Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons

Akila Peiris and **Nisansa de Silva**
Department of Computer Science & Engineering,
University of Moratuwa, Sri Lanka
{akila.21,nisansadds}@cse.mrt.ac.lk

Abstract

This paper introduces the Forgotten Realms Wiki (*FRW*) data set and domain specific natural language generation using *FRW* along with related analyses. Forgotten Realms is the defacto default setting of the popular open ended tabletop fantasy role playing game, Dungeons & Dragons. The data set was extracted from the Forgotten Realms Fandom wiki consisting of more than over 45,200 articles. The *FRW* data set is constituted of 11 sub-data sets in a number of formats: raw plain text, plain text annotated by article title, directed link graphs, wiki info-boxes annotated by the wiki article title, Poincaré embedding of first link graph, multiple Word2Vec and Doc2Vec models of the corpus. This is the first data set of this size for the Dungeons & Dragons domain. We then present a pairwise similarity comparison benchmark which utilizes similarity measures. In addition, we perform D&D domain specific natural language generation using the corpus and evaluate the named entity classification with respect to the lore of Forgotten Realms.

1 Introduction

Specialized and domain specific data sets are useful for a number of advanced tasks in the domain of Natural Language Processing (NLP). For example, recent studies have shown that the domain specificity significantly impacts vital NLP tasks such as measuring semantic similarity (Sugathadasa et al., 2017) and domain specific text generation (Lebret et al., 2016). Further, it has been shown that models developed using data from a generic domain do not seamlessly transfer to tasks in a specific domain (Rajapaksha et al., 2020). Fantasy domains could be considered an extreme case of domain specific data, as it is possible to observe the full spectrum of deviations from the non-domain specific (general domain) usage, both in the lexical and semantic perspectives. An example for lexical differences is the usage of *dwarves* as the plural form of *dwarf* in the fantasy genre[1] in place of the general domain spelling *dwarfs*. An example for semantic differences can be seen in the words *Ghost*[2] and *Wraith*[3]. In the Merriam-Webster dictionary, they are given as synonyms in the generic domain[4] while in the domain of the fantasy role playing game Dungeons & Dragons, they are defined as two distinct creatures. In this paper, we present the *FRW* data set, specific to the *Forgotten Realms* setting of *Dungeons & Dragons*. We expect our data set to be useful for in-domain tasks such as text generation (Zhang et al., 2019), information extraction (de Silva and Dou, 2021), and information retrieval (Sugathadasa et al., 2018). We also anticipate our data set being vital for cross domain tasks such as text alignment (Sanchez-Perez et al., 2014), style transfer (Fu et al., 2018), and summarizing (El-Kassas et al., 2020). As a primer for these usages, we introduce a pairwise similarity comparison benchmark and evaluate the domain-specific free text generation task.

1.1 Dungeons & Dragons

Dungeons & Dragons (D&D or DnD), is an open-ended pen and paper tabletop role playing game (RPG) which has been commercially available since Gygax and Arneson (1974) published the first version. The games are primarily based on fantasy genre. However, there is a plethora of other settings ranging from science fiction, post-apocalyptic to hollow world and much more. Even within a selected genre, it is highly customizable, for example, a fantasy setting might be in high or low fantasy. D&D has a predefined set of rules governing almost every aspect of the gameplay including the *setting*. A setting has a lore, species and artifacts among other components; which can be dissimilar

[1] This is inherited from the spelling used in the *Lord of the Rings* and other relevant publications by J. R. R. Tolkien
[2] https://bit.ly/DnDGhost
[3] https://bit.ly/DnDwraith
[4] https://bit.ly/3Z2YsHC

between settings. There are also several editions of D&D, with 5 (Crawford et al., 2014) being the latest. It is the version that our *FRW* data sets are predominantly based on. However, it does contain some information from earlier editions in cases where there have been changes to the lore between versions or in cases where information have been consistently brought forward.

1.2 Forgotten Realms Wiki

Forgotten Realms as mentioned, is a setting which is categorized under *high fantasy*, set in an alternate world filled with magical elements combined with larger than life themes, plots, and characters. It originated as a medieval European setting but over the years, has been influenced by other cultures including Middle Eastern and Asian. *Forgotten Realms* became the most utilized of all the official D&D settings after it became the de-facto default setting of the immensely popular (Whitten, 2021) 5th edition. Almost all of the official material published for D&D is based on this setting. Due to this, *Forgotten Realms* now has the most resources and information available from all the settings in D&D.

However, this massive amount of information is distributed among hundreds of official books and magazines making it intractable for a casual enthusiast of D&D. To remedy this problem and to curate and consolidate the information, the community of D&D enthusiasts voluntarily contribute and maintain the *Forgotten Realms wikia* [5]. A *Wikia* or a *Fandom Wiki* is a Wikipedia [6]-esque website (uses the same MediaWiki [7] collaborative documentation platform) hosted by Fandom, Inc. [8]. This is typically dedicated to a particular domain (e.g., Star Wars [9], Marvel [10], Harry Potter [11], Formula One Racing [12]). The *Forgotten Realms Wikia* has over 45,200 articlesas of September 2022 and keeps growing at a rapid pace.

