MINISTERUL EDUCAȚIEI NAȚIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREȘ

Acta Marisiensis. Seria Technologica

Vol. 20 (XXXIV) no. 2, 2023

sciendo

10.2478/amset-2023-0017

TOPIC IDENTIFICATION IN VOICE RECORDINGS

Zsuzsa SIMO[1]

[1] *Doctoral School, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, Romania, zsuzsa.simo@umfst.ro*

Abstract

The paper shows the understanding of a topic recognition problem like the speech recognition system based on Natural Language Processing (NLP) and the steps of its implementation of a rules-based approach, which is able to classify given audio materials based on predefined topics in real-time. During implementation, a statistical vocabulary was developed. Google Speech API (Application Programming Interface) *was employed for subtitling audio materials, and the most ideal time frame for reception was identified through several experiments. The motivation of this work is based on the deficiency of similar simple systems for Hungarian topic recognition, even though numerous international languages already utilize multiple Automatic Sound Recognition (ASR) systems.*

Key words: *speech recognition, statistics for performance analysis, natural language processing, machine learning, prediction in healthcare*

1. Introduction

With the advantage of topic definition in real time communication becomes more efficient. Regardless of when the conversation is taking place, tracking the current topic and its keywords in advance allows us to gain comprehensive content based on previous thoughts. In addition, different language patterns and peculiarities can be easily identified by analyzing spoken conversations [1,2].

During the lexical interpretation of the words, the goal is to extract those words with a single meaning that occur in most parts of speech [3]. In order to filter out unwanted features, several 'stop words' must be defined, as these do not contribute significantly to interpretation, while their frequency is also considerable, which would increase computing resources. In addition, it is important to lemmatize the words and convert them into dictionary form, because it is easier to recognize them later based on any inflected alternative [4].

The study is structured as follows: Section 2 presents a word-based approach technique for speech recognition, while Section 3 presents the architecture

and parameters of the proposed system. Finally, Section 4 summarizes the results of the research.

2. Word-based approach

This paragraph presents the methods applied to prepare the vocabulary, the open-source subtitling software, the extracted words, and the processing of the content related audio materials.

The Google Speech API (open-source speech-to-text software) was utilized for subtitling the audio materials, and thanks to the parallelism of its operation, receiving and subtitling could be done simultaneously [5]. To identify the correction of the Speech API, the average correctly recognized words were calculated [6] on the worst possible scenario, where the environment sounds were most of the time noisy and fast, where a recognition rate of 59.94% was gained.

When creating the dataset, it was important to consider the difficulty and characteristics of the problem to be solved. The problem is challenging because a few numbers of text can exist multiple topics [7].

To extract words defining groups, it was first necessary to collect a suitable volume. For this, audio channels were recorded, since the samples must be provided from an environment that the system will encounter later. After finding contents related to a specific topic, a time stamp had to be set for continuous audio recording. Since speech is not always clearly audible during recording, the microphone had to be constantly tuned to the ambient noise level.

In the case of a successful audio recording, the appropriate audio material was labeled and stored the result, otherwise down error messages were documented, to draw later generalizations about the accuracy of the audio recording, while it represented a valuable starting point in the basis of system parameterization.

From the successfully stored text, many of the most frequently occurring words were extracted, which can be related to the given topic. For this, it was essential to create a blacklist dictionary, most of which consists of sentence structure elements used in grammar. It was necessary to constantly expand this list to eliminate the words that least defined a group.

3. System architecture and methods

This paragraph presents the unified system architecture and the dataflow. It also presents the one-hot encoding and the rule based-system as well.

