International Journal of Intelligent Information Systems
2021; 10(4): 74-80
http://www.sciencepublishinggroup.com/j/ijiis
doi: 10.11648/j.ijiis.20211004.16
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Extracting Structured Data from Text in Natural Language

Zheni Mincheva, Nikola Vasilev, Ventsislav Nikolov, Anatoliy Antonov

Eurorisk Systems Ltd, Varna, Bulgaria

Email address:
jmincheva@eurorisksystems.com (Z. Mincheva), nvasilev@eurorisksystems.com (N. Vasilev), vnikolov@eurorisksystems.com (V. Nikolov), antonov@eurorisksystems.com (A. Antonov)

To cite this article:
Zheni Mincheva, Nikola Vasilev, Ventsislav Nikolov, Anatoliy Antonov. Extracting Structured Data from Text in Natural Language. *International Journal of Intelligent Information Systems*. Vol. 10, No. 4, 2021, pp. 74-80. doi: 10.11648/j.ijiis.20211004.16

Received: August 6, 2021; **Accepted**: August 20, 2021; **Published**: August 31, 2021

Abstract: Nowadays, the amount of information in the web is tremendous. Big part of it is presented as articles, descriptions, posts and comments i.e. free text in natural language and it is really hard to make use of it while it is in this format. Whereas, in the structured form it could be used for a lot of purposes. So, the main idea that this paper proposes is an approach for extracting data which is given as a free text in natural language into a structured data for example table. The structured information is easy to search and analyze. The structured data is quantitative, while the unstructured data is qualitative. Overall such tool that enables conversion of a text into a structured data will not only provide automatic mechanism for data extraction but will also save a lot of resources for processing and storing of the extracted data. The data extraction from text will also provide automation of the process of extracting useful insights from data that is usually processed by people. The efficiency of the process as well as its accuracy will increase and the probability of human error will be minimized. The amount of the processed data will no longer be limited by the human resources.

Keywords: Data Extraction, Structured Data, Unstructured Data, Automation, NLP, RASA

1. Introduction

The amount of data grows exponentially. For the year 2020 the data amount is up to a 64.2 zettabytes [1]. Figure 1 shows the data from 2010 with projection for the next 4 years in zettabytes.

Statistics shows that each person generates 1.7 megabytes of data in just a second. 80 - 90% of the generated data is in unstructured format [3, 6]. Unstructured data could be any information in text form [5], emails, social media data, mobile data as text messages and locations, any MS office documents and other [4]. While the data is unstructured it cannot be used for many proses. Since most of the produced data is in unstructured format many opportunities of the usage of the data are missed. This paper proposes an approach for conversion of unstructured data to a structured data. This extraction will enable the usage of the data for analyzing purposes and training of artificial intelligence networks.

There are several approaches for context-aware systems [12]. The suggested approach is to use and offers the opportunity to customize the key information which will be collected and transformed into records. The customization of the system implies extracting the only information, which is applicable for the specified context.