1.3 Wikipedia and other Wikis as Data Sources

Wikipedia and other Wikia, maintained by a community of volunteers, are treasure troves of domain specific knowledge (Ferrari et al., 2017). While there are endless debates regarding the validity of such community maintained knowledge-bases in scientific context (Cozza et al., 2016; Ferschke, 2014), there are still a number of ways they can be used to further the scientific frontiers (Ponzetto and Strube, 2007; Zesch and Gurevych, 2007; Zesch et al., 2008). One such usage in the field of Natural Language Processing is to use them as data sources which not only provide corpora of the relevant domains but also provides insight into community-based collaborative text maintenance (Ferschke et al., 2013).

The possibility of accessing as a freely available source in multiple languages (Nastase and Strube, 2013) (human translated), being extensive, and having special information content such as infoboxes[13] make Wikipedia and similar wikia rich resources for data. An infobox is the table-like structure typically found at the top-right side of a wiki article. It is a human annotated, tabular summary of the article, arranged in a key-value structure according to a template. According to Lange et al. (2010) about one third of all Wikipedia articles contain an infobox. While this is indeed a rich source of information, they are known to be noisy and sparse(Hoffmann et al., 2010). The wiki page itself only renders the pairs that contain values.

Another special information content is found in the first paragraph/ (lead section [14]) of a wiki article. According to the guideline, this is typically formatted as an abstractive summary to the entire page. In their study on wikipedia, Lange et al. (2010) report that there is a 92% chance to find any of the information summarized in the infobox within the first paragraph.

1.4 Domain Specific Text Generation

Domain specific text generation is an emerging area in NLP (Liu et al., 2018; Chen et al., 2021; Zhang et al., 2022; Amin-Nejad et al., 2020). The objective in this is to generate text which adheres to a given domain, in the sense that the content generated should be semantically and pragmatically truthful to the said domain. One of the reasons why domain specific text generation is difficult compared to generic text generation is that, in most cases this requires copious amounts of linguistic resources based on the domain in question. This hurdle is true even for fine-tuning a pre-trained

[5] https://forgottenrealms.fandom.com
[6] https://en.wikipedia.org
[7] https://bit.ly/3YHpG6K
[8] https://www.fandom.com
[9] https://starwars.fandom.com/
[10] https://marvel.fandom.com/
[11] https://harrypotter.fandom.com/
[12] https://f1.fandom.com/wiki/

[13] https://bit.ly/3lLcqiOen.wikipedia.org/wiki/Help:Infobox
[14] https://bit.ly/3IAzHNx

model which relatively demands less amount of data than training a model from ground-up (Zhang et al., 2021).

2 Related Work

2.1 Wikipedia and Other Wiki Data Sets

In recent times the availability of linguistic data sources have increased significantly. Especially Wikipedia based data sets such as Wit (Srinivasan et al., 2021), WCEP (Ghalandari et al., 2020), and SQuAD (Rajpurkar et al., 2016). Tools such as LUCHS (Hoffmann et al., 2010) and WOE (Wu and Weld, 2010) are capable of extracting information from Wikipedia pages to create such data sets. Both systems rely on the key-value structure of the infoboxes to guide the information extraction process from the natural language text. This guided process is akin to the widely used Ontology-Based Information Extraction (OBIE) (de Silva et al., 2017).

As mentioned, Fandom, Inc. is an organization which hosts wikis for a large number of entertainment media franchises and other areas as the general populace may desire. The Fandom wikis operate on the same technology and guidelines[15] as Wikipedia. They are good sources of domain specific data for different media franchises as they are written in the desired domain and offers a clear demarcation from in-domain and out-domain data as opposed to obtaining data from sources such as the common crawl (Kreutzer et al., 2022). The Critical Role Dungeons & Dragons Data set (*CRD3*) (Rameshkumar and Bailey, 2020) is a D&D domain-specific, narrative driven, multi speaker dialog data set that has been extracted from a similar Fandom Wiki[16]. This particular wiki is dedicated to the web series, *Critical Role*, a live D&D gameplay series. The data set consists of multi-speaker dialogue that form a narrative, paired with their abstractive summaries.