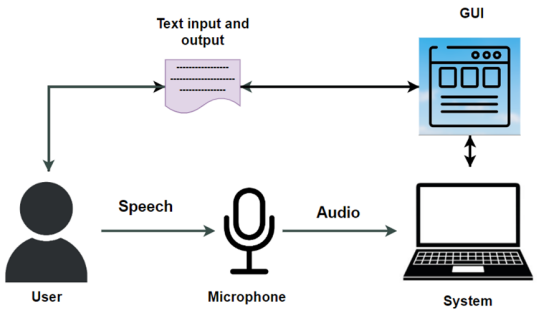

Fig. 1: Unified system architecture

Figure 1 illustrates the combination of hardware and software components. First, the microphone converts the user's voice into sound waves. The Speech API then processes the sound waves and converts them to text. The generated text is utilized to predict from the given words, and the output of its processing is displayed on the user interface when it is entered into the system. The user can see the result of this while he can also write text data on the screen. In this way, you can expand and list the current topics in the system.

To define a topic, multiple sentences and expressions were examined. The recording was prepared to handle various and sufficient audio as well. The text data was stored using one-hot encoding as a binary feature, where each word corresponded to a unique index in a binary vector, implementing a rule-based system [8] with specific combinations of words that indicate a certain topic. These rules are combined with the keywords and phrases of the categories, logical conditions and the tracking of the category's assignments.

The words obtained during continuous reception were stored after timestamping in a prediction buffer, which constantly changes its content, if it does not contain enough words, it must be expanded with empty elements and corresponding timestamps for the prediction to be realized. The system receives these words and their possible affix-free versions as input [9].

Since it was not clear in what size the audio recording frame of the speech should be analyzed,

measurements were performed to determine the ideal frame. The results of the measurements showed which time interval was the most suitable for the system, while it appeared that the other measurements provided worse results.

Another similar experiment was related to measuring the number of words that were used in the real-time prediction, where the applied function calculated the position of the highest and the arithmetic average.

4. Results

This paragraph shows the results of significant measurements with explanations like the results of sampling in diverse seconds, the percentage occurrences across categories and the occurrences of aggregated topics in previous time.

Figure 2 presents 210 second measurement with four-second samplings, which shows a weak measurement because of the low word number in the vocabulary.

Figure 3 shows a 210-second measurement with 5 seconds of sampling, which presented a better result in time selection than Figure 2. In this time interval, it was possible to identify the primary theme (*Gastronomy*) of the audio material ten times. Compared with Figure 2, the *Religion and Spirituality* category also appeared in the same time intervals. However, the *Advertisements* category was also detected, the presence of which can be attributed to the fact that the word found was not sufficiently specific to the topic.

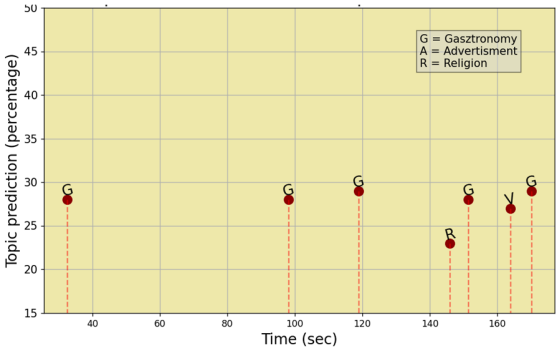

Fig. 2: Result of sampling every 4 seconds

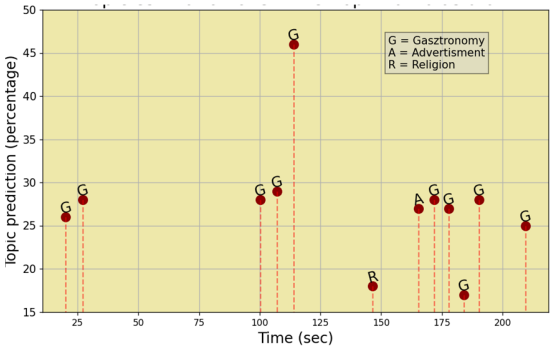

Fig. 3: Result of sampling every 5 seconds

Figure 4 shows the same 210-second measurement but with considering the possible suffixes and prefixes of the detected words. It was thirteen times possible to determine the main topic, and three times other possible topics, which is 30% more accurate than the result of Figure 3.