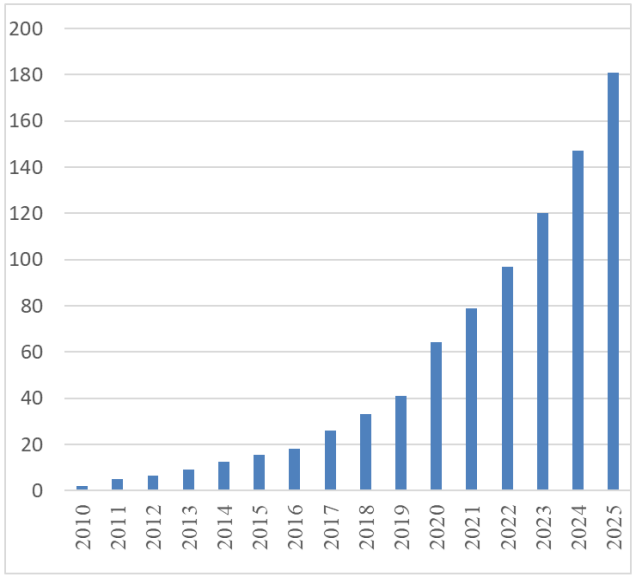

Figure 1. *Data amount projection from 2010.*

2. Related Work

There are several surveys on both data generation and extraction of meaningful information from free text. Synthetic generated datasets have big importance in computing testing and artificial intelligence. The generators can be different type. A survey [18] shows different type of them from the architecture point of view and their pros and cons. Another point from this survey is there must be a semantic connection between generator of the data and the targeted system. In addition, the data is often generated to comply and meet very specific need and certain conditions. These conditions might not be easily found into the original and real datasets. On the other hand, the size of the different dimensions can be crucial. An example for managing big dimensional data is weights and using different trends in the dataset [19].

The other big part in the algorithm is the data recognition. Extraction the information from structured data is known as data mining. However, methods in this paper are related to text mining techniques that are dedicated to extract the information from textual data. A survey about text mining techniques can be for example association extraction and prototypical document extraction. The first type is designed to answer specific queries from the user and the other type is pointed toward fining a class of repetitive structures that can be recognized [20].

Most of the stored information has a text format (80%) [21]. In addition, text mining is believed to have potential value in natural understating solutions. The other issue of this concept is Knowledge Discovery from Text. A previous survey shows different concepts, operation and application of text mining. It can be concluded that these solutions have a wide variety of applications [22]. The different operations could be future extraction, search and retrieval, supervised and unsupervised classification, summarization, etc.

3. Technical Environment

3.1. Python

The majority of modern solutions regarding natural language processing are created using Python [13]. Python is high-level programming language that supports a variety of well-developed and widely used frameworks for language processing, including spaCy [16], NLTK [17] and others. The solution presented in this paper is based primarily on Python 3.7, along with other Python libraries described below.

3.2. RASA

RASA offers tools for building natural language processing. It consists of two independent modules. The first one Rasa natural language understanding (Rasa NLU) for language undemanding and Rasa Core for dialogue management. In the paper is examined the Rasa module. It combines a lot of natural language processing modules and libraries for machine learning. [2, 7, 8]

The RASA platform uses a model for intent recognition and key entities extraction. The model is trained with the sentences generated from step 1. They contain key entities with the corresponding indications. The platform allows regular expression search while recognition, which is very convenient for the numeric values in this case.

4. Approach

The approach consists of several steps. Firstly, samples are generated. Then they are used for the training of a model for recognition. Synthetic data generators are useful for testing data analytics applications and especially machine learning algorithms [14]. Once the model is trained it can recognize intents and entities. When entities are recognized they are put into table rows and saved. The extracted information could be later used as a base for generating training data. Figure 2 shows the steps of the approach used for data extraction.

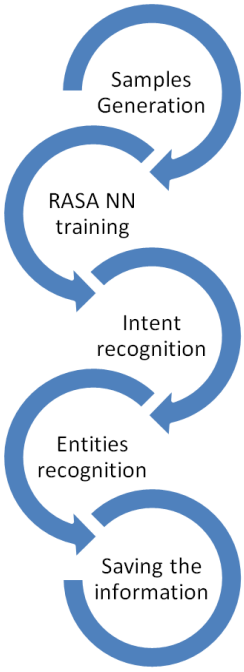

Figure 2. *Scheme of the approach for data extraction.*

4.1. STEP 1: Samples Generation

The requirements for this step are tables and database with description for each table. The table contains enumerable and numeric values. The enumerable column contains values of a finite set. Whilst the column area would contain a number therefore the column area is numeric column. The count of the items in the set of the distinct values in it is too large, so it cannot be covered as the enumerable. This type of information is contained in the additional database with the descriptions of the table that was mentioned earlier.