2.2 Domain Specific Text Generation

Text generation methodologies fall into three categories (Stent et al., 2004). **Template based** methods (Busemann and Horacek, 1998; Reiter and Dale, 1997; McRoy et al., 2003) are the most common variant. It uses pre-defined text templates applicable to different scenarios to generate text. It is a tedious and non-salable approach. Secondly, there is **Rule based** generation (Dale et al., 2003; Turner et al., 2009; Reiter et al., 2005). This has three inter-dependent phases: (1) text planning - governs the process of meaning representation retrieval from a knowledge base, (2) sentence planning - governs the words and their order to produce coherent sentences, and (3) surface realization - converts the sentence plan into actual sentences. Thirdly, the **Data driven** approach (Barzilay and Lapata, 2005; Liang et al., 2009). Unlike rule based approaches, data driven ones require more data. This burden is alleviated using pre-trained language models and transfer learning techniques. Open AI's Generative Pre-trained Transformer models GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) as well as the open source reproductions of these, the GPT-Neo models (Gao et al., 2020; Black et al., 2022) are such large pre-trained language models. In addition to having been trained on very large data sets, they are also large networks. These models are capable of generating highly sophisticated texts. With some fine-tuning, they can be adapted to do the same for specific domains.

3 Forgotten Realms Wiki (*FRW*) Data set

We introduce the Forgotten Realms Wiki (*FRW*) data set[17], extracted from the Forgotten Realms Fandom wikia. We have extracted multiple data sets from this textual resource and present them in multiple formats for different linguistic use cases. We also present several different embeddings for this data set including Poincaré hierarchical embedding and multiple word and document level embeddings. A summary of the data sets is shown in Table 1 and the individual statistics for each data set can be found listed under Table 2.

The plain text corpora (*FRW-P*, *FRW-J*) are devoid of special data structures and other markings. As for the links, the MediaWiki allows having an alternative text to display for the links instead of the exact page title for aesthetics of the writing, hence we extract that part for the plain text corpus when available. The *FRW-FJ* data set is composed of mainly the *lead sections*. Because of this, we can consider this as an abstractive summary of *FRW-J*. The *FRW-CL* links pages with categories. The categories themselves have rendered pages which aggregate the pages under each category. The infobox data in *FRW-I* are converted from markdown to

[15] https://www.mediawiki.org/

[16] https://criticalrole.fandom.com

[17] https://huggingface.co/datasets/Akila/ForgottenRealmsWikiDataset

JSON before being embedded in the overall JSON structure indexed by the page title. Each of the word and document embedding data sets (*FRW-W*, *FRW-D*, *FRW-FD*) have 2 different embeddings used. For word level embeddings CBOW and Skip-gram (Mikolov et al., 2013) are used. While PV-DBOW and PV-DM (Le and Mikolov, 2014) are used in document embeddings. Figure 1 shows the convergence of the Poincaré embedding data set, *FRW-PE*.

All of the data sets that use a JSON structure (*FRW-J*, *FRW-FJ*, *FRW-I*) use the same high level JSON schema. The pages are organized in a JSON array with *page* and *content* being the only two attributes in each element. The *page* attribute contains the article title while the *content* attribute contains the corresponding information extracted from the page. This information is in plain text format for *FRW-J* and *FRW-FJ*. In the case of *FRW-I*, it is a JSON dictionary containing the infobox content as the key-value pairs. Code 1 shows the top level JSON schema.

Code 1: JSON top level structure

```
[
  {
      "page": "page_title",
      "content": "page_content"
  },
  ...
]
```

Name	Description
FRW-P	Raw plain text corpus (no Markdown text markings)
FRW-J	A JSON structure with plain text indexed by article title
FRW-FJ	A JSON structure with only the first paragraph (plain text) of the articles indexed by article title
FRW-L	A directed graph indicating all the references in the articles to other articles
FRW-FL	A directed graph indicating the first references in the articles to other articles
FRW-CL	A directed graph indicating the category references in the articles to category pages.
FRW-I	A JSON structure for the Wikipedia infobox substructures indexed by article title
FRW-PE	Poincaré embedding of *FRW-FL*
FRW-W	2 Word2Vec models for *FRW-P* (CBOW and Skip-gram)
FRW-D	2 Doc2Vec models for *FRW-P* (PV-DBOW and PV-DM)
FRW-FD	2 Doc2Vec models for *FRW-FJ* (PV-DBOW and PV-DM)

Table 1: *FRW* data set

4 Use Case Analysis 1: Semantic Similarity Comparison

To illustrate both the consistency as well as the non-trivial nature of data sets we have collected, we have performed similarity calculations for a set of text pairs extracted from the data set using multiple different similarity metrics. By the high alignment of semantic similarity in similar perspective data sets, we show the consistency in our data sets. By the low alignment of the differing perspective data sets, we show that the individual data sets are not redundant and that they carry unique information that may not have overlaps with other data sets that we present in this work.