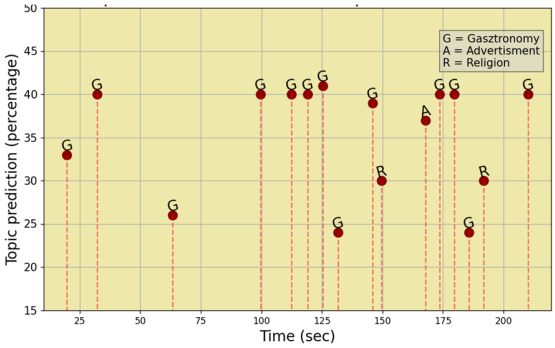

Fig. 4: The result of sampling every 5 seconds, considering the affixes of words.

In addition, a 10-times 600-second measurement was performed, and the results of True Positive and True Negative values were also considered.

The stacked bar chart in Figure 5 shows the percentage ratio of the categories per measurement compared to the whole according to the 600-second period, with the help of which the composition of the individual measurements can be seen. It is visible the overlaps between the categories, the successful topic identifications, and the extent to which some topics can appear in other topics.

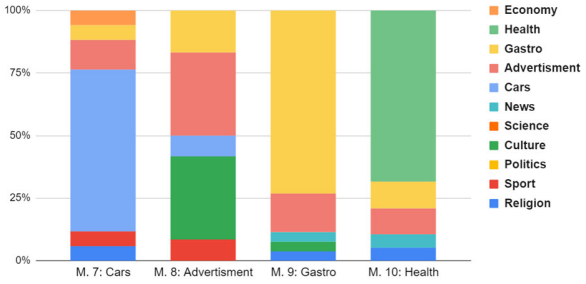

Fig. 5: Breakdown of percentage occurrences across categories

During this measurement, the root causes of the unexpected measurement were classified, like the presence of the False Positive values. As an example, if we consider *Measure 10:Health* the main topic was identified 13 times, the presence of the correct sub-topics are 2, and the number of incorrect sub-topics are 5, which leads to 75% precision.

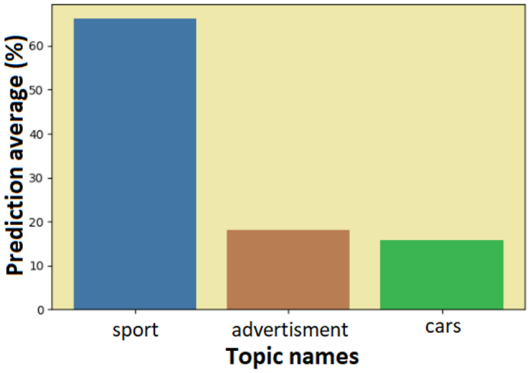

Fig. 6: Occurrences of aggregated topics in the previous one minute

Figure 6 shows what the constantly updated diagram looks like during a 60-second aggregate measurement, during which inactive topics were filtered out after 30 seconds and increased the percentage occurrence of topics that occurred more than once by normalizing. Figure 7 presents a 300-second aggregate measurement, during which the inactive topics after 150 seconds were filtered out.

The advantage of this is the fact that within a short period of time, the topics that are not present will be present in such a small percentage that they can be ignored in further summary.

