-maiden]$_x$ has escaped from me, and I believe [she]$_x$ has climbed up the pear-tree.	36.4% (78/214)	2.0% (2/100)
Changes in Entities	[*The Frog-King or Iron Henry*] "How [the silly frog]$_x$ does talk!" (...) But when [he]$_x$ fell down [he]$_x$ was no frog but a King's son with beautiful (...) the way [the King's son]$_x$ heard a cracking behind [him]$_x$ (...).	24.8% (53/214)	1.0% (1/100)
Foretelling or Wishes	[*Jorinda and Joringel*] At last he dreamt one night that he found [a blood-red flower]$_x$, (...) He sought until the ninth day, and then, early in the morning, he found [the blood-red flower]$_x$.	7.0% (15/214)	1.0% (1/100)
Lexical Variations	[*The Wonderful Musician*] "(...) I will fetch hither [a good companion]$_x$ for myself." Then he took [his fiddle]$_x$ from his back, and played (...).	23.4% (50/214)	35.0% (35/100)

Table 1: Annotation issues, their corresponding examples from Grimms' Fairy Tales, and the proportion of documents from FantasyCoref and OntoNotes, respectively, in which each issue has been found.

broadcast, and talk shows). The comparison between the two corpora shows that several issues are dominantly specific to fictional texts. Moreover, although OntoNotes shows higher proportion on *Lexical Variations*, lexical variety found in non-fictions often differ from those in fictions. While lexical variations in non-fictional texts such as news (e.g, *Iran - the city*, *the Ford - the company*) are relatively easy to assume from world-knowledge, this may not be the case for literary contexts: The linkage among metaphorical expressions and paraphrases need to be indirectly assumed from utterances and actions of characters, which may be far more complicated and confusing in coreference resolution tasks. Hence, it can be said that the existing guideline does not provide sufficient information on these literary-specific issues. We believe the following guidelines are highly necessary to achieve fine-quality annotation of literary texts. More examples and explanations of the issue categories are provided in Appendix A.

Asymmetry of Knowledge This occurs when the knowledge is shared differently (a) between the reader and the characters or (b) between characters or when (c) the reader's knowledge changes throughout the plot (e.g., a plot-twist) (Bamman et al., 2020). Furthermore, we elaborate this issue by categorizing it into *deception (lie, disguise)*, *mistaking*, *secret* and *a plot-twist*. In these cases, we follow the omniscient writer's point of view.

Changes in Entities In literary texts, changes often take place in the development of entities (e.g., transformations or changes in the status of a character). While this phenomenon has been addressed in the previous literature (Roesiger et al., 2018), they have not suggested a specific guideline for such. When there exist separate mentions referring to a human-character, an animal, an object, or a place before and after changes, we group them under a single entity as long as they hold the same identity throughout the story.

Foretelling or Wishes One of the characteristics of literature is that characters often face a prophecy, curse or dream, and what has been foretold becomes reality. This is parallel to cases where a character's wish is granted and an entity that has been wished for is realized. In this case, we link the first appearance of an entity in the foretelling (or wishes) to its following real-world counterparts.

Lexical Variations Literary texts frequently make use of lexical variations or paraphrases (Roesiger et al., 2018). Hence, a character or a noun phrase is often repetitively referred to as various expressions. We group these anaphoric (e.g., the dragon = monster) or metaphoric expressions (e.g., three girls = gifts) under a single entity.

2.3 Inter-Annotator Agreement

To ensure the quality of our annotation, the standard coreference evaluation metrics, MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_{\phi_4}$ (Luo, 2005), are used to estimate inter-annotator agreement (IAA). The Kappa score (Cohen, 1960) is considered the standard method of

evaluating IAA in NLP tasks; however, it evaluates IAA in a pair-wise fashion, while coreference resolution assigns a markable to a coreference chain. Therefore, to avoid the Kappa score over-penalizing the wrongly assigned markable (He, 2007), we used the aforementioned metrics instead.

