УДК 004.82 https://doi.org/10.15407/pp2022.03-04.154

EXTRACTING STRUCTURE FROM TEXT DOCUMENTS BASED ON MACHINE LEARNING

Kuzma Kudim, Galyna Proskudina

This study is devoted to a method that facilitates the task of extracting structure from the text documents using an artificial neural network. For the method to work it requires a set of manually labeled documents to train the network. The trained model can be used to extract sections of documents bearing similar structure.

Keywords: natural language processing, information extraction, machine learning, neural network.

Дослідження присвячене методу, що вирішує задачу автоматичного витягу структури з слабо структурованих текстових документів за допомогою штучної нейронної мережі. Для того, щоб цей метод працював, потрібен розмічений вручну набір документів для навчання мережі. Навчену модель можна використовувати для витягу розділів документів, що мають подібну структуру.

Ключові слова: обробка природної мови, видобуток інформації, машинне навчання, нейронні мережі.

Introduction

There are a lot of text documents that have rich representational formatting, easily readable and understandable by human but not intended for automatic processing. Examples are scientific papers, legal documents, books. All of them have implicit logical structure like title page with title and author, publisher's imprint, chapters, references. If we make this logical structure explicit then it can be automatically processed. And then it can be used either as metadata describing the document or as input for further fine-grained information extraction.

Here we describe a method that facilitates the task of extracting structure from the text documents using an artificial neural network. For the method to work it requires a set of manually labeled documents to train the network. The trained model can be used to extract sections of documents bearing similar structure.

Previously we already described two other methods of data extraction from semi-structured text documents. One based on detecting patterns using regular expressions and another based on linguistic rules [1, 2]. Both of these methods require special skills to set up them for a particular type of documents, and to update the system for the changed structure. The method based on machine learning described here has the benefit of not requiring programming skills for usage. The initial set up requires only an accurately labeled set of documents, and this labeling can be made by any person with basic understanding of the target structure of the document in the usual sense.

Overview

The paper consists of three main sections, as follows.

First of all, data should be prepared to train, validate and evaluate the model. Data preparation includes collecting corpora of documents, converting a variety of file formats into plain text, and manual labeling each document structure. Finally, the dataset is split into three subsets for model training, validation and test in 70/15/15 ratio respectively.

Building and training the model is the central part of the work. Document is split into tokens and then into paragraphs. The text paragraphs are represented as feature vectors to provide input to the neural network that consists of three fully connected layers. The model is trained and validated on the selected data subsets.

After the model is trained showing a good F1 score on validation dataset for the selected features, it's time to evaluate the results on a very new data, i.e. test dataset. The final performance is calculated per label using precision, recall, and F1 measures, and overall average.

Data preparation

Corpora. A selected subset from the thesis corpora from the National library of Ukraine by V.I.Vernadsky is used as a dataset. The whole corpora consists of nearly 65000 documents. A subset of 100 theses is selected and split into 70 documents as training set, 15 as validation set, and 15 as test dataset for final evaluation.

Conversion to plain text. The selected documents are in doc and rtf formats. As a preliminary step, this variety of file formats is converted to plain text using LibreOffice (https://www.libreoffice.org) from command line as follows:

```
soffice --headless --convert-to txt --outdir out_dir in_file
```

Output text files are in UTF-8 encoding with BOM signature at the file start, so additionally the first three bytes of each file are removed.

Labeling. Our goal is to select top-level sections of the document that are potentially useful for further information extraction. That means, from one side, we are not interested in thesis main thematic content, and, from the other side, we don't care of fine-grained data contained deeper in each section on this stage. The factual data extraction can be the next step after larger document sections are successfully extracted.

19 labels shown in Table 1 are selected to reflect the desired top level logical structure of the thesis document. Each label covers the whole section of the document, although sections can differ much in size and inner complexity. For example, a section labeled SPEC covers speciality digital code and name, or maybe a list of such records. Another section labeled PUBLICATIONS includes all listed publications as a whole section of the document. Fine-grained information extraction is out of scope of current work.

Special label O is used internally to represent absence of any specific label.