In addition to the types of the columns of the database also need to contain synonyms for each column. The synonyms are

words with which the column could be named or described in natural language. Each column should have at least one synonym in order for the column to be called. Here are some examples for synonyms.

Table 1. *Part of the Real Estate training table.*

EST TYPE	AREA	REGION	PRICE	B YEAR	CONSTR TYPE
maisonette	105	Egyptian	197551	2016	wooden
studio	116	Ukrainian	75964	2016	brick
3 bedroom	69	Welsh	59799	2007	wooden
maisonette	84	Scottish	131159	1987	brick
2 bedroom	134	Czech	53766	1990	panel
studio	65	Swedish	88844	1974	brick
studio	101	Japanese	192306	1993	brick
studio	88	Angolan	140748	2011	wooden
2 bedroom	152	Mexican	143697	1973	brick
office	160	Russian	91114	1982	wooden
…	…	…	…	…	…

Table 2. *Part of table with synonyms for the real estate context.*

EST TYPE	AREA	REGION	PRICE	B YEAR	CONSTR TYPE
type	area	region	price	construction year	construction type
property	space	locality	cost	construction	material
	field	city	amount	building	

The encountered requirements are used for the generation of a large set of training data for the neural network which is used for the information extraction. Each sample is generated using a record from the table. Sentence generator implemented in python selects random columns to take part in the produced sentence. For each selected column is randomly selected a synonyms that represents the column. For each column there is a set of synonyms that are used for its naming. Examples of generated sentences used for the training of the network.

1) Area (area synonym) and price (price synonym) for the storehouse (estate type key entity) property (estate type synonym), in Russian (region key entity) locality (region synonym) are 106 (number value) and 62222 (number value).
2) For the 3 bedroom (estate type key entity) property (estate type synonym) in Swedish (region key entity) region (region synonym) the amount (price synonym) is 137578 (number value).
3) 45845 (number value) is the amount (price synonym) for the 2 bedroom (estate type key entity) type (estate type synonym) in Belgian (region key entity) region (region synonym) and brick (construction type key entity) construction type (construction type synonym).

The selected synonyms and values from the record are placed into one of the predefined sentence templates. Those templates are natural language sentence with blank placeholders. In the placeholders are put values and synonyms so the sentence describes the information from the record. This is how the natural language sentences for the training are generated.

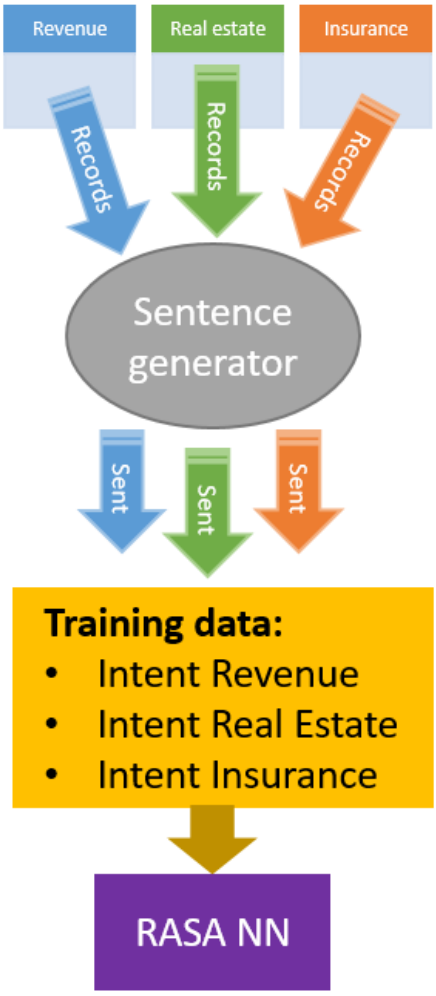

Figure 3. *Visual representation of sample generation and neural network training.*

4.2. STEP 2: RASA NN Training

The RASA platform provides model for intent recognition and key entities extraction. The model is trained with the sentences generated from step 1. They contain key entities with the corresponding indications. The platform allows regular expression search while recognition, which is very convenient for the numeric values in this case.

