TEXT BASED SPEAKER IDENTIFICATION

[1]Mohit Agarwal, [2]Mayank Chaudhary, [3]Purwa Maheshwari, [4]Kaushal Kishor

[1,2,3,4] Department of Information Technology, ABES Institute of Technology,
Vijay Nagar, Ghaziabad,
Uttar Pradesh, India

***Abstract* : This paper presents an approach to identify the speaker based on non-acoustic content. However, the present article reports for text-based speaker identification task from dialogues obtained from the TV show called "FRIENDS" which includes only textual features only. In our research, we only focused on what the speaker speak not how the speaker speak. We have used a text corpus of TV show Friends. As we know different characters exhibit different linguistic styles while speaking. We have integrate the stylistic features that are extracted from the turns of various characters. We are using part-of-speech (POS) that reads text in some language and assigns parts of speech to each word. The k-nearest neighbor (k-NN) have been employed for identifying the speakers from the dialogue scripts at turn levels.**

Keywords- corpus, non-acoustic, stylistic features, part-of-speech

1. INTRODUCTION

Speaker recognition termed as automatically recognizing who is speaking on the basis of individual information in a set of known speakers. The process of speaker identification is the best example of pattern recognition which is a branch of machine learning. For identifying the speakers training phase is required. The system is provided with the dataset that belong to one of the speakers that have been learned by the system. The process of text based speaker identification is done for known set of speakers. It has been observed from earlier studies that mostly the speaker identification research using only speech signals. As our knowledge not much research is done on speaker identification which is only based on textual features. In our research on speaker identification we are using the dialogues of TV show, FRIENDS. In this research our aim is focussed only on what the speaker speak not how the speaker speak.

Speaker Recognition, Speaker Identification and Speaker Verification are looking similar but all the terms have some difference between them. Speaker recognition can be divided into Speaker Identification and Speaker Verification. Speaker identification determines which registered speaker provides a given utterance from amongst a set of known speakers. The research paper that are related to this topic are using the movie scripts which are the good resources for the speaker identification research. In this different datasets which belong to the different speakers are provided to the system which helps in system training. The required output of the system is to identify one of the speakers from trained speakers, or rejection if the test data is not belong to any trained speakers. The main challenge that we face in the fact that no verbal information is present in the utterance. To identify the speaker we are using only the textual features which is quite difficult task. As we know different speakers have different linguistic styles of speaking, so we associate these linguistic features of speakers at turn level and used them in various supervised techniques of learning.

2. RELATED WORK

We observed that different researchers used different approaches for speaker identification. Reynolds and Rose (1994) introduced the Gaussian Mixture Models (GMM) for robust text independent speaker identification. GMM approach is applied on different datasets and provide better result. Amitava Kundu and Dipankar Das (2012) prefer to use the KNN Algorithm, Naive Bayes Classifier and Conditional Random Field (CRF) to organize speakers in the film dialogues based on linguistic stylistic features. By using these approaches and achieved good result. Kaylin Ma, Catherine Xiao and Jingo D. Choi (2017) proposed a Convolutional Neural-Network model for text based speaker identification on multiparty dialogues retrieved from the TV show, Friends. With the help of CNN model they achieved the great result.

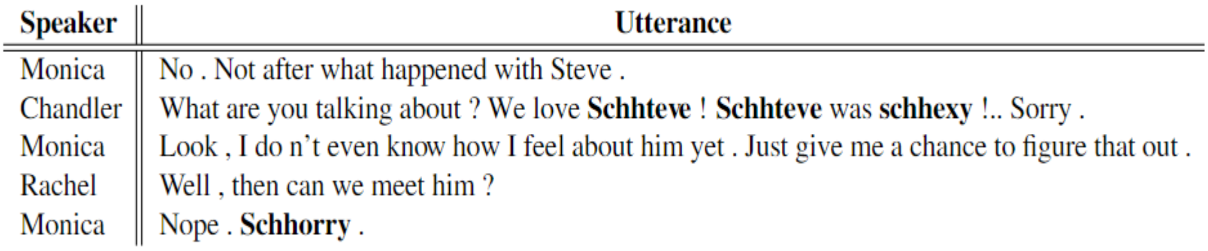
3. CORPUS

In this project we are using transcripts of famous TV Show "Friends" available in JSON format for all 10 seasons by The Character Mining Project (reference for link of github). Table 1 represent some dialogues of characters present in the script. Transcript is made available separately for each season, and contains many features like Episodes, Scenes, Utterances, Speakers, etc.

Each season contains a number of episodes, and each episode is comprised of separate scenes. The scenes in an episode, in turn, are divided at the utterance level.

We are using first 8 seasons as training/learning data, and rest of two seasons for testing purpose. JSON format seasons are saved in POJO using Jackson API and henceforth used in object form for training/identification/grouping/classification purposes.

As per analysis, there are 6 major characters with approx. 11-15% dialogues of every individual hence we are taking only 7 characters (All others are regarded as others) here to maintain the balance of utterances for all. In total, this corpus consists of 194 episodes, 2,579 scenes and 49,755 utterances. The percentages for major speakers are fairly consistent. However, the other speaker has a larger percentage in the dataset than any of the major speakers.