Table 1. Structural labels for thesis document

Label	Document section
MAIN_ORG	Organization this document is related to in general, at the top of the title page
AUTHOR	Thesis author
UDK	UDC classifier
TITLE	Thesis title
SPEC	Thesis speciality code and name
DEGREE	Target scientific degree of the thesis
CITY_YEAR	City and year in the footer of the title page
WORK_ORG	Author's work organization
SUPERVISOR	Scientific supervisor
OPPONENTS	Scientific opponents
LEAD_ORG	Leading organization for the thesis
DEFENSE	Information about thesis defense event
LIBRARY	Where the thesis manuscript is stored
SENT	When participants were notified by mail
SECRETARY	Scientific secretary
PUBLICATIONS	Author's publications for the thesis
ABSTRACT_UK	Abstract in Ukrainian
ABSTRACT_EN	Abstract in English
ABSTRACT_RU	Abstract in Russian
O	Used internally to represent empty label

All 100 documents from the corpora are manually labeled using Label Studio (https://labelstud.io/) open source data labeling tool as shown in Figure 1, and exported in JSON format.

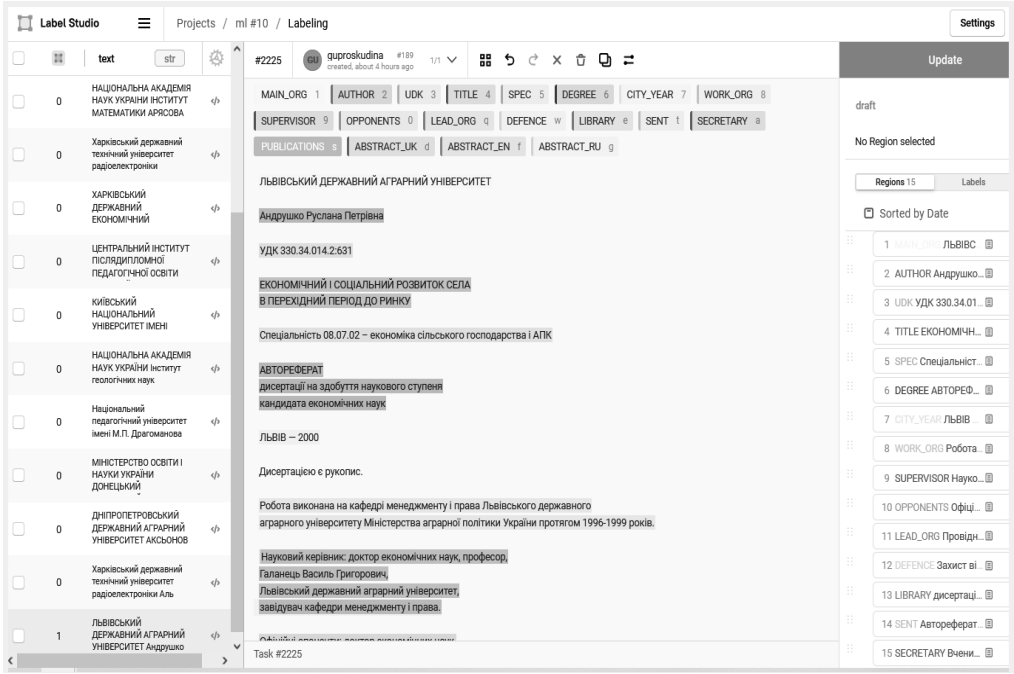

Fig. 1. Manually labeling process using Label Studio

Model

Feature vector representation of paragraph. Document is represented as a sequence of paragraphs, and each paragraph is converted to a feature vector of N dimensions. Paragraph features are listed in Table 2. First Ns = 12 features are quite simple, each reflecting one statistic value in a paragraph [3]. For untrivial features the explanation follows.

Amongst other features a vector representing dictionary word count is used. A short dictionary of Nd = 105 words is built of the most frequent words met in labeled sections, 10 most frequent words for each label over all documents. The dictionary word vector is concatenated to the main feature vector. This feature adds Nd dimensions to the feature vector.

The same goes for character frequencies in a paragraph. Dictionary for characters from the training set contains Nc = 293 characters. Here the paragraph is considered as a bag of characters and the frequency of each character is calculated. It is also concatenated to the main feature vector adding Nc dimensions.

Another special feature represents a non-empty label preceding the current paragraph in the document. This feature catches the global order of labels. This feature has Nl = 20 dimensions that is equal to the count of non-empty structural labels. When using a window of nearby paragraphs for model training then this global feature is concatenated only once to the input vector. How the window is used is described in the next section.