Once the sentences from each table are generated, they are fed to the RASA neural network. The network is also set up with additional settings for numeric values recognition via regular expression.

Figure 3 shows the visual representation of the first two steps of the approach.

4.3. STEP 3: Intent Recognition

After the training the model is able to produce percentage of confidence for the input data to belong to each intent. The intent is the meaning of the sentence [10]. The intent with highest percentage is taken into account. If the information is not related to the any of the topics all of the produced confidence indexes are approximately equal. In that case the information is considered not related to any of the observed topics. Once the intent is recognized the information extracted from the input will be referred to the corresponding table (topic).

4.4. STEP 4: Entities Recognition

As well as the intent the network recognizes key entities [9, 11]. They represent the concrete values of the tables and synonyms for the columns of the referred from the intent OLAP table [15]. Each entity contains name, value and confidence. The name and value are the affiliation of the entity and the fond concrete value. For example:

Input: The property type is 2 bedrooms

Result:

1. Intent: real estate
2. Entities:
   A. Entity 1:
      a) Name: estate type synonym
      b) Value: property type
   B. Entity 2:
      a) Name: estate type key entity
      b) Value: 2 bedrooms

4.5. STEP 5 Information Saving

Lastly the records are sent to Data Extractor also implemented in Python. The Data extractor produces records from the neural network results. The records are the structured view of the unstructured input data (sentences in natural language).

If there is additional information (like the underlined sentence in the input example) which is not related to the context and has no corresponding column in the table, it is not recognized as entity or in other words as value that could be included in the table so they are skipped.

Figure 4 shows visual representation of the last three steps of the approach.

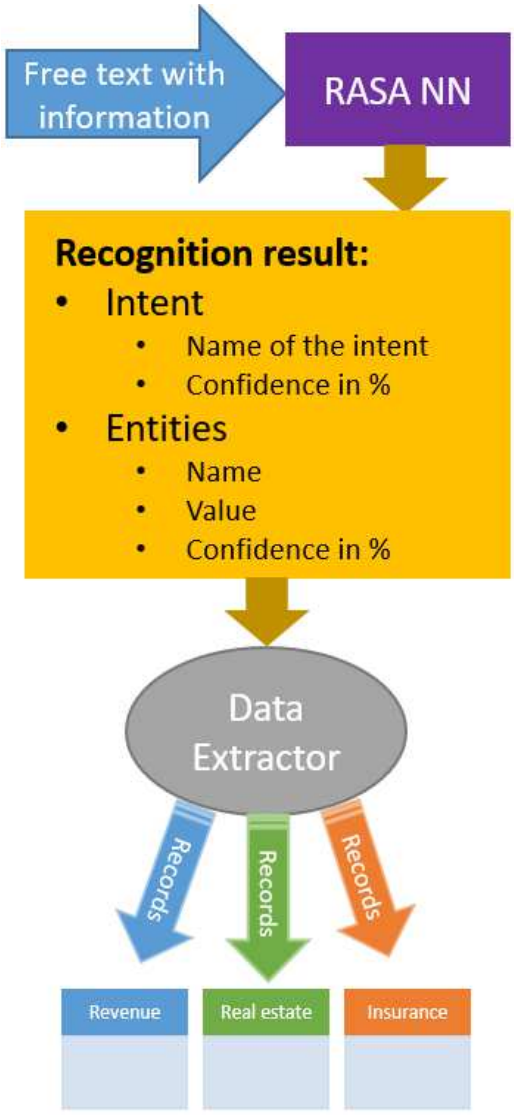

Figure 4. *Visual representation of step 3 and step 5.*

5. Results

For the test results are used three topics. Each has its own context that has nothing in common with the others. This shows that the system is independent from the context as soon as the needed information for the training is available. The environment used for the testing of the approach used 200 sentences for each context. They are generated by randomly selecting records from the training tables. Each training table has 10000 rows.