The agreement scores are calculated in four stages. First three stages were on GFT and the last stage was on AFT. In each stage, the annotators have been divided into three groups consisting of two annotators (i.e., first group of annotator 1 and 2, second group of annotator 1 and 3, third group of annotator 2 and 3), and each group worked independently on two random stories. Then, sets of scores calculated by the three groups are averaged into the final score for each stage. Table 2 shows the IAA scores in three stages. After the third stage, a high IAA score of 87% is achieved. The score of the last stage is similar with the third stage's even though the it used stories other than *Grimm Stories*.

Stage	MUC	B^3	$CEAF_{\phi_4}$	Avg-F1
1	90.34	84.48	74.21	83.01
2	92.25	86.70	73.62	84.19
3	91.72	86.87	**82.54**	**87.04**
4	**92.62**	**87.14**	78.45	86.07

Table 2: Inter-annotator agreement scores measured by the evaluation metrics, MUC, B^3, and $CEAF_{\phi_4}$, in three stages and the averaged score for each stage.

3 Corpus

3.1 Corpus Overview

FantasyCoref is a novel dataset of 367,891 tokens comprising of GFT and AFT, and is publicly available. The statistical information of our corpus is shown in Table 3.

	Average per document		Total	
	GFT	AFT	GFT	AFT
# of sentences	57	327	12,128	980
# of tokens	1,614	9,115	340,546	27,345
# of entities	28	116	5,829	347
# of mentions	270	1,529	56,968	4,587
# of mentions per entity	9	14	-	-

Table 3: Basic statistics of FantasyCoref.

3.2 Corpus Analytics

FantasyCoref is larger than the other existing public coreference resolution corpus on literary texts in English: LitBank (Bamman et al., 2019, 2020). *FantasyCoref* covers all types of noun phrases while LitBank annotates only six subsets of entity types (people, facilities, locations, geo-political entities, organizations and vehicles).

Compared with other benchmark datasets such as OntoNotes (Hovy et al., 2006), *FantasyCoref* shows a larger number of tokens per document. Regarding this characteristic, we calculated the average distance between mentions by the number of tokens between a mention and its antecedent within a single entity. *FantasyCoref*'s average distance between consecutive mentions and spread(distance between the first mention and the last mention within the same entity) are longer than OntoNotes'.

On the qualitative side, *FantasyCoref* presents a new perspective of coreference resolution by introducing an omniscient writer's point of view to the corpus which can handle multiple perspectives between timelines and characters that exist in literature. Furthermore, *FantasyCoref* consists of fantasy texts, which include various situations such as metamorphoses of characters and prophecies becoming to reality. Such issues make coreference resolution more complex compared to the other genre of the literature.

	# of tokens	# of docs	# tokens per document
FantasyCoref	370K	214	1,700
OntoNotes	1,600K	2,384	700
LitBank	210K	100	2,100

Table 4: Comparison between FantasyCoref and other coreference resolution corpora. Tokens per document are rounded to the nearest ten.

	Antecedent Distance		Spread	
	Fantasy Coref	Onto Notes	Fantasy Coref	Onto Notes
mean	57	48	484	171
std	191	112	823	280
min	1	1	1	1
median	13	14	186	46
max	7,086	4,362	11,010	4,876

Table 5: Comparison of mention distance between FantasyCoref and OntoNotes.

4 Experiment

4.1 Model

We use the state-of-the-art NLP model, Emory Language and Information Toolkit (ELIT) (Xu and Choi, 2020), to examine how well the existing coreference resolution model accounts for our annotated corpus, *FantasyCoref*. We experiment on end-to-end coreference systems with and without higher-order inference (HOI) approaches implemented in ELIT: attended antecedent (AA), entity equalization (EE), span clustering (SC), and cluster merging (CM).

