# Exploring Narrative Structure with MMORPG Quest Stories

**Emmett Tomai, Eric Martinez and Lucian Silcox**
University of Texas – Pan American, 1201 W. University Dr. Edinburg, TX, 78539
tomaie@utpa.edu, emmartinez4@broncs.utpa.edu, lucian.silcox@gmail.com

### Abstract

In this paper we present a corpus of quest stories taken from a popular commercial Massively Multiplayer Online Role-Playing Game (MMORPG). These stories are open-ended narrative, but anchored to formal, in-game actions and entities, providing valuable constraint as a narrative corpus. We present two preliminary experiments establishing baselines for evaluating similar patterns and content across the corpus.

## Introduction

The ability to reason about narrative has long been considered a fundamental cognitive tool and an important challenge for artificial intelligence (cf. Schank 1977; Bruner 1991). But even though video games are well established as both a growing storytelling medium (cf. Jenkins 2006) and a powerful domain for AI research (cf. Laird 2001), we are not aware of any work that has used video game stories as a corpus to develop computational models of narrative. In this paper, we present a corpus of short stories taken from a Massively Multiplayer Online Role-Playing Game (MMORPG). These *quest stories* are delivered to players by in-game characters as first-person narrative. Each corresponds to a structured set of *objectives* that the player must achieve in the game to make progress. The quest stories allow the designers to flesh out the setting and characters, and to motivate and justify the objectives. They are open-domain stories, but they all motivate activities supported by the game simulation, such as combat, exploring and gathering materials. They also adhere to genre tropes, deliberately encouraging a sense of familiarity and comfort in the players. As a result, these quest stories demonstrate at fairly large scale how human storytellers can turn very similar events into very different stories. We believe that this makes a promising corpus for modeling and learning structures and patterns in narrative discourse.

Deep narrative structures were a large part of early work on narrative understanding (cf. Dyer 1983), but were intractable at scale. Statistical NLP has proven robust at very large scale (cf. Voorhees & Buckland, 2009), but in part by omitting deep understanding. Recently, however, there has been more work using scalable, shallow statistical tools to identify patterns and map to script-like narrative event structures. Chambers and Jurafsky (2009) showed that next-event predictions could be extracted from a gigaword news corpus. Li et al. (2012) used statistical clustering over crowd sourced short narratives on common events (e.g. a movie date) to automatically generate scripts. In contrast to those corpora, our quest story corpus covers a wider range of creative narrative techniques, while still having useful constraint to in-game objectives. This can enable investigation beyond sequences of events to the way that they are presented in real, commercially relevant stories. Here we present two preliminary experiments to characterize similarities among the quest stories and establish baseline evaluations for future work.

## Activity Identification

We collected over 7000 in-game quests from *World of Warcraft*, copyright Blizzard Entertainment[1], which are publically available on many reference websites for players. Each quest in the corpus includes the quest story text, a brief objective statement and bullet points that explicitly detail the objectives. The majority of the bullet points consist of a single verb indicating the player activity needed to fulfill the objective, and unique identifiers for target in-game entities. A human player reading a quest story will understand generally what activities are being proposed, even without the explicit objectives. We believe that different narrative structures are used to motivate those

[1] http://blizzard.com

*Table 2. Comparison of classifiers and features for KILL label classification.*

| | All Words | | | Verbs Only | | | Verbs Only (Stemmed) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F-Score** | **Precision** | **Recall** | **F-Score** | **Precision** | **Recall** | **F-Score** |
| **Naïve Bayes** | 0.679 | 0.670 | **0.674** | 0.645 | 0.552 | 0.595 | 0.719 | 0.639 | **0.676** |
| **Random Forest** | 0.752 | 0.635 | **0.688** | 0.669 | 0.572 | 0.616 | 0.826 | 0.487 | 0.613 |
| **Gradient Boosted** | 0.787 | 0.440 | 0.564 | 0.704 | 0.390 | 0.502 | 0.750 | 0.496 | 0.597 |

few activities in many different ways. Our first question, therefore, is to what extent the shallow, unstructured semantic content of the stories is sufficient to identify those activities. A strong positive answer to that question would diminish the usefulness of this corpus in exploring narrative structures.

