IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791(Impact Factor: 1.479)
www.ijreat.org

Storyboard-A new Approach for extraction of key scenes from novels

Aarthika.R[1], Malathy.C[2]

[1]CSE Department, SRM University, Kanchipuram, Tamilnadu, India

[2]Professor, CSE Department, SRM University, Kanchipuram, Tamilnadu, India

Abstract

Novel-based movies brings large expectations leading to a much greater hype on novel-based games creating an intense curiosity among people who expect it to be as real as possible and be in par with the content. This has challenged the game developers the need to create the games that have the maximum possibility of satisfying the content of the novels. Here, it is proposed to extract the key information's from novel. This is done by parsing and tagging the entire document with the help of POS taggers such as Monty-lingua .Tf-idf frequency algorithm method is used to calculate the frequency of each term in the document and the terms with lesser frequency are discarded. Clustering technique is used to extract the pertinent elements. The information is then presented in a timeline format, which is used as an input for Storyboarding.

Keywords: *Parsing, Tagger, Natural language processing, Extraction.*

1. Introduction

No other sector has experienced such an explosive growth as video game industry [1]. It is analyzed that video game industry will have a growth of 20B US$ in the year 2014 and it is expected to cross 80B US$ in a couple of years [2].so this growth of game industries has lead the game developers in need of new ideas for developing games, this leads the game developers to develop games based upon adaptations of fictional novels. Some of its examples are. This includes extraction of characters, scenes and events from the novel. Instead of reading out the entire novel we describe a system in which novel is parsed to identify characters, scenes and events [3].

Some of the previous works are text to scene systems in which text is taken as the input and generates a scene depicting the text [4], extracting scenes and characters from short stories [5].The current work is based on extracting characters, scenes and events and providing statistical information based on the importance.

2. Architectural diagram

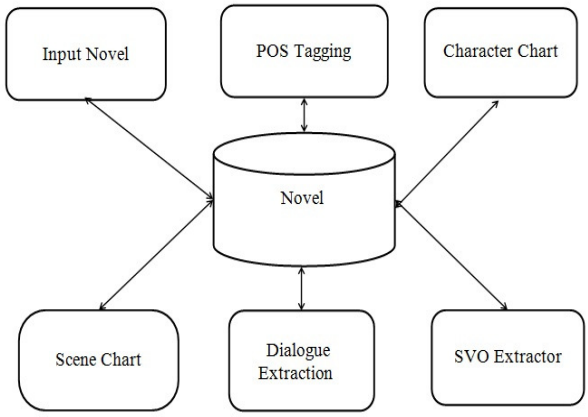

Fig 2 Architecture diagram

Figure 2 clearly depicts the design of the tool. In which the input novel is parsed and tokenized and those tokens are labeled with the help of penn tree-bank tag set. Character chart is used to depict the characters of the novel. On the other hand the objects/location are obtained from the scene chart. A brief summary of objects is obtained in scene summary. Dialogue importance can be found in the timeline chart. The SVO extractor is used to obtain the verb, subject and object of any particular sentence.

3 Modules

3.1 GUI Compilation

In this module, all the other modules are compiled and presented in a single interface. The initial GUI contains the login page where the user name and password is taken as the input. Once the enter button is clicked it opens another GUI. This GUI has the buttons for each module which shows the output of character chart, scene chart, scene summary, dialogue timeline chart and storyboard tool.

3.2 Extraction of Characters

Character Extraction is done to extract key characters present within the novel and shortlist them according to their importance in the story using the obvious measure of frequency of occurrence. Using this method a cast amount of data is highly reduced to provide within the necessary character names along with their importance.

3.3 Extraction based on Frequency

Term frequency-inverse document frequency is a numerical statistics which reflects how important a word is in a document. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document to the number of times a word appears in the document.

The frequency count is calculated by using the equation (1) and (2)

$$Tf - idf(t, d, D) = tf(t, d) \times idf(t, D)$$

$$tf(t, d) = \frac{tf(t,d)}{\max\{f(w,d) : w \in d\}} \quad (1)$$

$$Idf(t, d) = \frac{\log |D|}{|\{d \in D : t \in d\}|} \quad (2)$$

$t \in d$- the number of documents where the term t appears.

