INNOVATION-NEXT@ ARTINEX'15: A MOVIE GENRE IDENTIFICATION CHALLENGE

Saurabh Bhagvatula[1], K.J.Shreyas[2], Saurav Gupta[3]
Student, B.E.(7th Semester), Department of Computer Science & Engineering,
Bhilai Institute of Technology, Durg
Email:Saurabhbh21@gmail.com[1],shreyaskj10@gmail.com[2],sourav77gupta@gmail.com[3]

Abstract— The present communication laid by the above mentioned ArtInEx participants aims to report about the system incorporated towards the genre identification from the given plot summary of the movie. The noun-phrases and verbs generated from the plot summary are match against the existing verbs and noun-phrases available in genres' dictionary. Also, the care was taken for any keyword belonging to multiple genres and giving different weight age to every genre it belongs. For each keyword match in a genre dictionary, a weight is added toward movie belongs to that genre and genres' having its total weight above threshold are belonging to the genres. But, the proposed model fails to produce the desired result when it encounters the keyword which is not available in any of the genres' dictionary.

Index Terms— Genre Identification, Machine Learning, P.O.S. tagger, Plot Summary, Type Dependency parser.

I. INTRODUCTION

This piece of work is performed as part of classifying new movies to multiple genres based on synopsis or plot summary. The genres evolve periodically. New genres can get added or some old ones might get removed over a period of time. Such evolutions of genres make classifications slightly difficult since we don't have fixed classes and we need to reclassify all the old documents. The problem can also be considered as a topic modeling problem of identifying topics which can correspond to genres as opposed to a classification problem So, the infrastructure had to be developed in which the noun-phrases and verbs should be available within the 19 genres verb and noun dictionary, which are available over the time and classifying genres based on threshold aggregate value of evaluation genre matrix. For instance, any available movie plot summary, the noun-phrases and verbs are generated from P.O.S. tags and type dependencies of plot summary. Then, the Evaluation Genre Matrix is generated, which gives genre-wise weight aggregation of noun-phrase and verbs matched in the genre. Those genres having aggregate value above the threshold aggregate value are belonging genres..

II. THE DATASETS AND CLASSIFICATION CLASSES

The ArtInEx participants were asked to download and utilize any dataset which they might feel be needed to develop the working prototype of movie genres' identification based on the plot summary of the movie. For this purpose we have downloaded the dataset of 1600 records containing the name of movie, link containing its plot summary along with a vector

(of 19 characters as 0 or 1) which contains information of set of genres belonging to a particular movie from a set of genres as shown in figure1.

Looking at the dataset one can only view the plots of the movies and its corresponding genre vector delivering the information of belonging genres. In order to execute the task, the web links are stored in a separate file and this plot are scraped from the web using the links of the plot summary and stored in a file. Then genre vectors are stored in another file in same serial as the movie plot summary.

After this, the final two files i.e. the plot summary file and the genre vector file are combined to train the system for plot and genre fetched from the files.

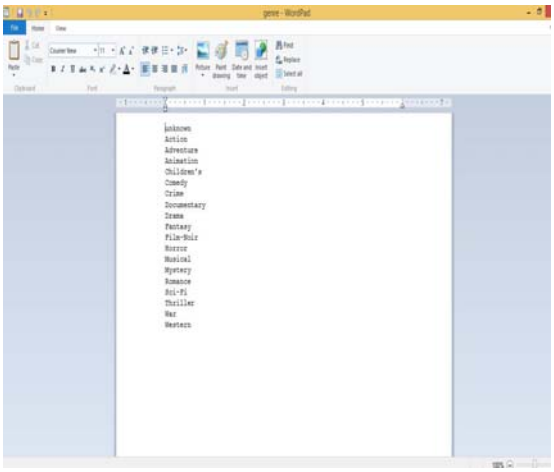

Figure 1: File having list of genres

III. THE PROPOSED APPROACH

The figure 2 shows the entire processing steps and also indicate the training and the testing phase flow through two different types of arrows. The training phase is described in this section and the testing phase in the next section.