The end-to-end coreference system is based on *c2f-coref* model (Lee et al., 2018) and SpanBERT (Joshi et al., 2020), and we adopt the "*independent*" splitting variant for long documents introduced by Joshi et al. (2019). The higher-order inference is a widely adapted method for global optimization of coreference links. Both AA and EE refines mention representation by aggregating its antecedents' information. While AA uses the distribution over antecedents from the span-ranking process as attention mechanism for refinement, EE aggregates all mentions in the cluster thus equalizing all representations of mentions in the same cluster. SC also refines mention representation from spans in cluster where it belongs, but differs from EE that it constructs the actual clusters from true predicted entities. CM ranks antecedents by sequentially merging entity clusters (see Xu and Choi (2020) for more details).

Both the OntoNotes model, which is trained on OntoNotes dataset, and our model, which is trained on GFT, are evaluated on the test set.

4.2 Data Processing

Development and test set are chosen from 211 stories (10 stories each) in GFT using stratified sampling method so that the train/dev/test set could have homogeneous distributions in terms of number of sentences, tokens, entities, mentions, and number of issues in Table 1. The rest of the stories (171 stories) are used as a train set to train our model. Note that all stories in AFT are only used as a separate test set to evaluate the generalization ability of our model on additional fictional texts.

Furthermore, in the case of the dev/test set, additional partitioned version is constructed to reduce the number of tokens to be similar to that of the OntoNotes dataset (467 tokens). The purpose is to compare the models' performance on both long and short version of the dataset, as it is widely known that the performance of coreference resolution on long documents is relatively poor. The partitioning process was done towards preserving the original entity chain. Each story is partitioned by finding the case where the sum of the number of entities in each partition is minimum so that the original entity chains are not cut in the middle and the partitioning does not produce extra number of entities in each partition. In addition, the stories were not partitioned in the middle of pair quotation marks to avoid starting or ending the partition in the middle of one's utterance. The train set is not partitioned, since feeding the model as many sentences as it can handle would be more desirable to learn coreferent links between distant mention pairs.

For brevity, the original versions and the partitioned versions of the dev/test set are referred to as follows for the rest of the paper:

- split 1: the original version of the GFT dev/test set
- split 2: the partitioned version of the GFT dev/test set
- split 3: the partitioned version of the AFT test set (We do not evaluate our model on the original version of AFT because we encounter a memory issue due to their large number of tokens, averaging to 9199).

The statistical information of the train/dev/test set and the comparison before and after partitioning is shown in Table 6.

			# of tokens per doc	# of docs
GFT	train	original	1,611	171
	dev	original (split 1)	1,652	20
		partitioned (split 2)	525	63
	test	original (split 1)	1,601	20
		partitioned (split 2)	525	61
AFT	test	original	9,119	3
		partitioned (split 3)	516	53