The first test is made with the fallowing paragraph:

For the revision year of 2019 the revenue in the Swiss region is 600. The revenue in the German region for revision year 2018 is 550. The employees are 10.

Found entities could be seen in table 3.

The entities from table 3 are transformed into records that are shown in table 4.

Table 3. *Entities extracted from the first text.*

Name	Value	Confidence
revision_year	revision	0.949
measure_val	2019	0.9947
revenue	revenue	0.8759
region_val	Swiss	0.8891
region	region	0.8919
measure_val	600	0.9996
revenue	revenue	0.8759
region_val	German	0.9566
region	region	0.9536
revision_year	revision	0.9922
measure_val	2018	0.9993
measure_val	550	0.9996
employees_num	employees	0.9585
measure_val	10	0.9997

Table 4. *Records from the first text.*

Name	Region	Revenue	Revision year	Employees number
	Swiss	600	2019	
	German	2018	550	10

Another test is made with a text consisting of seven sentences that refers to the real estate topic.

The price is 28286.19 EUR. The property type is 2 bedrooms. The construction stage is under construction The number of floors is 1 The floor is first residential floor It is located in the German region. The area is 54 m^2. The type of construction is brick. The year of construction is 2022.

The extracted information is organized into records in table 5.

Table 5. *Records from the second text.*

EST TYPE	AREA	REGION	PRICE	B YEAR	CONSTR TYPE
2 bedroom	54	German	28286.19		brick