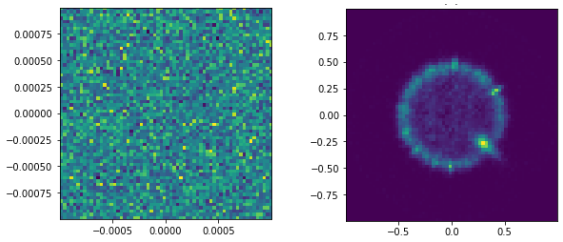
4.1 Text Pairs for Evaluation

To ensure that these different metrics are comparable, we have used the same set of text pairs for all of the similarity calculations. Hierarchical similarities are measured using article titles and the embedded vector distance based similarities are calculated using article contents. We use the *FRW-FJ* data set to generate the text pairs. Since the *FRW-FJ* data set is a subset of *FRW-J*, it ensures that a) all the pairs correspond to actual wiki articles b) has valid text content c) full document and first paragraph only data sets are available for different similarity calculations, and d) almost all nodes (page titles) are significant/"worthy of notice"[18] to the domain as per the Wikipedia guidelines.

4.1.1 First Link

First link is the first internal reference link (refers to another article in the same wikia) found in an article that is not a broken link or a miscellaneous link such as the pronunciation guide. According to the wikimedia guidelines, the lead section of a typical Wikipedia article contains links to other articles that provide context to the article in question i.e., the references in lead section point towards more generalized concepts and/or any other concepts related to the context of that article. We use this arrangement to measure the similarity or relatedness of topics. This leads to an interesting pattern where clicking the the first link of a random Wikipedia page and doing so repeatedly on the subsequent pages will 97% of the time (Lamprecht et al., 2016) lead to a cycle containing the article "Philosophy"[19] . The rest of these *first link*

[18] https://bit.ly/3S9HHbc
[19] https://bit.ly/3xyBnRf



(a) Initial (0 epochs) (b) After 50 epochs

Figure 1: Poincaré Embedding convergence

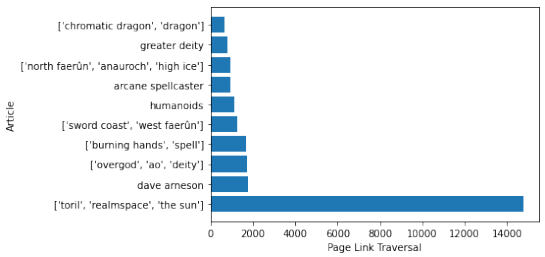


Figure 2: First link traversal graph

Statistic	Value
Total number of tokens (excluding titles)	9,189,536
Total number of tokens (including titles)	9,287,670
Total number of unique tokens	145,624
Total number of sentences	517,248

(a) *FRW-P*

Statistic	Value
Total number of articles	48,892
Average number of tokens per sentence	17.77
Average number of tokens per article	187.96
Average number of sentences per article	10.58

(b) *FRW-J*

Statistic	Value
Total number of articles	41,204
Total number of tokens	980,047
Total number of sentences	98,244
Average number of tokens per sentence	9.98
Average number of tokens per article	23.78
Average number of sentences per article	2.38

(c) *FRW-FJ*

Statistic	Value
Average number of attributes per infobox	40.54
Average number of completed (filled) attributes per infobox	10.40
Total number of articles containing infoboxes	35,923

(d) *FRW-I*

Statistic	Value
Total number of nodes	46,910
Total number of edges	570,857
Average number of edges per node	12.16

(e) *FRW-L*

Statistic	Value
Total number of nodes	43,329
Total number of edges	41,213
Number of nodes not referenced by others	34,881
Number of nodes with no references	2151

(f) *FRW-FL*

Table 2: Statistics of different sub data sets of the *FRWdataset*

traversals exhibit one of the following shortfalls: 1) contain no internal links, 2) contains a self loop, 3) ends up in an isolated tree, 4) form a cycle with a few other pages. The Forgotten Realms wiki also abide by the same principle. The *center* of the Forgotten Realms wiki universe is a cycle composed of the articles, "*Toril*", "*Realmspace*" and "*The Sun*". However, unlike in the case of Wikipedia, in Forgotten Realms wiki, this only applies to around 30.2% of the articles. Figure 2 lists the 10 most commonly traversed articles using this method. The ones enclosed in "[ ]" refer to cycles, for example ['*toril*', '*realmsphere*', '*the sun*'] refers to a first link cycle between the three corresponding articles.

4.1.2 Issues with Category Links for Semantic Similarity Evaluation

Even though it is the de-facto categorical hierarchy, there are many issues with using category links as a measure for semantic similarity. The most prominent bottleneck of *FRW-CL* data set is that it is mostly a flat hierarchy. So any set of node pairs would have almost identical distance measures no matter how different they are semantically. Secondly, the Categories are not consistent across all articles, i.e,. while some articles may have an abundance of Categories, others may have have little to none. Finally, Category pages do not necessarily have article content as a typical page does, hindering the ability to perform effective word and document embedding.

4.1.3 Generating Text Pairs for Evaluation

We created 1,000,000 unique text pairs using 41,000 nodes from *FRW-FL*. We have also ensured that there are no interchangeable duplicates. To ensure that the selected pairs have better representation, we have used a weighted random sampling technique with dynamically updated weights. The sampling was done with replacement. The probability of an item i getting selected for the sample pair set is given in Equation 1, where N is the total

number of pairs we generate and k_j is the number of times the jth item has already been selected.