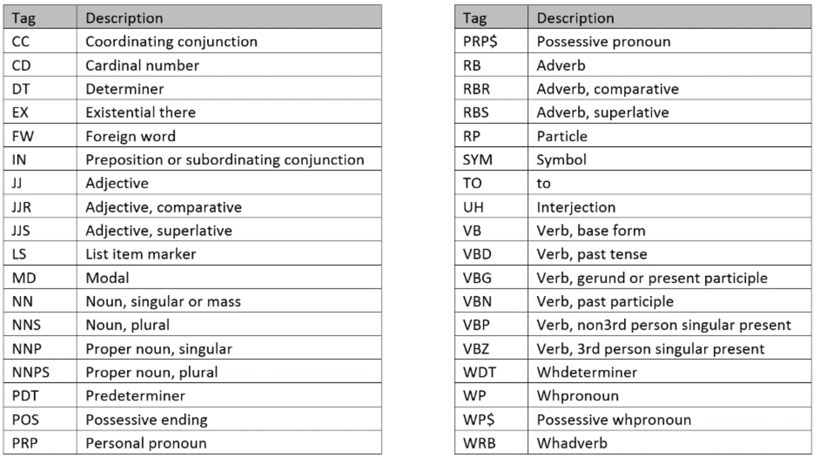
Speaker	Utterance
Monica	No . Not after what happened with Steve .
Chandler	What are you talking about ? We love **Schhteve** ! **Schhteve** was **schhexy** !.. Sorry .
Monica	Look , I do n't even know how I feel about him yet . Just give me a chance to figure that out .
Rachel	Well , then can we meet him ?
Monica	Nope . **Schhorry** .

Table 1: Dialogues from transcripts to the TV show Friends.

4. FEATURE ANALYSIS

As our best knowledge the stylistic features have been used in the previous research for speaker identification of textual data based on the premise that different speakers have different styles of writing. In our project, we have attempt to identify the speakers on the basis of their speaking styles. We are using the POS tagger which can read each word of the dialogue of each speaker and assign a part of speech to every word. In our research, we are using the "Stanford part-of-speech" tagger (http://nlp.stanford.edu/software/tagger.shtml).

In figure 2 we can provide the Penn Treebank tagset in which number of stylistic features are present which are used in our project.



Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Fig.1 Penn Treebank tagset

5. TECHNIQUES

5.1 k-nearest neighbour (k-NN) approach

In Amitava Kundu and Dipankar Das (2012), the good result is achieved by implementing the K Nearest Neighbor algorithm (KNN). In our research we are also implement K-NN algorithm and selected as basis approach for this paper. In our project, each turn is presented as a point in the vector space. Around every point in vector space many neighbours are present, to find the nearest neighbours of that point we require the k-NN approach. K-Nearest Neighbor is one of the most basic yet essential classification algorithm. K-Nearest Neighbor is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM). As according to corpus, each utterance is served as one sample, and 50 distinct linguistic features are extracted from elicit from each sample. The locality has been calculated with the help of a 'similarity' metric, the cosine similarity measure. As cosine similarity is used to discover the 50 nearest neighbours to each sample. The advantage of cosine similarity is that it is used to measure the text similarity. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbors. Fig. 2 represent the how k-NN approach work.

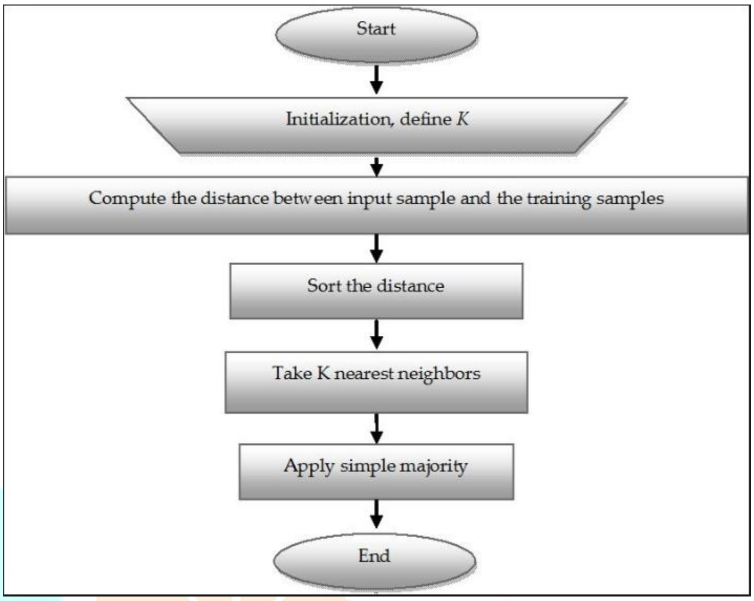


Fig.2 Work flow of k-NN approach.

6. CONCLUSION

In the present paper a K-NN approach is implemented to identify the speaker in script dialogues depend only on textual data. As Amitava Kundu, Dipankar Das (2012) applied different classification approaches and uses 8 stylistic features for speaker identification by using text scripts only. The features used in the experiment provide better result. In our experiment, we are using more number of stylistic features and applied K-NN approach on textual script which provide us more accurate results. Although speaker identification and verification is more based on acoustic features which is a helpful tool. Speaker identification task become easy by using acoustic features as using textual features only. So using only textual features for speaker identification become the more interesting area of research. As other classifiers like Support Vector Machine (SVM) is also useful in text-based speaker identification. Apart of future work we are devising to implement other classification techniques with different numbers of stylistic features and compared the results.

7. REFERENCES

[1] Amitava Kundu, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Speaker identification from film dialogues. IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction.

[2] Kaixin Ma, Catherine Xiao and Jinho D. Choi (2017).Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks.

[3] Elisabeth Lex, Andreas Juffinger, Michael Granitzer. (2010) .A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs. 2010 Workshops on Database and Expert Systems Applications.

[4] D.A. Reynolds and R.C. Rose. 1994. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing.

[5]Iulian V. Serban and Joelle Pineau. 2015. Text-Based Speaker Identification For Multi-Participant Open Domain Dialogue System. Department of Computer Science and Operations Research, Universite´de Montreal ´.

[6] http://nlp.mathcs.emory.edu/ character-mining