In this experiment, the objective text and bullet points were used to label the activities, and only the story text was used in the classification. By comparing the verbs used in the objective text and bullets across the entire corpus, we identified nine initial activity labels. Each quest may have any number of activities. Three of the nine labels were omitted due to small positive example count, and a fourth was so semantically broad as to be meaningless. The remaining five labels and the number of instances in the corpus are given in Table 1.

*Table 1. Quest activity labels and annotation counts.*

| Label | Instances | Description |
|---|---|---|
| KILL | 3110 | Kill a NPC |
| GATHER | 1994 | Collect Items |
| SPEAK | 1171 | Social interactions with a NPC |
| USE | 337 | Use an Item |
| HELP | 318 | Protect/Save an NPC |

For each activity, we preformed an independent binary classification to test the ability to predict whether the quest story implies it. This is a document classification problem, so we tested it with three common classifiers in that space: Multinomial Naive Bayes, Random Forest of Trees and Gradient Boosted Trees. Although it has been shown that the tree methods can outperform Naive Bayes (Caruana & Niculescu-Mizil 2006), we expected that the relatively small sample sizes would be a greater factor. The features were unstructured bag-of-words frequency counts covering the vocabulary of the training set (stopwords and low-frequency terms filtered). We tested three feature set conditions: All Words, Verbs Only and Verbs Only (Stemmed). Verbs were tested because of their importance in narrative structure. However, only part-of-speech tagging was used to extract the verbs, to emphasize simple, shallow techniques. That combined with the relative shortness of the stories lead us to hypothesize that the All Words condition would be most effective.

Table 2 shows the results for predicting the KILL activity for each condition. For each, we used a random 50/50 split of the corpus into testing and training, and calculated the accumulated precision, recall and f-score over 10 trials. The top three f-scores are highlighted in bold. We note first that the Gradient Boosted condition widely underperforms the other two classifiers. Boosting methods are susceptible to noise (Bootkrajang & Kabán 2013), and the large number of different words combined with short stories (low occurrences) may have contributed to that. Among the top two classifiers, there is a statistically significant difference between the results for the All Words feature set (1-tailed student's paired t-test, $p<<0.01$), but it is a negligible difference from the point of view of an initial exploration. In general, the two Verbs Only conditions significantly under-performed All Words (1-tailed student's paired t-test, $p>0.01$), except in the case of Naive Bayes with Verbs Only (Stemmed). As the Verbs Only feature vectors are much more compact, this is a strong candidate for further exploration.

Table 3 shows the best results for all five labels. The Naive Bayes classifier with the All Words feature set was significantly the best performing condition for all the labels apart from KILL. As expected, the size of the positive samples strongly impacts the quality of classification. Following these results, we attempted to add sentiment analysis features as additional semantic information, but saw no significant improvements. We also ran Latent Semantic Analysis on the stories, transforming them into principle component vectors. It appeared in visualization that the components for verb co-occurrence clustered strongly with activity labels, but using the component vectors as features also failed to improve classification.

*Table 3. Naïve Bayes classification results for All Words.*

| Label | Precision | Recall | F-Score |
|---|---:|---:|---:|
| KILL | 0.752 | 0.635 | 0.688 |
| GATHER | 0.629 | 0.626 | 0.627 |
| SPEAK | 0.545 | 0.512 | 0.528 |
| USE | 0.215 | 0.171 | 0.190 |
| HELP | 0.417 | 0.251 | 0.312 |

We present these results as a useful baseline, at different sample sizes, for the extent of the predictive power in the shallow, unstructured semantic content of the stories. We conclude that there is room for productive exploration of more sophisticated structural models that can be evaluated against this data.