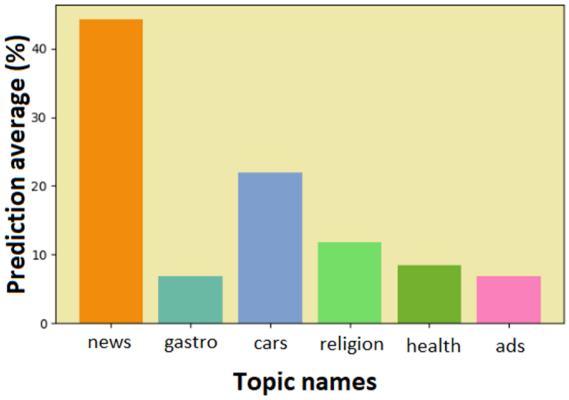

Fig. 7: Occurrences of aggregated topics in the previous five minute

When comparing the proposed system with an available topic modeling-based system, with twelve measurements each in 60 seconds, the same text was provided which gained a result of 69% accuracy for the proposed system and 87% precision for the NLP system, which presents that the proposed system with the mentioned development, can compete.

5. Conclusion

After completing the measurement analysis, this paragraph outlines our conclusions.

From the summary of the system's responses in a real-world setting, a comprehensive summary of its accuracy was gained. This way, it becomes evident that continuous development of the system is necessary while focusing on improvements that can effectively filter out most of the inappropriate cases. It should be noted that the categories "*Advertisements*" and *"News and current events"* presented numerous inaccuracies and weaker measurements. This is due to their broad coverage of diverse topics.

During the measurements, known sample sound recordings were chosen, and the categories were also predetermined. This allowed checking and summarizing the results by time segment. In addition, a solution to the problem was proposed during which the category of incoming audio materials can be determined for a longer time interval by continuously updating the time-stamped topics. To illustrate the system, a user interface was created to facilitate the display and expansion of existing samples.

Implementing a rule-based NLP approach helped to identify the main themes and topics of the subtitled text. Consideration of Hungarian language peculiarities was developed to collect base ideas that can be used in the future work of this project.

Frequently the experimental evaluation results are not in-depth analyzed by treating aspects like the variability and the apparition of outliers (statistical extremes). In such cases could appear even misinterpretations of the experimental evaluations.

A possible research direction can be to extend the experimental evaluations and apply some advanced analysis of the experimental results based on the methods presented in the scientific literature, like: [10, 11] comparison of two samples of experimental evaluation results; [12, 13] comparison of any number of samples of experimental evaluation results ([12] based on measurements made in pairs, [13] without measurements made in pairs); [14] identification of statistical extremes in the experimental evaluations that could alter significantly the experimental evaluation results.

Another research direction can consist in the application in healthcare. Paper [15] presents a multilingual healthcare chatbot using concepts of NLP with different Machine Learning techniques for disease prediction. Another similar idea is presented in work [16], where a chat and voice bot is implemented for prediction tasks in heart and ENT (ear, nose, and throat) diseases. Their proposed system collects data and makes diagnosis remotely with NLP. Paper [17] presents an NLP-based early prediction of medical specialties using structured data with unstructured textual notes recorded at the triage stage.

Acknowledgment

We would like to thank the Research Center on Artificial Intelligence, Data Science, and Smart Engineering (ARTEMIS), from George Emil Palade University of Medicine, Pharmacy, Science and Technology of Târgu Mureș, Romania for the research infrastructure support.

References

[1] G. Saon and J. -T. Chien, "*Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances*," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 18-33, Nov. 2012, doi: 10.1109/MSP.2012.2197156

[2] J. M. Baker et al., "*Developments and directions in speech recognition and understanding, Part 1 [DSP Education]*," in IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 75-80, May 2009, doi: 10.1109/MSP.2009.932166

[3] Z. Huang, P. Li, J. Xu, P. Zhang and Y. Yan, "Context-dependent Label Smoothing Regularization for Attention-based End-to-End Code-Switching Speech Recognition," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, 2021, pp. 1-5, doi: 10.1109/ISCSLP49672.2021.9362080.