Table 6: Statistics of the train/dev/test set and before and after partitioning

	split 1										
	MUC			B^3			$CEAF_{\phi_4}$				
	P	R	F1	P	R	F1	P	R	F1	Avg. F1	Avg-M
SpanBERT + AA (OntoNotes model)	81.02	86.61	83.72	66.76	69.72	69.21	58.13	**66.16**	61.89	71.27	-
SpanBERT (Ours)	84.47	86.01	85.23	**70.86**	72.69	**71.76**	**67.28**	62.09	**64.58**	**73.86**	**73.21 (±0.60)**
+ AA	83.90	86.19	85.03	69.16	69.67	69.41	66.23	59.10	62.46	72.30	71.74 (±0.56)
+ EE	**84.90**	62.83	72.21	70.22	44.46	54.45	63.56	34.80	44.98	57.21	56.94 (±0.29)
+ SC	83.70	86.24	84.95	70.42	71.29	70.85	65.87	60.78	63.22	73.01	72.62 (±0.39)
+ CM	82.87	**88.04**	**85.38**	66.79	**76.28**	71.22	66.60	62.41	64.44	73.68	73.18 (±0.37)
	split 2										
	MUC			B^3			$CEAF_{\phi_4}$				
	P	R	F1	P	R	F1	P	R	F1	Avg. F1	Avg-M
SpanBERT (OntoNotes model)	81.01	87.82	84.27	69.17	75.21	72.06	64.96	65.31	65.13	73.82	-
SpanBERT (Ours)	83.64	87.92	85.73	71.15	77.69	74.28	72.97	62.20	67.15	75.52	75.11 (±0.66)
+ AA	83.33	86.69	84.98	71.88	75.22	73.51	70.82	62.53	66.42	74.97	74.77 (±0.22)
+ EE	**84.04**	87.72	85.84	**74.60**	75.80	**75.19**	72.37	**67.07**	**69.62**	**76.88**	**76.55 (±0.37)**
+ SC	83.78	87.06	85.39	71.90	76.58	74.17	71.55	60.70	65.68	75.08	74.92 (±0.13)
+ CM	83.83	**88.05**	**85.89**	72.06	**78.04**	74.93	**72.98**	62.11	67.11	75.98	75.49 (±0.35)
	split 3										
	MUC			B^3			$CEAF_{\phi_4}$				
	P	R	F1	P	R	F1	P	R	F1	Avg. F1	Avg-M
SpanBERT + SC (OntoNotes model)	87.12	**87.68**	87.40	77.39	**78.80**	**78.09**	68.45	55.02	61.00	75.50	-
SpanBERT (Ours)	88.64	82.94	85.69	76.78	72.44	74.55	**73.21**	48.66	58.47	72.90	-
+ EE	**88.66**	86.39	**87.51**	**79.86**	73.93	76.78	55.19	**73.76**	**63.14**	**75.81**	-

Table 7: The best performance result of different approaches on *FantasyCoref*, measured by the standard coreference evaluation metrics, MUC, B^3, and $CEAF_{\phi_4}$. The main evaluation metric is the averaged F1 of the three metrics. (P: Precision, R: Recall, Avg-M : the mean of Avg-F1 and the standard deviation from three developments)

4.3 Experiment Details

The experiment is comprised of three parts:

1. (for every approaches in Section 4.1) we take the best performing model on the GFT split 1 dev set and compare the result with the OntoNotes model on the GFT split 1 test set.

2. (for every approaches) we take the best performing model on the GFT split 2 dev set and compare the result with the OntoNotes model on the GFT split 2 test set.

3. We take the best performing model from part 1 and 2 and compare the result with the OntoNotes model on the AFT split 3 test set.

4.4 Results

Table 7 shows the results of the OntoNotes model and our model evaluated on the GFT test set (split 1, split 2) and the AFT test set (split 3). For the OntoNotes model, only the result of the best model among the five approaches are reported. In the case of our model, we report the mean scores and the standard deviations after three repeated developments for precise measurement.

The effect of training the model on GFT dataset is clear, when comparing the performance of the OntoNotes model with ours, for both split 1 and 2. The Avg-F1 score improves 2.59% (71.27% to 73.86%) and 3.06% (73.82% to 76.88%) for split 1 and 2, respectively. It can be interpreted that the model captures and learns the characteristics of coreferent links specific in fictional texts. In fact, we observe that the error types specific to fictional texts decreases after fine-tuning, which will be discussed in more detail in Section 5.

SpanBERT with entity equalization (EE; HOI suggested by Kantor and Globerson (2019)) shows a considerable performance difference in split 1 and 2 (Table 7). According to the published code of EE[5], the number of candidate spans is limited to

[5] https://github.com/lxucs/coref-ee

300 for the reason that the implementation of EE requires O(k^2) memory with k being the number of mentions extracted, while other HOI approaches requires O(k). Under this limited setting, the poor performance (especially the recall of metrics) is expected for long documents whose number of mentions in gold labels exceeds 300. EE is a limited approach for long documents in terms of computational limitation.

Also, the evaluation on AFT shows that the best performing model on split 2 still maintains its performance with slight decrease of 1.07% (76.88% to 75.81%), which still surpasses the OntoNotes model. This shows the generalization ability of our model trained on GFT to fictional texts.