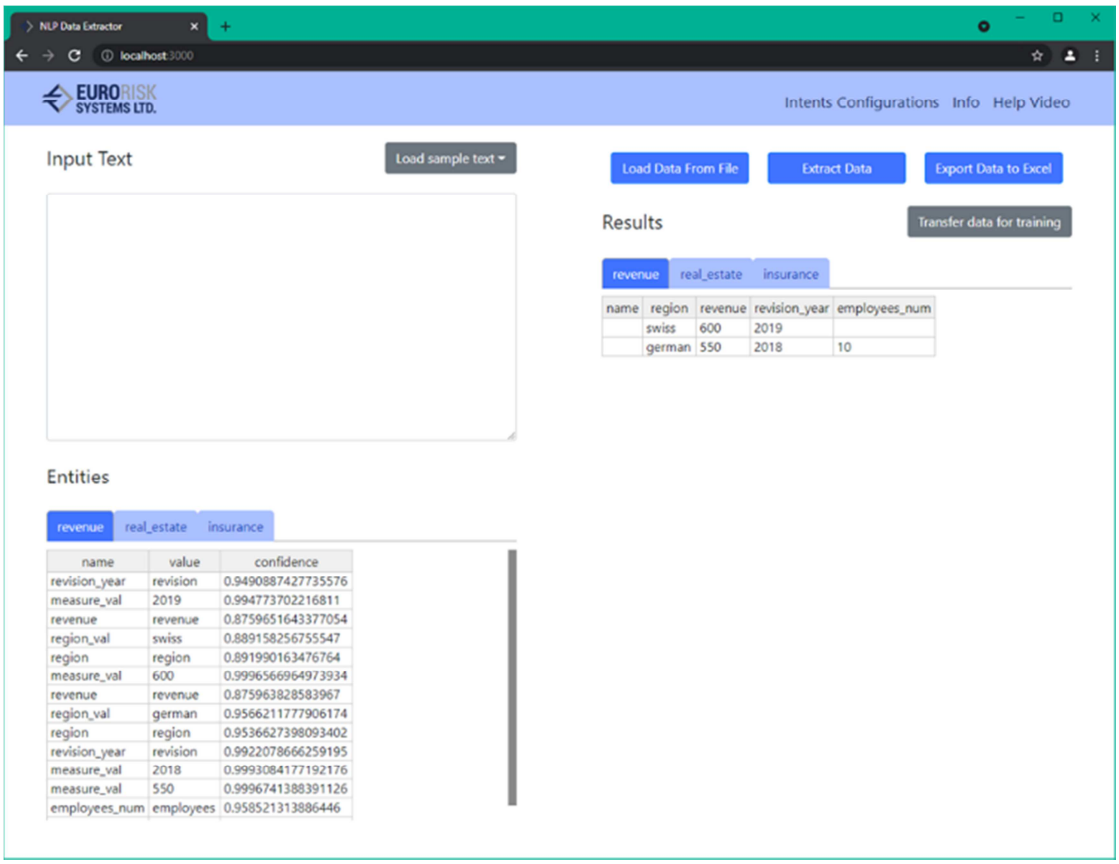

Figure 5. *Home page of the application.*

6. Application

A simple web application was developed for easier access to the data extraction functionalities. It also provides opportunities for changing some of the configurations of the system and to test the system for three predefined contexts.

6.1. Data Extraction

The home page which is show in figure 5 has text box for the unstructured data in text format. The data could be e either typed in the box or loaded from file. When the data is extracted the results table shows the records (structured data) that was generated from the text. There are three tabs each for a context. The contexts that network understands are companies' revenues, real estates and insurances. Those are used just for the demo version. The system can be set up for recognizing any context. That depends on the initial training data. In the entities table are shown the recognized entities. The name of the entity is the name of group it belongs to. For example, Swiss and German are values for the region column. The value is the concrete word matched as entity and confidence is the probability of the value to belong to the group displayed in the first column (name). The entities with lower confidence that a predefined threshold value are not taken into account when creating the records.

6.2. Network Configurations

The editable network configuration settings could be found in figure 6. The training and synonyms tables is used for the generation of the sentences used for the training of the model. Editing the training table will reflect the distinct values that the network recognizes. Changes in the synonyms table will also reflect the network recognition ability. The synonyms refer to the columns of the training table. They are quite important for the recognition of the numeric entities and placing them into the results records. The network will work with the words displayed as synonyms and distinct values.