A. USE OF PART-OF-SPEECH TAGS

The concept related with the aspect of providing the space to the machine to understand, the governing parameters are contributed by the extraction of Noun / verb phrases parsed from natural language syllabus text-strings. An efficient Part-Of-Speech Tagger is offered by Stanford NLP group,it is an open-source tool generated by the computational linguistic communities. The work of POS Tagger is to read English text and assign parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally applications use more fine-grained POS tags like 'noun-plural'. Similarly, in the current scenario, all possible Noun variants in the pool of Part-Of-Speech tags shall be used to extract the entity-specific noun phrases of the happenings as displayed in newswire.

A. EXPLORING THE TEXT SEMANTICS

Our task revolves around the main objective of finding the degree of semantic similarities between the mentioned news article and the referenced knowledge-base node. After deep survey it was found that the Stanford typed-dependencies which provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used to extract meaningful relations by the computational linguist-expertise. The most important feature, with which it excites the community, is that it represents all sentence relationships uniformly as typed dependency relations between pairs of words - simple, uniform representation that can be tailored to any application realm in Natural Language Processing tasks. However some degree of pre-processing is needed to correctly operate this selected NL Parser.

B. EXTRACTION OF VERBS AND NOUN-PHRASES

The Stanford typed-dependencies of the plot summaries are manipulated so that noun-phrases and verbs, may be existing in form of single-word nouns or n-word compound noun phrases in shape of n-gram patterns and the verbs, are generated in series, revealing semantically related chains of nouns and the verbs that dominate the major event vocabulary happening in that plot summary. For accomplishing the above, at one end, Part-Of-Speech tagged of the plot is fetched and at the other end, the seed-words that exist as query-names along with the alias-names (if any) trigger the generation of these noun-phrase chains and verbs.

.

A. GENERATION OF GENRE DICTIONARY AND WEIGHTED GENRE DICTIONARY

The noun-phrases and verbs are generated are the basis for delivering the meaning any summary of context holds. The nouns and verbs in the plot deliver the genres a particular plot summary belongs to. In order to create a genre vocabulary (both verb and noun), the verbs and nouns of the training plots are stored in the belonging genres' verb and noun dictionary respectively. Also, there is possibility that a particular word/ phrase are stored in multiple genre dictionary but have different degree of contribution to each dictionary. So, a genre weighted dictionary is created, when all training plots are used and the genre dictionary is created. For instance the noun-phrases of action genre are stored in a file named as action_noun is shown in figure-3. Similarly, the weighted genre matrix for the Action (noun) having filename action_nounwt is shown in figure-4.

B. EVALUATION GENRE MATRIX AND THRESHOLD AGGREGATE VALUE

The evaluation genre matrix is the 19*3 matrix, where 19 rows represent the available set of genres (19 here). The three columns are for noun, verb and aggregate value of genre for a particular movie. The noun, verb columns represent the total weighted match of verbs and noun in the plot with that of weighted genre matrix. Since, the verbs are more contributing towards the genre identification as verbs represents actions and other relevant activities. On the other hand, the nouns are less contributed for the genres as are more viable to change like persons' name, place, etc. So, for calculating the verb value are given more weight age and aggregate value is calculated. For instance table1, shows the evaluation genre matrix where Xi, is the total weighted match of nouns, Yi is that of verbs and the aggregate value is Zi =(0.8*Xi+1.2*Yi)/2.