4.2 Hierarchical Similarity Measures

We have used the *FRW-FL* data set as the base for similarity measures using hierarchical similarity evaluation methods. Although the *FRW-FL* is already devoid of any self-loops, there are cycles and isolated trees while also lacking a common root node. We process this and convert into a directed cyclic graph.

Let G be a graph in the set of disconnected graphs $G = (V, E) \in G_1, G_2, ..., G_n$ where $E \in E'$ and $V \in V'$. G_c represent a subgraph of a given graph $G_c = (V_c, E_c) \in G$ where $e_1, ..., e_n$ is a trail with vertex sequence $a_1, ..., a_n$ (cyclic graph). Then $\forall G \in G_1, ..., G_n$ apply Equation 2 to obtain the final unified graph, $G' = (V', E')$.

For the intermediate node v', we use a comma separated combination of the names of the nodes forming the cycle. Using an intermediate node helps us retain the relatedness they had with one another up to a certain degree before reaching the root node. We use algorithmic measures such as Wu and Palmer (1994) similarity metric and Jiang and Conrath (1997) distance measure, both of which use the Least Common Ancestor (LCA) as the basis for the calculations. Apart from this, we have also evaluated with the hierarchical embedding using Poincaré (Nickel and Kiela, 2017) method. While other embedding methods such as ones using metadata (Xing and Paul, 2017; Zhou et al., 2015) can be experimented with, we have chosen the Poincaré embedding since we are measuring hierarchical similarities. This allows us increase the comparability between the different types of measurements in our experiment.

It should be noted that, for the sake of this comparative analysis, we have converted the Jiang-Conrath distance measure (Jiang and Conrath, 1997) into a similarity measure ranging from 0 to 1 as shown in the Equation 3 where the $LCA(a, b)$ function returns the least common ancestor of the nodes a and b, the $IC(d)$ function returns the Information Content of the node d, and c_i is the node in the hierarchy representing the term t_i.

4.3 Embedding Based Similarity Measures

We have performed both word embedding on *FRW-P* corpus and document embedding for the *FRW-J* data set to create *FRW-W* and *FRW-D* data sets. Both of these data sets contain essentially the same content albeit the format. In addition, we have created *FRW-FD* data set using *FRW-FJ* which only contains the first paragraph of each page to evaluate the effectiveness of the first paragraph in comparison to the whole text. For embedding vector distance based similarity, we have used the *FRW-W* data set containing CBOW and Skip-gram (Mikolov et al., 2013) model embeddings. For word embedding, when a title is given the corresponding article is retrieved from *FRW-J*. Then for all the words in the article, the word vectors are fetched from the *FRW-W* data set and a single vector is obtained via average pooling. Cosine similarity is defined as shown in Equation 4, where t_i is a string in the domain and v_i is the corresponding vector in the embedded vector space.

$$P(i) = \frac{\sqrt{N} - k_i}{N\sqrt{N} - \sum_{j=1}^{N} k_j} \tag{1}$$

$$E' = \begin{cases} E \cup (root, v) & \text{if } deg^-(v) = 0 \text{ and } v \neq root \\ E \cup (root, v') \cup (v', v) & \text{if } v \in G_c \\ E & \text{otherwise} \end{cases} \tag{2}$$

$$JC\_s(t_1, t_2) = \frac{1}{1 + |2 \times IC(LCA(c_1, c_2)) - (IC(c_1) + IC(c_2))|} \tag{3}$$

$$cosine\_similarity(t_1, t_2) = \frac{v_1 \cdot v_2}{||v_1|| ||v_2||} \tag{4}$$

For Doc2vec (Le and Mikolov, 2014), we have used both *FRW-D* and *FRW-FD* data sets with each having a `PV-DBOW` and a `PV-DM` model (Le and Mikolov, 2014) resulting in four embedded models altogether. The four models are trained with the article title as the tag for the content of the article. Hence the document vector itself can be fetched directly from the model using the article title (text phrase).

We have briefly mentioned at the start of this section, we specifically used the first paragraph only text to assert for its goodness compared to the whole text. The rationalé for this as follows: as discussed in subsection 4.1.1, the first paragraph or the lead section of a wiki article is an abstractive summary of the entire article. Hence, if this showed comparable results to full text, the full text document embedding can be substituted by this. Which requires less computational resources due to its much smaller size. For comparison, the word count on data sets *FRW-J* and *FRW-FJ* are 9,189,536 and 980,047 respectively, which is a 10:1 compression ratio. Further, this would open the door to future in-domain text summarizing research.