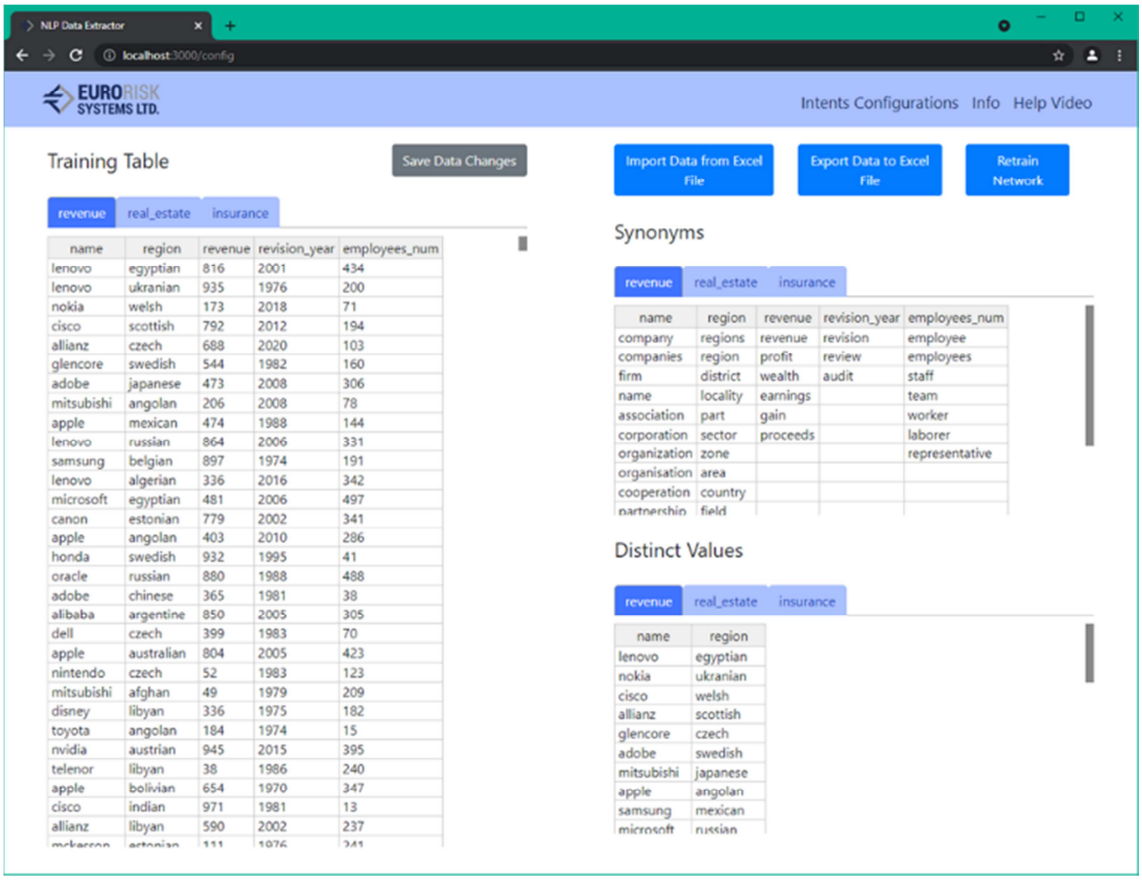

Figure 6. *Configuration setting page of the application.*

7. Conclusion

The presented in this paper approach enables the conversion unstructured data to a structured one. Since the format of the data is crucial most of the time, the extraction will make use of the data which until now is considered as useless. The approach consists of several steps. The first step is generation of samples for different topic which are used for the training of a neural network. Once the network is trained, recognition of entities from a text is allowed and they can be extracted to the topics that are set up previously. The extracted entities are structured and stored into new records.

When the extracted information is voluminous enough it

can be used as a new table for the generation of the sentences. This operation will improve the accuracy of the training network. But for its execution is needed a large set of collected records. On figure 7 is shown the process of the circularity when the extracted information is used as a training data.

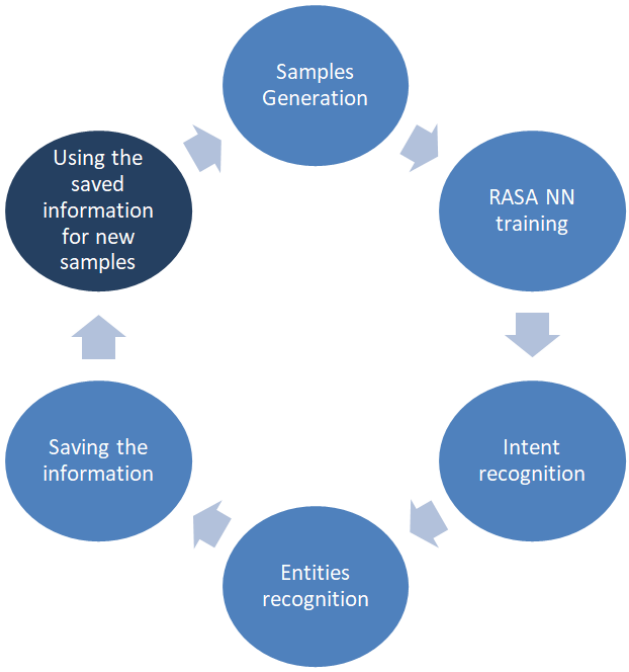

Figure 7. *Complete circularity of the steps in the process.*

References

[1] Holst A. (2021, June 30). Amount of data created, consumed, and stored 2010-2025. https://www.statista.com/statistics/871513/worldwide-data-created/

[2] Bocklisch T., Faulkner J., Pawlowski N., Nichol A. (2017). Rasa: Open Source Language Understanding and Dialogue Management.