In order to classify the genres of the plot summary from evaluation matrix, we have to develop the threshold value which is used for genre identification. For this purpose the evaluation genre matrix having the belonging genres', its aggregate value as in table1 is used and minimum of which is calculated and stored in a file and minimum of all values stored correspond to training plot summaries(as shown in figure 5) is the required threshold values of Evaluation Genre Matrix(aggregate value) for being classified to a particular class(genre

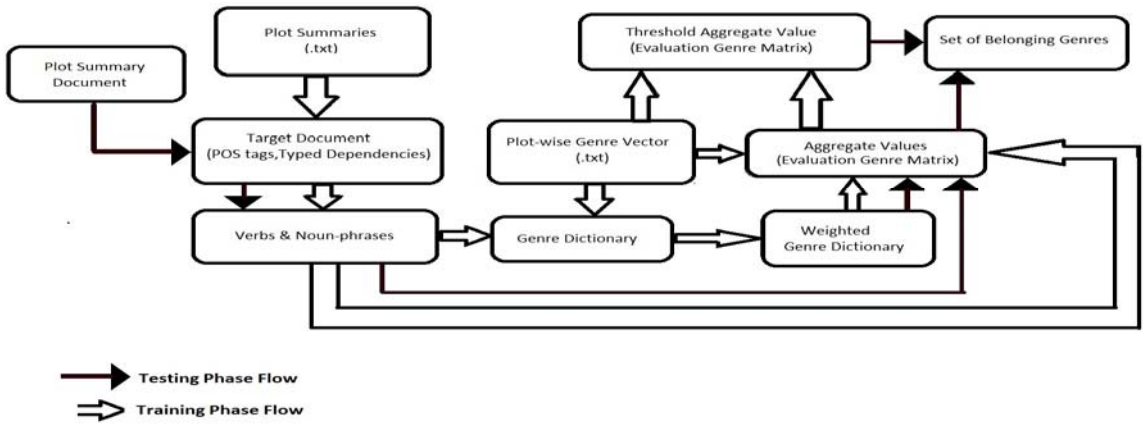

Figure 2: Process Flow Diagram

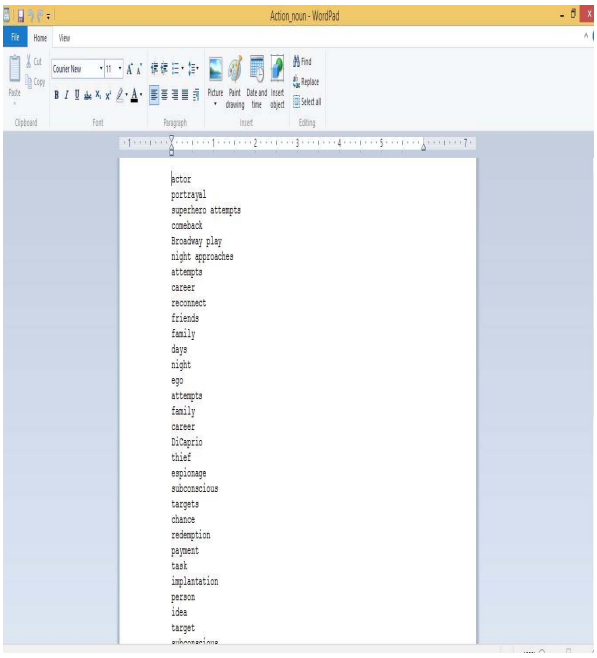

Figure 3: Keywords, which are stored in noun dictionary of Action genre

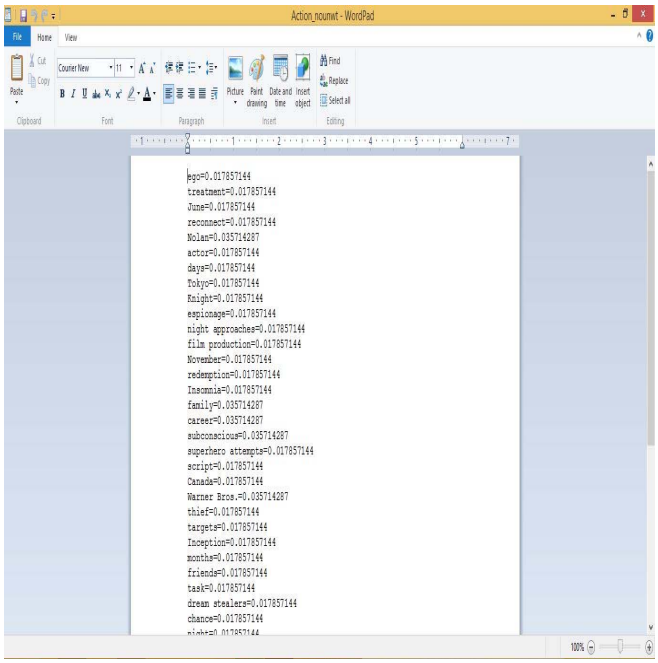

Figure 4: Every unique keyword with its weight of Action(noun)

