Environmental Audio Immersion from Fictional Narratives

Leon Derczynski, Nanna Inie

University of Sheffield
S1 4DP, UK
leon@dcs.shef.ac.uk

Aarhus University
Aabogade 34, 8200, Denmark
inie@cavi.au.dk

Abstract

Experiencing a narrative can be made more enjoyable and powerful by increasing the level of the immersion in the story. This can be achieved through contextual cues, such as sounds and images. In order to provide these cues automatically, we attempt to detect parts of the environment described in a fictional narrative. We present machine learning approaches to this problem, compare them against to a rule-engineered baseline, and evaluate them over a dataset of fictional texts. In addition, we analyse the performance of this system over the output of a text-to-speech system taking the same narratives as input. This audio dataset is made available. The system is able to recognise environmental context and respond appropriately.

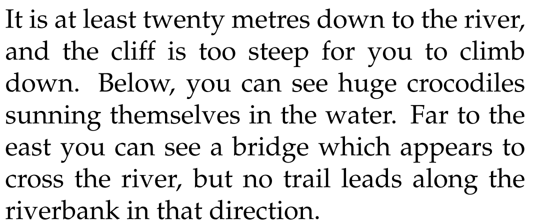
1. Introduction

Non-verbal cues can improve the experience of a narrative. For example, adding sound qualitatively impacts the recipient's mood, and is useful for increasing immersion and positive narrative experiences (Ermi and Mäyrä, 2005; Madden and Logan, 2009; Huiberts, 2010). An immersive environment is most engaging when participants believe their actions affect the environment; reliable and responsive cues such as sounds are important for this (Bobick et al., 1999). However, providing such cues can be an expensive manual process, and many textual narratives do not have accompanying contextual media such as background sounds or images. An automatic approach could be taken to building such context. The research question we address is: how can we automatically identify in narratives environments and events that can be associated with external cues?

Taking narratives as input, we attempt to determine what the appropriate environmental component of the context is, which can then be used to provide cues (such as ambient sounds). There is a diverse range of potential environments from which cues can be generated, even in a constrained corpus of narratives. In addition, effects (e.g. video or audio cues) are required for surprise events described in a story, like an approaching horse or a thunderstorm. This is similar to streamed topic extraction (Allan, 2002; Preotiuc-Pietro et al., 2012). We determine a corpus, frame the task by giving a specific set of environments for which cues are available, and then evaluate system performance over this dataset. Finally, we create a new language resource of audio recordings and automatic transcriptions, used to evaluate the system on input closer to that it might receive when operating in real-time.

2. Background

- NANNA CLEVER WORDS HERE -
What's "immersion"?
Why should we care?
What's important when creating environmental cues for improved narrative immersion?



It is at least twenty metres down to the river, and the cliff is too steep for you to climb down. Below, you can see huge crocodiles sunning themselves in the water. Far to the east you can see a bridge which appears to cross the river, but no trail leads along the riverbank in that direction.

Figure 1: Example narrative text

3. Method

In general, we cast the problem as one of document classification. Some special characteristics are included, namely that documents can have more than one class; and that documents may have not belong to any class at all, i.e., not every text requires or activates an environmental cue.

3.1. Dataset

Our ideal requirements are that a dataset be structures as instances, each being a focused passage of text, with most containing descriptions of environments or events. For this dataset, we used paragraphs from a set of role-playing books.An example text is given in Figure 3. Each paragraph is set in a distinct location, each possibly combined with the description of an event, making manual environment annotation simple and distinct. In comparison, the bounds of passages of text relevant to a particular cue are harder to detemine in running narrative text from e.g. a fictional novel . The set of potential environmental items is:

- **Environments:** Mountain; Hill; Forest; Swamp; Meadow; Road; Town; Crowd; Tavern; Underground.
- **Ambience:** Windy; Blizzard; Rain; Lightning; Stream; River; Campfire; Night.
- **Events:** Trotting horse; Galloping horse; Many galloping horses; Thunder.

Paragraphs were labelled with one or more of these labels by a human annotator. In total, 203 paragraphs were labelled, with 13884 words.

3.2. Spoken dataset

An eventual use of this system is to provide automatic sound effects in real-time for a story that is read out loud by a human narrator. Accordingly, we attempt to train and evaluate the system based on the output of a speech recognition system. This requires an audio dataset and the automatic transcription of audio content, and the subsequent analysis of the text output.

Paragraphs were read by an English native speaker, and recorded. A speech recognition system (Lamere et al., 2003) then interpreted these readings and generated a textual version of each paragraph. The labels used in the text input corpus were then associated with these outputs. This constitutes the transcribed dataset.

3.3. Baseline

As a baseline, we use a gazetteer trigger words that match the name of the cue. If a word occurs in a candidate passage, then that cue is triggered. For example, if there is a cue named "swamp", the passage "He marches through the swamp, sweating from his brow" will trigger that cue using the baseline system.

3.4. Features and Classification

It is better to provide environmental responses as soon as possible after input. Therefore, we break paragraphs into sentences and label at sentence level, allowing a response to be given after processing a sentence instead of waiting for the paragraph to end. To build sentence-level representations, we use the sentence2vec tool to map candidate sentences into a $k-$dimensional continuous embedded space. This relies on the word2vec tool, which takes large tracts of unlabeled text as input and builds vector representations of words based on their distributional properties.