5 Error Analysis

For both OntoNotes model and our model, the predicted results of 9 stories, randomly selected from the test set, have been examined based on a number of error types. The error categories were added or modified by the annotators throughout the analysis, and the finalized error types are found in Table 8. The distribution of error types for OntoNotes and our model are shown in Figure 1 and Figure 2, respectively.

5.1 OntoNotes model

(simplified).pdf (simplified).pdf

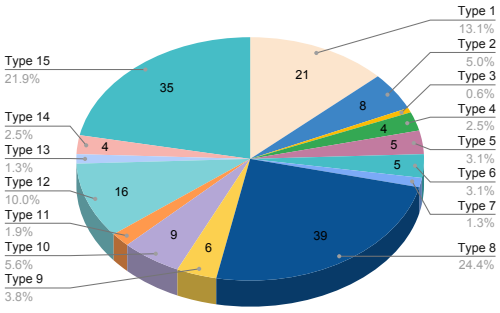

Figure 1: The distribution of error types for OntoNotes model. See Table 8 for details of the error types.

For OntoNotes model, 160 errors are observed and 46.3% of errors are found in Type 8 (Lexical Variations) and Type 15 (Miscellaneous) combined. Large proportion of Type 8 errors indicate that the ELIT system has a limited performance in capturing various metaphoric expressions or lexical variations as single entities, which frequently appear in literary texts. Moreover, Miscellaneous type errors mostly include cases in which ELIT captures an entity correctly but its referents only partially, while negligence of abstract concepts (e.g., *the power of three giants* and *their power*) or temporal expressions (e.g., *yesterday*) are also found.

Errors specific to literary texts are shown by Type 9 (Foretelling, Wishes, 3.8%), Type 5 (Plot-Twists, 3.1%), Type 6 (Changes in Appearance, 3.1%), and Type 4 (Mistakes, 2.5%), which consist of 12.5% of errors combined. The proportion is not dominant; however, the errors show that the current coreference resolution model (trained on OntoNotes) has limitations on linking referents such as follows: an imaginative concept to its realization, asymmetric information, and a character to its new identity after change.

5.2 Our model

(simplified).pdf (simplified).pdf

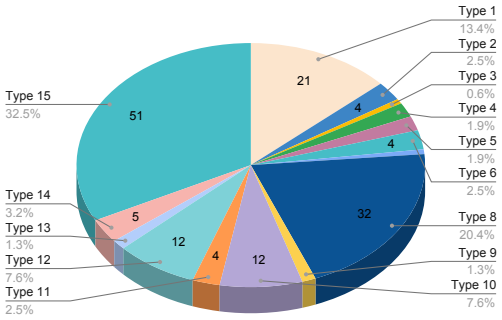

Figure 2: The distribution of error types for our model. See Table 8 for details of the error types.

For our model, the number of errors is 106, approximately 33.8% reduced compared to the result of OntoNotes model. The proportion of Type 15 (Miscellaneous, 32.5%), Type 8 (Lexical Variations, 20.4%), Type 1 (General Objects, 13.4%), Type 12 (Different entities, 7.6%), and Type 10 (Long Modifiers, 7.6%) combined are still dominant (81.5%), while the increased ratio of Type 15 is noticeable. The increase in the Type 15 errors and the reduced ratio of most of the other categories indicate the following: the model trained on our dataset better distinguishes separate entities and better predicts entities with respect to literary-specific issues, and now, more errors are of 'missing referents', rather than 'missing entities'.

Comparing our model to OntoNotes model, the performance improves with respect to Type 2 (Long Distance, 5.0% to 4.0%). The average distance of