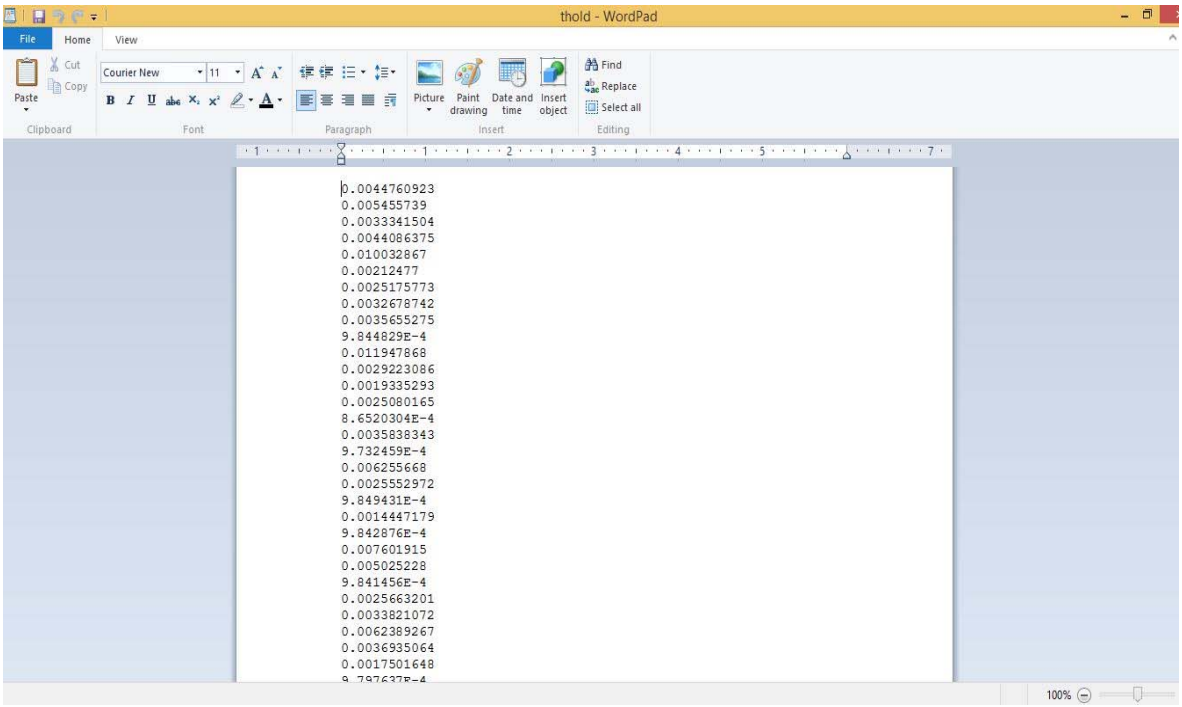

Figure 5: Threshold value of Evaluation Genre Matrix(aggregate value) of training plots stored in file

Table 1: Evaluation Genre Matrix Table

Name of the Genre	Noun-weight (=Xi)	Verb-weight (=Yi)	Aggregate- Value Zi=(0.8*Xi+1.2*Yi)/2
Unknown	X1	Y1	Z1
Action	X2	Y2	Z2
Adventure	X3	Y3	Z3
Animation	X4	Y4	Z4
Children's	X5	Y5	Z5
Comedy	X6	Y6	Z6
Crime	X7	Y7	Z7
Documentary	X8	Y8	Z8
Drama	X9	Y9	Z9
Fantasy	X10	Y10	Z10
Film-Noir	X11	Y11	Z11
Horror	X12	Y12	Z12
Musical	X13	Y13	Z13
Mystery	X14	Y14	Z14
Romance	X15	Y15	Z15
Sci-Fi	X16	Y16	Z16
Thriller	X17	Y17	Z17
War	X18	Y18	Z18
Western	X19	Y19	Z19