The genre used to generate these had a significant effect on result quality. For example, using word2vec embeddings generated from two billion Google News articles, the top-ranking non-punctuation match for the sentence *"The going underfoot becomes muddier, until eventually you reach an area where bulrushes tower over your head."* was the phrase *Maria_Sharapova* – not an intuitively relevant result. News is intrinsically constrained in topic and style, and out-of-domain for fictional narratives. Further, with creative writing, we expect more out-of-vocabulary words originating from authors'/speakers' extended use of language when conjuring imagery. To this end, we anticipate that representations which rely on lexical items directly may suffer from degraded performance.

Therefore, we used embeddings learned from the multi-topic Brown corpus (Francis and Kucera, 1979), which includes large amounts of narrative and fictional text. The sections used were F (Popular Lore), K (Fiction: General), and N (Fiction: Adventure and Western). In addition, we used a digital copy of *The Lord of the Rings*. This yield a sample of roughly 820K words over which distributional embeddings were induced.

We frame this as a multi-label, multi-class learning problem. That is, there are many potential classes in terms of sounds to be played, and also, more than one of these may apply to a given input. We take a one-vs-all approach to classification, training a binary classifier for each label, and apply all classifiers to a sentence as it comes in.

The Stanford PTB tokeniser is used for splitting texts.

4. Evaluation

We note that false positives have a higher penalty than false negatives. It is disruptive to receive the wrong cue. For example, hearing battle noises when one should hear a babbling brook (false positive) is detrimental, and worse than hearing nothing (false negative). In order to bias classifications in this direction, we experiment with the SVM Cost parameter (Morik et al., 1999).

Further, we represent this immersion disruption in the evaluation measure. Given precision P and recall R, typically an F-score is drawn from F_β with $\beta = 1$.

$$F_\beta = (1 + \beta^2)\frac{PR}{(\beta^2 P) + R} \qquad (1)$$

When $\beta = 1$, precision and recall are balanced in a harmonic mean, e.g. F1-score. That is, false positives and false negatives impact results equally. To score away from false positives, i.e. wrong cues, we set $\beta = 2$.

Firstly, we try a bag-of-words representation using a multinomial Bayes classifier. In addition, we represent paragraphs as average vectors in n-dimensional space by taking the cosine product of embeddings of words in the paragraph, and then learn a binary SVM for each cue based on these representations.

Note that there is an assumption of orthogonality here which is not necessarily appropriate: if "oncoming horses" is the target label, a classifier that returns "trotting horses" does not really mis-perform as severely as one that returns "torrential rain".

5. Results

Our evaluation is using precision and recall. This is not a plain multi-way classification task; spurious cues and missed cues are both possible, as are multiple cues per text passage.

How to present these?

Results are given in Table 4..

The effect of tuning the cost parameter is shown in Table 4..

error analysis

per-class accuracy (coarse, fine)

6. Related Work

There is extensive literature on data-intensive approaches to document classification (Sebastiani, 2002). Given that the representation chosen is dervied from neural net approaches, it may improve things to use a neural net for the classification itself. We already know that neural nets are good at binary doc classification (Derczynski, 2006). Prior work has also examined document classification using SVM (Isa et al., 2008). However, little work has address the case of streaming classification of small documents, being primarily focused on batch indexing and retrieval tasks.

Approach	Precision	Recall	F_1 score	F_2 score
Baseline: trigger words				
BoW + Multinomial Bayes				
s2v + SVM				

Table 1: Classification accuracy

Cost C	Precision	Recall	F_1 score	F_2 score
0.1				
0.2				
0.5				
0.8				
1.0				
1.25				
1.5				
2.0				
2.5				
5.0				

Table 2: Modulating the cost function to reduce false positives; SVM with s2v representation

7. Conclusion

Acknowledgments This research has received partial support from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 611233, PHEME, and also from The Danish Council for Strategic Research under grant agreement No. 1311-00001B, CIBIS.

8. References

James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer.

Aaron F Bobick, Stephen S Intille, James W Davis, Freedom Baird, Claudio S Pinhanez, Lee W Campbell, Yuri A Ivanov, Arjan Schütte, and Andrew Wilson. 1999. The kidsroom: A perceptually-based interactive and immersive story environment. *Presence: Teleoperators and Virtual Environments*, 8(4):369–393.

Leon Derczynski. 2006. Machine learning techniques for document selection. Master's thesis, University of Sheffield.

Laura Ermi and Frans Mäyrä. 2005. Fundamental components of the gameplay experience: Analysing immersion. *Worlds in play: International perspectives on digital games research*, 37.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Sander Huiberts. 2010. *Captivating sound the role of audio for immersion in computer games*. Ph.D. thesis, University of Portsmouth.

Dino Isa, Lam Hong Lee, V Kallimani, and Rajprasad Rajkumar. 2008. Text document preprocessing with the Bayes formula for classification using the support vector machine. *Knowledge and Data Engineering, IEEE Transactions on*, 20(9):1264–1272.

Paul Lamere, Philip Kwok, William Walker, Evandro B Gouvêa, Rita Singh, Bhiksha Raj, and Peter Wolf. 2003. Design of the CMU Sphinx-4 decoder. In *Proc. INTERSPEECH*.

Neil Madden and Brian Logan. 2009. Collaborative narrative generation in persistent virtual environments. In *Proc. AAAI*.

Katharina Morik, Peter Brockhausen, and Thorsten Joachims. 1999. Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on real-time analysis and mining of social streams*.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.