## Sentence Ordering

The second question we ask is whether there are common, consistent patterns in the order in which events are presented. If there are, then it should be possible to identify more and less likely sequences of sentences, even without deep understanding of what is taking place. In this experiment, we use shallow features to identify correct sentence orderings from incorrect. For a story with *n* sentences, there are $n^2 - 2n + 2$ alternative orderings that can be created by repositioning one sentence. While some alternative orderings would still form coherent stories, these should be rare given the short length of the quests. Once the alternative orderings are generated as negative examples, the task is to identify the correct ordering of a set of sentences, given all the possible orderings for a quest story. For any random half of this corpus, the expected accuracy for randomly guessing on this task is 0.20.

Our first approach, Paired Sentence Order, trains on a quest story by taking all possible sentence pairs within that story, labeled according to their correct ordering as *before* or *after*. It uses the same unstructured bag-of-words features for All Words and Verbs Only as in the first experiment. A Gaussian Naïve Bayes classifier was trained on a random 50% of the corpus. For testing, the trained model is used to predict the probability of each sentence pair order in each candidate ordering. The score for an ordering is the sum of the predicted probabilities for that ordering. We hypothesized that even though verb chains are key to narrative flow, the Verbs Only condition would under-perform the All Words condition due to the sparseness of verbs in the sentences.

We compared these conditions with a more sophisticated metric of Verb Chain Similarity. Li et al (2006) demonstrated the effectiveness of a sentence similarity metric that combines semantic and ordering similarity metrics. The former is calculated as cosine similarity between lexical semantic content vectors (generated using corpus information content and WordNet path lengths). The latter also uses vector-based similarity, for a novel vector representation of word order. This metric easily applies to verb chains, as sequences of words with both semantic and ordering similarity. Verb chains were extracted using simple part-of-speech tagging. For a given quest story, the feature vector is the sentence similarity between that story's verb chain and each other story's verb chain in the training set. A Gaussian Naïve Bayes classifier was trained with the correct orderings as positive samples, and the alternative orderings as negative samples.

The three conditions were run 10 times each with a random 50% training/testing split, and the mean accuracy and standard deviation are reported in Table 4. As expected, Verb Only significantly underperforms All Words (1-tailed student's paired t-test, p<<0.01). Verb Chain Similarity significantly outperforms the simpler conditions (1-tailed student's paired t-test, p<<0.01) in spite of using only rough verb information. This suggests that there are significant similarities in the presentation of verbs in the corpus which can be explored and exploited. It also provides a better baseline for future improvement.

*Table 4. Accuracy in the Sentence Ordering task.*

| | **Accuracy** | **Std. Dev.** |
|---|---|---|
| **Verb Chain Similarity** | **0.442** | 0.030 |
| **Paired Sentence Order, All Words** | 0.369 | 0.008 |
| **Paired Sentence Order, Verbs Only** | 0.331 | 0.006 |

## Conclusion

Quest stories combine computationally friendly constraints and in-game meta-knowledge with open-ended narrative. They demonstrate how many different versions of the same stories can be told. The results we have presented here using accessible, shallow techniques aim to begin characterizing the consistent patterns of content across the corpus. The next step for this project is to use those insights to guide annotation efforts for deep structural elements in the stories.

## References

Bootkrajang, J., & Kabán, A. (2013). Boosting in the presence of label noise.*arXiv preprint arXiv:1309.6818*.

Bruner, J. (1991). The narrative construction of reality. *Critical Inquiry*, 18, 1–21.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*.

Chambers, N., & Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*.

Dyer, M. *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. Cambridge, MA: MIT Press, 1983.

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. NYU press.

Laird, J., & VanLent, M. (2001). Human-level AI's killer application: Interactive computer games. *AI magazine*, *22*(2), 15.

Li, Boyang, Stephen Lee-Urban, D. Scott Appling and Mark O. Riedl. (2012). Crowdsourcing Narrative Intelligence. *Advances in Cognitive Systems*, Palo Alto, California.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, *18*(8), 1138-1150.

Schank, R., and R. Ableson. (1977) *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.