[4] Khurana, A. Koli, K. Khatter, and S. Singh, "*Natural language processing: state of the art, current trends and challenges*", Multimedia Tools and Applications, vol. 82, pp. 3713–3744, 2022

[5] P. Maergner, A. Waibel and I. Lane, "*Unsupervised vocabulary selection for real-time speech recognition of lectures*," 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 4417-4420, doi: 10.1109/ICASSP.2012.6288899

[6] L. A. Kumar, D. K. Renuka, S. L. Rose, M. C. Shunmuga priya, and I. M. Wartana, "*Deep learning based assistive technology on audio visual speech recognition for hearing impaired*," International Journal of Cognitive Computing in Engineering, vol. 3, pp. 24-30, 2022

[7] H. C. Ozmutlu and F. Çavdur, *"Application of automatic topic identification on Excite Web search engine data logs*," Information Processing & Management, vol. 41, no. 5, pp. 1243-1262, 2005, doi: 10.1016/j.ipm.2004.04.018

[8] P. R. Nivedha and V. P. Sumathi, "A Survey on Text Mining Tools and Techniques support early testcase prediction," *2021 Int. Conf. on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICAECA52838.2021.9675761.

[9] D. A. McFarland, D. Ramage, J. Chuang, J. Heer, C. D. Manning, and D. Jurafsky, "*Differentiating language usage through topic models*," Poetics, vol.

41, no. 6, pp. 607-625, 2013. doi: 10.1016/j.poetic.2013.06.004

[10] L. B. Iantovics, C. Rotar, and M. A. Niazi, "*MetrIntPair-A Novel Accurate Metric for the Comparison of Two Cooperative Multiagent Systems Intelligence Based on Paired Intelligence Measurements,*" Int. Jou. of Intelligent Systems, vol. 33, no. 3, pp. 463-486, 2018. doi:10.1002/int.21903

[11] L. B. Iantovics, L. Kovacs, and C. Rotar, *"MeasApplInt - a Novel Intelligence Metric for Choosing the Computing Systems Able to Solve Real-life Problems with a High Intelligence,"* Applied Intelligence, vol. 49, pp. 3491–3511, 2019. doi:10.1007/s10489-019-01440-5

[12] L. B. Iantovics, *"Black-Box-Based Mathematical Modelling of Machine Intelligence Measuring,*" Mathematics, vol. 9, no. 6, p. 681, 2021.doi: 10.3390/math9060681

[13] L. B. Iantovics, M. Dehmer, and F. Emmert-Streib, *"MetrIntSimil-An Accurate and Robust Metric for Comparison of Similarity in Intelligence of Any Number of Cooperative Multiagent Systems,*" Symmetry, vol. 10, no. 2, p. 48, 2018. doi:10.3390/sym10020048

[14] L. B. Iantovics, R. Kountchev, and G. C. Crișan, *"ExtrIntDetect-A New Universal Method for the Identification of Intelligent Cooperative Multiagent Systems with Extreme Intelligence,*" Symmetry, vol. 11, no. 9, p. 1123, 2019. doi: 10.3390/sym11091123

[15] S. Badlani, T. Aditya, M. Dave and S. Chaudhari, "Multilingual Healthcare Chatbot Using Machine Learning," *2021 2nd Int. Conf. for Emerging Technology (INCET)*, Belagavi, India, 2021, pp. 1-6, doi: 10.1109/INCET51464.2021.9456304

[16] B. Dinesh, P. Chilukuri, G. P. Sree, K. Venkatesh, M. Delli and K. R. Nandish, "Chat and Voice Bot Implementation for Cardio and ENT Queries Using NLP," *2023 Int. Conf. on Innovative Data Communication Technologies and Application (ICIDCA)*, Uttarakhand, India, 2023, pp. 124-130, doi: 10.1109/ICIDCA56705.2023.10099942

[17] É. Arnaud, M. Elbattah, M. Gignon and G. Dequen, "NLP-Based Prediction of Medical Specialties at Hospital Admission Using Triage Notes," *2021 IEEE 9th Int. Conf. on Healthcare Informatics (ICHI)*, Victoria, BC, Canada, 2021, pp. 548-553, doi: 10.1109/ICHI52183.2021.00103.