				Hierarchical			Embedding					
				WP	JC	P	Word2Vec		Doc2Vec			
							(*FRW-J*)		(*FRW-FJ*)		(*FRW-J*)	
							CBOW	SG	DM	DBOW	DM	DBOW
Hierarchical			WP	1.0000								
			JC	0.6346	1.0000							
			P	0.0097	0.0624	1.0000						
Embedding	Word2Vec	*(FRW-J)*	CBOW	0.0581	0.0212	0.0013	1.0000					
			SG	0.0553	0.0188	-0.0043	0.9412	1.0000				
	Doc2Vec	*(FRW-FJ)*	DM	0.0040	-0.0298	0.0548	-0.0626	-0.0791	1.0000			
			DBOW	0.0466	0.0155	0.0359	0.0362	0.0222	0.5691	1.0000		
		(FRW-J)	DM	0.0259	0.0186	0.0175	-0.1865	-0.2593	0.1724	0.1484	1.0000	
			DBOW	0.0361	0.0287	0.0453	-0.0896	-0.1601	0.1511	0.1825	0.5493	1.0000

Table 3: Pearson Correlation for the pairwise text similarities across multiple similarity metrics: 1) WP – Wu & Palmer similarity 2) JC – Jiang-Conrath Similarity 3) P – Poincaré metric 4) CBOW – Continuous Bag of Words 5) SG – Skip-gram 6) DM – Distributed Memory 7) DBOW – Distributed Bag of Words

5 Use Case Analysis 2: Forgotten Realms Free Text Generator

Our work on free text generation[20] is divided into two phases. In the first phase, we try to train a GPT-Neo (Gao et al., 2020) model capable of generating text in the Dungeons & Dragons (Forgotten Realms) domain. Once the text generator was in place, the second phase of the work involves classifying in-domain named entities present in the generated text. We used the trained model from step 1 and fine-tuned it using data from the Forgotten Realms Wiki JSON (*FRW-J*) database. We tokenized each of the the documents in the *FRW-J* data set separately. The model that was used contained 1.3 billion parameters. Using this fine-tuned model, we were able to generate texts that are grounded in the D&D domain.

5.1 Generated Text

An example text generated using the Forgotten Realms Free Text Generator we have trained using the *FRW* data set is shown in Figure 3. To make comparisons as to the effect of training on the *FRW* data set (fine-tuning, we have also listed some text samples generated using the unmodified (vanilla) version of the same language model (Figure 4). All of these samples were generated based on the same input prompt "*A Bard, a Wizard, and a Paladin walk into a tavern*".

We can observe, that the vanilla model refers to real world entities such as *Northumberland*, *Emerald Tablet*, and *Alderney*. It is also worth noting that the vanilla model refers to other works of literature as can be observed in Figure 4a where *Order of the Phoenix*[21] was mentioned. Here, the entity was correctly identified as a group or a society in accordance with its literature and has been used appropriately in the generated text.

In comparison, the text samples generated from

A Bard, a Wizard, and a Paladin walk into a tavern, and debrief the group on their travels. The trio discuss their lives and lives. Bard: "In actuality, the only things I have gained in my life that I truly value are knowledge and comradeship, both of which seem more valuable than wealth." Wizard: "Clan Bloodforge was once my clan; I guess I'm considering it my clan after death to boot.". Paladin: "There are two separate communities of people living in this land, one in the North and another in the South, and the northern people came to the west to the Sword Coast about five centuries ago." One of Bloodforge's leaders meets with the leaders of Clan K'yorlhan, the tribe from Tethyr that fights the bard raider bandit leader Korgh, and discovers that most of the members of Clan Bloodforge do not want to join Clan K'yorlhan. In the end, he leads them to the hidden Bloodforge, and joins them in battle. Monk: "Korgh is a madman, but he's got a point about the invaders.". Bard: "That's true; otherwise, why'd he attack all across Tethyr? That's something Clan Bloodforge are going to have to consider." During the homecoming party, Bard, Wizard, and Paladin trade news of their adventures with each other. Bard: "Clan Bloodforge and Clan K'yorlhan, now that's something I'm going to have to watch carefully.". Wizard: "I'd better keep in mind that Clan K'yorlhan is already taken.". Paladin: "We were also allied with the tanarukks of Norland, but I hope I'm wrong, because they keep countless spies hidden around here.". Paladin - Son of Blackstaff - joins Bloodforge as a young man. Bard: "Artemis Entreri is not the most forward-thinking of consorts, but still, he's a good man." Paladin - A young warrior from the north. Bard: "Welcome to Bloodforge." Paladin - A young man from the west. Bard: "The Wyrmskull Thron is a noble house in Calimshan.". Wizard: "I'd have to say that the War Wizard of Calimshan is a good man."