IV. GENRES' PREDICTION

In order to predict the genres from its plot summary with the help of the predictive model described above, a set of steps are followed utilizing some of the modules which are described in the process approach

A. GENERATION OF VERBS AND NOUN-PHRASES

The plot summary input for prediction of set of genre, at first its P.O.S. tags and type dependencies are created and are utilized for further process. From the generated P.O.S. tags and semantic dependencies available with us, we can generate the noun-phrases and verbs from this, which represents the major keywords of the plot and the genre dictionary having its vocabularies, are matched with for existence of genres.

.

B. GENERATION OF EVALUATION GENRE MATRIX AND GENRE PREDICTION

The evaluation genre matrix is derived from the nouns and verbs from the plot of the movie given to predict the genre. The weighted genre matrix of nouns and verbs are utilized for this purpose. The aggregate value of each genre in the matrix are compared with the threshold aggregate value, those having value above the threshold value are the genres', to which the given movie plot are classified. From the figure-6, the genres having aggregate value of evaluation genre matrix above threshold are stored in a file.

```
if(eval_mat.get(ks)[2]>min)
{
 pw1.println(ks);

}
```

Figure 6: Snippet of code checking genre belongs to movie or not

V. RESULT

The generated model is tested for 50 plot summaries with 175 classification instances. After the model was tested for these instances and analyzing the output generated from the system, the system accuracy is mentioned below in table2.

Table 2: Classification result on testing dataset

Classification Category	Classified Instances (in number)	Classified Instances (in percentage)
Correct Classification	53	30.2
Incorrect Classification	24	13.7
Missed Classification	122	69.7

VI. CONCLUSION

The developed genre identification system is first of its prototype. The design series encourages the research group to work only upon increasing the robustness of module design. In doing so successful match for unknown keywords are obtained resulting in 90% accuracy ,as it was observed in table 2 that missed classification was 69.7% due to the reason that genre dictionary doesn't have the sufficient keywords. The group is focusing on to develop system model which can improve accuracy without including external dictionary.

VII. ACKNOWLEDGEMENT

The authors are thankful to Dr. A. Rawal, Professor, Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg and Dr. M.V. Padmavati, Head, Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg for their continuous encouragement guide and support towards the research

REFERENCES

[1] Ani Thomas, Arpana Rawal, M K Kowar, Sanjay Sharma, Sarang Pitale and NeerajKharya, "**Bhilai Institute of Technology Durg at TAC 2010: Knowledge Base Population Task Challenge**", Proceedings of Text Analysis Conference, 2010, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.

[2] Arpana Rawal, Ani Thomas, M K Kowar and Sanjay Sharma, "**ARPANI@BIT_DURG : KBP English Slot-filling Task Challenge**", Proceedings of Text Analysis Conference, 2013, organized by Information Technology Laboratory, National Institute of Standards and Technology, USA.

[3] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning on "**Generating Typed Dependency Parses from Phrase Structure Parses (2006)**", PROC. INT'L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC),444-23 self.

[4] Marie-Catherine,de Marneffe and Christopher D. Manning, "**Stanford typed dependencies manual**" , IEEE/ACM International Conference on Automated Software Engineering (ASE 2015), Lincoln, Nebraska, USA; 11/2015.

[5] Kristina Toutanova and Christopher D. Manning,"**Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger (2000)**", EMNLP/VLC 2000, 142-4 self.

[6] Kristina Toutanova, Dan Klein, Christopher D. Manning and Yoram Singer, "**Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network**", Computational Linguistics on Human Language Technology - Volume 1, Pages 173-180,May-2003