Error Types	Description
1. General Objects	A general object that does not appear frequently in a text and thus not captured as an entity.
2. Long Distance	A referent that is far apart (more than 700 tokens) from its preceding referent is not linked to its preceding referent.
3. Lies	Two or more referents involved in a character's lying (deception) are not correctly linked.
4. Mistakes	When a character mistakes an object or a person for something or someone else (but the referents need to be categorized under a single entity based on OWV), referents involed in such case are not correctly predicted.
5. Plot-Twists	Two or more referents that are involved in a plot-twist are not correctly linked.
6. Changes in Appearance	Referents are not correctly linked when they refer to a single entity which has undergone a change in its appearance (e.g., metamorphosis, physical growth)
7. Changes in Status	When a character goes through a change in status (e.g., demotion, promotion), terms of address used and after the change are not correctly linked.
8. Lexical Variations	Two or more terms of address referring to a single entity (e.g., anaphora, metaphorical expressions) are not correctly linked.
9. Foretelling, Wishes	A referent which seem separate from its entity due to a character's lying, but in fact is linked to its entity is not captured by the model.
10. Long Modifiers	A referent is not correctly linked to its entity when the referent is an NP with a long post-modifier (e.g., *a great neck-kerchief of silk embroidered with gold*).
11. NP+VP+Rel Clause	A referent having the structure of (NP+VP+relative clause modifying NP) is not marked as an entity (e.g., *a sword was hanging on the wall which was made of pure silver*).
12. Different Entities	A model groups separate entities and their referents under a single group.
13. Verbs & Nominalizations	A verb and its nominalizations (e.g., *flew-flight*) are not linked.
14. Nouns w/o Determiners	An NP used without a determiner is not linked with other referents (e.g., *table-the table*).
15. Miscellaneous	Errors which do not fall into the types above are categorized as 15. Miscellaneous.

Table 8: Error types used during the error analysis and the description of each error type.

the correct group becomes 43.34 tokens from 41.65 tokens and that of the error group becomes 88.96 tokens from 94.62 tokens. This implies that the model has improved towards capturing the mentions of the same entity in long distance after training with our dataset.

Moreover, a couple of errors from literary-specific categories, (Type 4 (Mistakes), Type 5 (Plot-Twists), Type 6 (Changes in Appearance), and Type 9 (Foretelling, Wishes) are now predicted correctly, which shows that training with our dataset has an effect of remedying issues caused during the coreference resolution of literary styles or concepts. As an example of Type 4, in the story *Frederick and Catherine*, a group of robbers under a tree mistakes *Catherine* for *the devil*, when she drops an object from up the tree. While the two referents, *Catherine* and *the devil* are grouped under separate entities in OntoNotes model, they are correctly captured under the *Catherine* cluster in our model. Moreover, a Type 5 error is exemplified by the story *The Girl without Hands*. In the text, *an old man* comes across *a stranger* on the street, who turns out to be *the devil* in the latter part of the story. The two referents, *a stranger* and *the devil*, which used to be in separate entity clusters are grouped under a single cluster in our model. Finally, a Type 6 error is shown by the story, *The Sea-Hare*: In the story, *a fox* transforms into a human *merchant*, and the two categories are correctly linked in our model.

The low proportion of errors specific to literary (fictional) texts may also be due to characteristics of texts randomly selected from the test set. Hence, large data sets of literary texts are in need to train the pre-exsiting coreference resolution systems and boost their performance with respect to literary-specific features.

6 Conclusion

In this work, we present *FantasyCoref*, a new English annotated dataset of 367,891 tokens from Grimms' Fairy Tales and additional fantasy literature. *FantasyCoref* is larger in size compared to the existing coreference corpora in literature, and takes into consideration a number of issues specific to this genre. These issues are organized as guidelines having four categories of the following: 1) Asymmetry of Knowledge, 2) Changes in Entities, 3) Foretelling or Wishes, and 4) Lexical Variations. The state-of-the-art coreference system trained on our annotated corpus results in a significant improvement on fictional texts and a decrease in genre-specific errors. However, the best performance with 76.88% still has room for improvement compared to the those shown in the OntoNotes dataset. Hence, we believe that *FantasyCoref* is

a highly valuable resource that can be utilized to develop the current coreference resolution models and, by extension, language comprehension tasks.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5B1104865).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Andreas Cranenburgh, van. 2019. A dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9.