[3] Petrov. C. (2021, June 30). 25+ Impressive Big Data Statistics for 2021. https://techjury.net/blog/big-data-statistics/#gref

[4] Taylor. C. (2021, June 30). Structured vs. Unstructured Data. https://www.datamation.com/big-data/structured-vs-unstructured-data/

[5] Lomotey RK, Deters R. RSenter: terms mining tool from unstructured data sources. Int J Bus Process Integr Manag. 2013; 6 (4): 298.

[6] Gantz J, Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView IDC Analyze Future. 2012; 2007 (2012): 1–16.

[7] Jiao, A. (2020). An intelligent Chatbot system based on entity extraction USING Rasa NLU and neural network. *Journal of Physics: Conference Series*, *1487*.

[8] Bagchi, M. (2020). Conceptualising a Library chatbot using open Source Conversational artificial intelligence. DESIDOC Journal of Library & Information Technology.

[9] RASA. (2020, July 27) Introducing DIET: state-of-the-art architecture that outperforms fine-tuning BERT and is 6X faster to train. https://blog.rasa.com/introducing-dual-intent-andentity-transformer-diet-state-of-the-art-performanceon-a-lightweight-architecture/.

[10] Wochinger, T. (2019, June 4). *Rasa NLU in DEPTH: INTENT CLASSIFICATION*. The Rasa Blog: Conversational AI Platform, Powered by Open Source. https://blog.rasa.com/rasa-nlu-in-depth-part-1-intent-classification/.

[11] Wochinger, T. (2019, June 4). *Rasa NLU in DEPTH: Entity recognition*. The Rasa Blog: Conversational AI Platform, Powered by Open Source. https://blog.rasa.com/rasa-nlu-in-depth-part-2-entity-recognition/.

[12] Baldauf, Matthias & Dustdar, Schahram & Rosenberg, Florian. (2007). A Survey on context-aware systems. Information Systems. 2. 10.1504/IJAHUC.2007.014070.

[13] Zola, A. (2021, March 31). *The 5 best programming languages for AI*. Springboard Blog. https://www.springboard.com/blog/ai-machine-learning/best-programming-language-for-ai/.

[14] Mendonca, Sandro & Brito, Yvan & Santos, Carlos & Lima, Rodrigo & Araujo, Tiago & Meiguins, Bianchi. (2020). Synthetic Datasets Generator for Testing Information Visualization and Machine Learning Techniques and Tools. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2991949.

[15] Wrembel, Robert, and Christian Koncilia. Data Warehouses and Olap: Concepts, Architectures, and Solutions. IRM Press, 2007.

[16] *spaCy · INDUSTRIAL-STRENGTH natural language processing in Python*. · Industrial-strength Natural Language Processing in Python. (2020, July 30). https://spacy.io/.

[17] Loper, E., & Bird, S. *Nltk: The natural Language Toolkit*.

[18] Popić, Srđan & Velikic, Ivan & Teslic, Nikola & Pavkovic, Bogdan. (2019). Data generators: a short survey of techniques and use cases with focus on testing. 10.1109/ICCE-Berlin47944.2019.8966202.

[19] G. Albuquerque, T. Lowe and M. Magnor, "Synthetic Generation of High-Dimensional Datasets," in IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2317-2324, Dec. 2011, doi: 10.1109/TVCG.2011.237.

[20] Rajman M., Besançon R. (1998) Text Mining: Natural Language techniques and Text Mining applications. In: Spaccapietra S., Maryanski F. (eds) Data Mining and Reverse Engineering. IFIP — The International Federation for Information Processing. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-35300-5_3

[21] Hotho, Andreas & Nürnberger, Andreas & Paass, Gerhard. (2005). A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology. 20. 19-62.

[22] Gupta, Vishal & Lehal, Gurpreet. (2009). A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. 1. 10.4304/jetwi.1.1.60-76.