Figure 3: Sample text generated from the Forgotten Realms Free Text Generator (trained on *FRW* data set)

[20] https://huggingface.co/Akila/ForgottenRealmsFreeTextGenerator

[21] https://bit.ly/3Z4vOFI

A Bard, a Wizard, and a Paladin walk into a tavern, It was the middle of the night, and - like the rest of the town - Bard, Wizard, and Paladin had already found a place that they could call their own. It was a place where they could help young minds and provide a positive role model for young people. "Do you mind if I sit?" a BaldDwarf asked. "The kids love your games," the Bard replied. The Bard was an active member of the Order of the Phoenix, the Order of the Emerald Tablet, an Amity Group, the Viscount of Ce's Guild, and anointed as a Paladin by a bard named D'Cel. He was also a Wilderness Shield Master and a beardless man. He couldn't wear any clothing he didn't own, because - to him - the Wilds were his home. The Wizard and Paladin should never have come into a tavern, but the Bard was a musketeer, and a Tourney-member. He was just trying to get a look at the other two.

(a) Sample 1

A Bard, a Wizard, and a Paladin walk into a tavern, in rural Northumberland. The Bard and the Wizard pay and are received in the company of a Salmon, and the Paladin is greeted as a traveling merchant by merchants and peasants alike. The Bard is either still in his role as an entertainer, or has passed into the service of some people who know what they are doing. The Wizard has given up his peddling. He is now performing services as an entertainer, singing and dancing and acting as peacemaker or mediator. The Paladin has come here to seek the largess of the Bard, to borrow from him with interest, and to bestow upon him the reward of his protection and the services of his wizardry. The Bard has accepted his terms, and has given his lodgings - a mansion built in a ruined mill, called Cinann House, on the banks of the River Barrow, once belonging to the Earl of Northumberland, who was the brother of stood as guardian of the booth of Alderney who came to trade with Scandinavia to the east.

(b) Sample 2

Figure 4: Sample text generated from the vanilla text generator

the Forgotten Realms Free Text Generator show more D&D domain specific characteristics. It uses established entities from the D&D lore such as *Bloodforge*, *Norland*, and *Calimshan*. It also identifies and uses *Norland* as a location which is part of *Sword Coast* in accordance with the domain data. Another thing to note is that the model even generates fake names and characters that are not mentioned in the data set such as *Korgh* and *K'yorlhan* that fit in well with the fantasy genre and build narratives around those characters. Despite, some minor issues with cohesion, overall, it generates satisfactory results.

5.2 Named entity classifier

Although the Forgotten Realms Free Text Generator managed to create text based on D&D domain, when observed carefully by domain experts, there were some inconsistencies with the established lore of the domain. For example, according to the Forgotten Realms lore, *Artemis Entreri* is an *assassin* and not a *consort* while the *Wyrmskull Throne* is a physical object, not the name of a house. To assess the categorical validity of the named entities generated in the text, we have trained the same model on a data set where each row contains a full text generated by the Forgotten Realms Free Text Generator, a named entity in that text, and the matching category extracted from the *FRW-I*. By performing 5-fold cross validation, we were able to train our model to identify the category for a named entity in a generated text. For this basic analysis, we created 100 instances each containing on average 351.4 words and 19.07 sentences. The model was capable of predicting the correct category with 99.3% accuracy on average, attesting to the power of GPT-Neo (Gao et al., 2020) as well as the potential in domain specific text generation. Since the correct classifications are a set of rules declared by the *FRW-I* data set, and the GPT-Neo model uses a data driven training approach, this can be the first step towards creating a conditional text generator that will bridge the traditional rule-based text generation methods and the novel data-driven methods.

As for the vanilla model, we were unable to perform any meaningful entity classification in relation to the D&D domain, as there were no D&D specific entities that were mentioned in the generated text.

6 Conclusion and Future Work

When performing domain specific text generation, it is important that the output stays true to source material. For this, sufficient data from the domain is required. Other than the raw corpora, additional supplementary data structures such as tabular summaries can help ease the process of evaluating the consistency of generated text in context to the domain. In this paper we present a data set based on the D&D domain and a system that is capable of generating free text that stays consistent to the domain. Apart from this, the named entity classifier model shows promising results as part of a guided text generation system. We hope that the *FRW* offers a convenient unique data set for the D&D domain. We hope that the data set can also be enhanced in the future including an improved linked list to measure evaluation.

References

Amin-Nejad, A., Ive, J., et al. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th LREC*, pages 4699–4708.

Barzilay, R. and Lapata, M. (2005). Collective content selection for concept-to-text generation. In *Proceedings of the conference on Human Language Tech-*

nology and Empirical Methods in Natural Language Processing, pages 331–338.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Brown, T., Mann, B., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Busemann, S. and Horacek, H. (1998). A flexible shallow approach to text generation. *arXiv preprint cs/9812018*.

Chen, H., Takamura, H., and Nakayama, H. (2021). Scixgen: A scientific paper dataset for context-aware text generation. *arXiv preprint arXiv:2110.10774*.

Cozza, V., Petrocchi, M., and Spognardi, A. (2016). A matter of words: NLP for quality evaluation of Wikipedia medical articles. In *International Conference on Web Engineering*, pages 448–456. Springer.