Tian Ye He. 2007. Coreference resolution on entities and events for hospital discharge summaries. *Thesis (M. Eng.)–Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90% solution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*. The Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S Weld. 2019. Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. 2016. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia. European Language Resources Association (ELRA).

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Nils Reiter. 2018. CorefAnnotator - A New Annotation Tool for Entity References. In *Abstracts of EADH: Data in the Digital Humanities*.

Ina Roesiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138, Santa Fe, New Mexico. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, page 45–52, USA. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Michael Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Rosé. 2021. FanfictionNLP: A text processing pipeline for fanfiction. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 13–23, Virtual. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Appendix

We provide several paragraphs from *Grimms' Fairy Tales* which serve as an example for each annotation guideline.

A.1 Asymmetry of Knowledge

(1) (*Deception*) "If thou wilt invite [us]$_x$ to the wedding, not be ashamed of [us]$_x$, and wilt call [us]$_x$ thine aunts, (...) "I have [three aunts]$_x$," said the girl, "(...) allow me to invite [them]$_x$ to the wedding, and let [them]$_x$ sit with us at table."

In *The Three Spinners*, the girl lies that 'the three spinners' are her aunts. Nevertheless the spinners are not the girl's aunts in fact, the reader knows that the girl is referring to the same people. We connect 'the spinners' and 'the girl's aunts'.

(2) (*Mistaking*) He was dreadfully frightened, and ran to the back-door, but [the dog, who lay there]$_x$ sprang up and bit his leg; and as he ran across the yard by the straw-heap, [the donkey]$_y$ gave him a smart kick with its hind foot. (...) Then the robber ran back as fast as he could to his captain, and said, "Ah, (...) and by the door stands [a man with a knife, who stabbed me in the leg]$_x$; and in the yard there lies [a black monster, who beat me with a wooden club]$_y$ (...)"

In *The Bremen Town-Musicians* the robber mistakes animals for different things though the animals were not intended to deceive him. In (3), 'dog' is mistaken as 'a man with a knife' and 'the donkey' is mistaken as 'a black monster'. As the reader knows that the robber refers to the same thing, we annotate them as co-referents.

(3) (*Secret*) Then he took him about everywhere, up and down, and let him see all the riches, and the magnificent apartments, only there was one room which he did not open, that in which hung [the dangerous picture]$_x$. (...) "Ah, no," replied the young King, "if I do not go in, it will be my certain destruction.I should have no rest day or night until I had seen [it]$_x$ with my own eyes.

In *Faithful John*, Faithful John keep in secret to the young King that the dangerous picture is in the chamber. The young King says he will not move until he sees 'it'. Even though the young King does not know whether 'it' is the dangerous picture, the reader knows that, therefore we treated them as the same entity.

(4) (*Twist*) [He]$_x$ said to her kindly, "Do not be afraid, [I]$_x$ and [the fiddler who has been living with you in that wretched hovel]$_x$ are one. For love of you [I]$_x$ disguised [myself]$_x$ so; and [I]$_x$ also was the hussar who rode through your crockery. (...)"

In *King Thrushbeard*, it turns out that 'king Thrushbeard', 'the fiddler' and 'the hussar' were the same person at the end of the story. The reader may not notice it in the middle of the plot since it is the twist, however, we annotate all as the same referents with the knowledge after reading through all the text.

A.2 Changes in Entities

(5) (*Metamorphosis*) [The lion]$_x$, however, was an enchanted prince (...) a ray about the breadth of a hair fell on [the King's son]$_x$, and when this ray touched [him]$_x$, [he]$_x$ was transformed in an instant, and when she came in and looked for [him]$_x$, she did not see [him]$_x$, but [a white dove]$_x$ was sitting there.

In *The Singing, Springing Lark*, 'the lion' changes into 'a white dove' by being touched by 'a ray'. Since the identity of 'the lion' has not changed by its appearance, we annotate 'the lion' and 'a white dove' under the same entity.