$|D|$ - total number of documents

$f(w, d)$- weightage of the document

Example:

Consider the word “Apple” that appears 10 times and the word “the” appears 20 times in a document which contains a total of 100 words .

$$|D| = 1$$

$$Tf - idf(apple, 1) = \frac{10}{100} \times \log\frac{1}{1}$$

$$Tf - idf(the, 1) = \frac{20}{100} \times \log\frac{1}{1}$$

3.4 Extraction of Scenes

The whole book is processed and important scene description objects are extracted . These objects consist of location, scene objects and necessary characteristics. The input for this is book in text format . The steps that are followed first to obtain objects are

1) Extract the whole book in POS tagger.
2) Look out for preposition in the whole input book with the help of POS taggers.
3) Prepositions are always connected to nouns.
4)Eliminate nouns other than common nouns.
5) This represents the objects/locations.

The above five steps are followed to extract the objects. Thus finally objects are obtained. Again over here using the frequency count algorithm the importance of objects is obtained.

3.5 Scene Summary

The input given here is the information extracted from scene module. The scene objects and locations obtained from the scene module is taken as the keyword and the paragraphs containing those words are extracted.

3.6 Dialogue Extraction

Now-a-days games are created with important dialogues. Interaction between characters is shown in games for clear understanding of games. Thereby for this information, dialogue is needed. Q clustering algorithm is used to get important dialogues from a novel.

1) Look out for open and close quotes in the document.
2)That particular sentence is extracted and is put into clusters.
3)Using the concept of Q clustering clusters are formed.
4)Timeline chart for dialogues is obtained.

3.7 Clustering Algorithm

1) Read the entire document clustering done from the first sentence in the list.
2) Sentences that are related to other sentences are put into an initial cluster.
3) Each cluster should contain only one topic. There must be at least two sentences in a cluster and more than two if possible.

4) Continue clustering and build up the clusters while going through the list of sentences.
5) Clusters can be re-arranged at any point.
6) When clustering is finished check whether that all important information from the document is covered by the clusters.
7) Clusters are ranked by importance

3.8 Subject Verb Object Extractor

An intermediate GUI that helps in storyboarding is created. This is used to find action verb, object and subject. This tool will be useful for the game developers and graphic artist to develop games. The obtained timeline chart is fed as an input into this GUI which extracts subject, action verb and object.

4. Conclusions

This paper proposes a method of extracting Key Scenes i.e. characters ,scenes, events from novels and providing a statistical information based on the importance. This is achieved by using the Monty-lingua tagger and Q Clustering technique. Advantage of using this method is it saves the time of the game developer instead of reading out the entire novel. Some improvements have to be done in order to increase the accuracy level. This method had been checked with these novels Pride and Prejudice, Adventures of Huckleberry Finn, Women in Love , Adventures of Tom Sawyer .This method failed to response for kids fairy Tale Book

References

[1]http://www.theesa.com/facts/pdfs/esa_ef_2013.pdf
[2]http://www.qj.net/xbox-360/news/analyst-us20b-growth-for-video-game-industry-in-2014-new-ps-in-development-already.html
[3] Richard Cant, Caroline S. Langensiepen,Purcell.J, ”Fictional Simulation, Automatic Extraction of Key Scenes From Novels”, Proceedings of the 13th International Conference on Modelling and Simulation,pp.99-103,ISBN:978-0-7695-4376-5,March 2011.
[4]Xiaojin Zhu, Andrew B. Goldberg, Mohamed Eldawy, Charles R. Dyer, andBradley Strock,’A text-to-picture synthesis system for augmenting communication’ In Proceedings of the 22nd national conference on Artificial Intelligence (AAAI'07) Press, 2007. ISBN 978-1-57735-323-2, pages 1590-1595.
[5]Chandra,k. Automatic Extraction of characters & scenes from Natural Language text. Master thesis 2009. Nottingham Trent University