Crawford, J., Wyatt, J., Schwalb, R. J., and Cordell, B. R. (2014). *Player's handbook*. Wizards of the Coast LLC.

Dale, R., Geldof, S., and Prost, J.-P. (2003). Coral: Using natural language generation for navigational assistance. In *Proceedings of the 26th Australasian computer science conference-Volume 16*, pages 35–44.

de Silva, N. and Dou, D. (2021). Semantic oppositeness assisted deep contextual modeling for automatic rumor detection in social networks. In *Proceedings of the 16th Conference of the EACL: Main Volume*, pages 405–415, Online. ACL.

de Silva, N., Dou, D., and Huang, J. (2017). Discovering Inconsistencies in PubMed Abstracts Through Ontology-Based Information Extraction. In *Proceedings of the 8th ACM BCB*, pages 362–371.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2020). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, page 113679.

Ferrari, A., Donati, B., and Gnesi, S. (2017). Detecting domain-specific ambiguities: an nlp approach based on wikipedia crawling and word embeddings. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, pages 393–399. IEEE.

Ferschke, O. (2014). The quality of content in open online collaboration platforms: Approaches to nlp-supported information quality management in wikipedia.

Ferschke, O., Daxenberger, J., and Gurevych, I. (2013). A survey of nlp methods and resources for analyzing the collaborative writing process in wikipedia. In *The People's Web Meets NLP*, pages 121–160. Springer.

Fu, Z., Tan, X., Peng, N., et al. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference*, volume 32.

Gao, L., Biderman, S., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ghalandari, D. G., Hokamp, C., et al. (2020). A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Gygax, G. and Arneson, D. (1974). *dungeons & dragons*, volume 19. Tactical Studies Rules Lake Geneva, WI.

Hoffmann, R., Zhang, C., and Weld, D. S. (2010). Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 286–295.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Kreutzer, J., Caswell, I., et al. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *TACL*, 10:50–72.

Lamprecht, D., Dimitrov, D., Helic, D., and Strohmaier, M. (2016). Evaluating and improving navigability of wikipedia: a comparative study of eight language editions. In *Proceedings of the 12th international symposium on open collaboration*, pages 1–10.

Lange, D., Böhm, C., and Naumann, F. (2010). Extracting structured information from wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1661–1664.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Lebret, R., Grangier, D., and Auli, M. (2016). Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Liang, P., Jordan, M. I., and Klein, D. (2009). Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 91–99.

Liu, T., Wang, K., Sha, L., et al. (2018). Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference*.

McRoy, S. W., Channarukul, S., and Ali, S. S. (2003). An augmented template-based approach to text realization. *Natural Language Engineering*, 9(4):381–420.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nastase, V. and Strube, M. (2013). Transforming wikipedia into a large scale multilingual concept network. *AI*, 194:62–85.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In Guyon, I., Luxburg, U. V., Bengio, S., et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

Radford, A., Wu, J., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rajapaksha, I., Mudalige, C. R., et al. (2020). Rule-Based Approach for Party-Based Sentiment Analysis in Legal Opinion Texts. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 284–285. IEEE.

Rajpurkar, P., Zhang, J., et al. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 5121–5134, Online. ACL.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Reiter, E., Sripada, S., Hunter, J., et al. (2005). Choosing words in computer-generated weather forecasts. *AI*, 167(1-2):137–169.

Sanchez-Perez, M. A., Sidorov, G., and Gelbukh, A. F. (2014). A winning approach to text alignment for text reuse detection at pan 2014. In *CLEF (Working Notes)*, pages 1004–1011.

Srinivasan, K., Raman, K., et al. (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Stent, A., Prasad, R., and Walker, M. (2004). Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the ACL)*, pages 79–86.

Sugathadasa, K., Ayesha, B., et al. (2017). Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6. IEEE.

Sugathadasa, K., Ayesha, B., et al. (2018). Legal Document Retrieval using Document Vector Embeddings and Deep Learning. In *Science and information conference*, pages 160–175. Springer.

Turner, R., Sripada, S., and Reiter, E. (2009). Generating approximate geographic descriptions. In *Empirical methods in natural language generation*, pages 121–140. Springer.

Whitten, S. (2021). Dungeons & Dragons had its biggest year ever as Covid forced the game off tables and onto the web.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the ACL*, pages 118–127.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Xing, L. and Paul, M. J. (2017). Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49, Copenhagen, Denmark. ACL.

Zesch, T. and Gurevych, I. (2007). Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 1–8.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth LREC*.

Zhang, H., Song, H., Li, S., et al. (2022). A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Zhang, T., Kishore, V., Wu, F., et al. (2019). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, X., Jiang, Y., Shang, Y., et al. (2021). DSGPT: Domain-Specific Generative Pre-Training of Transformers for Text Generation in E-commerce Title and Review Summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2146–2150.

Zhou, H., Chen, L., et al. (2015). Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP (Volume 1: Long Papers)*, pages 430–440, Beijing, China. ACL.