(6) (*Change in Status*) There was once on a time [a girl who was young and beautiful]$_x$, (...). At last [she]$_x$ hired [herself]$_x$ to a farmer as a cow-herd, and buried [her]$_x$ dresses and jewels beneath a stone.(...) [she]$_x$ said,"Little calf, little calf, kneel by [my]$_x$ side, And do not forget [thy shepherd-maid]$_x$, As the prince forgot [his betrothed bride, Who waited for him 'neath the lime-tree's shade]$_x$."

In *The True Sweethearts*, a girl become a shepherd-maid and prince's bride during the story. Though a girl is not a shepherd-maid nor a bride in the beginning of the story, we group 'a girl', 'a shepherd-maid' and 'a bride' together.

(7) (*Growth*) (...) from [the two pieces]$_x$ that were buried in the ground [two golden lilies]$_x$ sprang up, (...)

In *The Gold-Children*, 'the two pieces (of a golden fish)' is buried, then grows into 'two golden lilies'. Although the appearance has changed throughout the growth, no further mention of 'the two pieces' appear after their growth. Hence, we consider 'the two pieces' and 'two golden lilies' to be the referents.

A.3 Foretelling and Wishes

(8) (*Prophecy*) "you must not drink [the wine which will be brought to you at night]$_x$, and must pretend to be sound asleep." (...) the eldest came and brought him [a cup of wine]$_x$ (...)

In *The Shoes that were Danced to Pieces*, 'an old woman' foretells and warns 'a poor soldier' that he must not drink 'the wine' that will be served that night, and 'a cup of wine' is indeed offered to him as the warning said. We link the first appearance of an entity in a prophecy, curse, or dream (in this example, 'the wine which will be brought to you at night') to its realization in the real-world (in this example, 'a cup of wine').

(9) (*Wish*) "Will you wish for [a new house]$_x$ instead of this old one?" "Oh, yes," said the man; "if I can have [that]$_x$, too, I should like it very much." And the Lord fulfilled his wish, and changed their old house into [a new one]$_x$, (...) on the opposite side of the way, [a new clean-looking house with red tiles and bright windows where the old hut used to be]$_x$.

In *The Poor Man and the Rich Man*, the man wishes for 'a house' and this wish is fulfilled as a form of 'a new clean-looking house with red tiles and bright windows'. When an entity that is wished or requested for in a story is realized into a real-world entity, we link the two as referents.

(10) (*Favor*) (...)"Here, put on this dress and go out into the wood, and fetch me [a little basketful of strawberries],—I have a fancy for some." (...) she answered, "I am to look for [a basketful of strawberries], and am not to go home until I can take them with me."

In *The Three Little Men in the Wood*, the mother makes the girl fetch her some strawberries and later, the girl says about strawberries. In this case, since 'a basketful of strawberries' are general, we do not annotate them as the same entity. If the mother asked the girl to find some 'special strawberries', strawberries should be annotated.

A.4 Lexical Variations

(11) (*Variations in terms of Address*) When Fir-twister was busy cooking, [a little shrivelled-up old mannikin]$_x$ came to him (...) "Be off, [sly hypocrite]$_x$," (...) But how astonished Fir-twister was when [the little insignificant dwarf]$_x$ sprang up at him, (...) "(...) they may just try their chance with [the little scrubbing-brush]$_x$;" (...) Then [the malicious dwarf]$_x$ wanted to spring on him (...)

In *Strong Hans*, 'a little shrivelled-up old manikins' is referred to as various expressions according to how other characters perceive him in the story or how the narrator depicts his appearance or personality. Regardless of the variation, when a reader can perceive that several anaphoric terms of address point to a single character, these terms are grouped under the same entity.

(12) (*Metaphoric Expression*) Then at last [the children]$_x$ became so impatient, (...) and ran away. But when church was over, the nix saw that [the birds]$_x$ were flown, and followed [them]$_x$ with great strides.

In *The Water-Nix*, 'the birds' is used as an metaphoric expression referring to 'the children' who have run away. We link metaphoric expressions to its referent when the connection between them can be perceived by the reader.