From the above we can see that the feature vector representing the paragraph has Np = Ns + Nd + Nc + Nl = 430 dimensions. Specific numbers of simple features, word and character dictionary size, label count can vary depending not only on a task in question but also when optimizing trained model scores.

Table 2. Paragraph features

Feature	Comment
Paragraph start position	Paragraph position measured in character
Paragraph size	Paragraph size measured in tokens
Words count	Count tokens consisting of cyrillic and latin letters only
Numbers count	Count tokens consisting of digits
Lower-cased word count	Count words with all characters in lower case
Capitalized word count	Count tokens with first character in upper case
Uppercased word count	Count words with all characters in upper case
Dots count	Count of dot characters in a paragraph
Commas count	Count of comma characters in a paragraph
Starts with upper-cased word	The first word of a paragraph is in upper case
Starts with capitalized word	The first word of a paragraph has the first char in upper case
Starts with number	The first token of paragraph is a number
Dictionary word counts	Vector with each element equal to dictionary word frequency in a paragraph
Character counts	Vector with each element equal to character frequency in a paragraph
Previous label	Vector representing label of the previous section in the document

The dictionary of the most frequent words in all labeled regions of the training corpus is shown in Table 3.

Table 3. Dictionary of the most frequent words in labeled regions

.	1	харків	відбудеться
університет	2	одеса	о
україни	3	донецьк	засіданні
інститут	та	львів	спеціалізованої
академія	на	дніпропетровськ	можна
державний	в	виконана	дисертацією
національний	,	робота	бібліотеці
і	у	університеті	ознайомитись
імені	з	освіти	університету
наук	–	науки	6
аль	01	державному	розісланий
-	05	науковий	р
анатолій	спеціальність	керівник	‘‘
а	00	професор	’’
миколайович	02	доктор	«
михайлівна	4	опоненти	року
‘	здобуття	кафедри	_
володимирівна	дисертації	офіційні	секретар
сергійович	наукового	провідна	вчений
микола	ступеня	установа	с
удк	кандидата	м	//
:	автореферат	кафедра	и
0	технічних	захист	что
)	5	ради	of
(	київ	вченої	the

Neural network training

Window of w = 3 consecutive paragraphs is used as input to the neural network [3]. The previous label feature is added only for the current paragraph. That gives us the input layer size of w*(Np-Nl) + Nl = 1250. Hidden layer size was chosen empirically to be of 40 nodes. Output layer size is equal to Nl = 20, it is defined by chosen labels count.

To train the neural network, 70 documents of training corpora are converted into vectors. Due to the chosen window of width 3 , each document is augmented with one padding paragraph at the beginning and one at the end.

For each paragraph in the document, the window consists of one paragraph before, the current paragraph, and one paragraph after as one sample input for training. Then these three paragraphs are converted to the input vector by concatenating their feature vectors. And the vector of the previous label feature is concatenated to these three.

A vector representing the label of the current paragraph is used for the training sample output. The vector consists of 0 in each position except of 1 in the position representing the label of the current paragraph (Fig. 2).

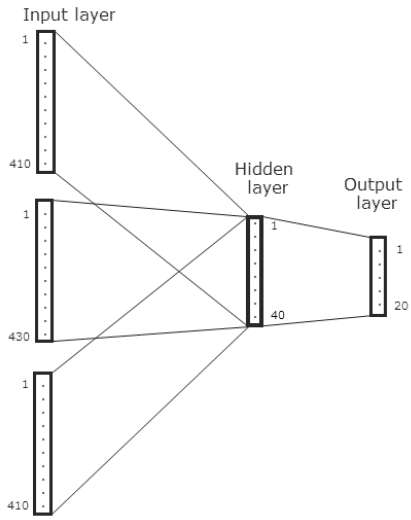

Fig. 2. Neural network for our example

In this way 70 documents of training dataset provide 18423 training samples. Neural network is trained with RPROP method implemented in FANN (Fast Artificial Neural Network - http://leenissen.dk/fann/wp/) library [4,5], it is an adaptive back propagation method which doesn't require to set learning rate explicitly. Mean square error is calculated once per epoch for the whole training set. It takes less than 50 epochs to achieve mean square error less than 0.001.

We used the validate set of 15 manually labeled documents to run the trained model, compare labeling results and empirically select features to use (see Fig 3a, 3b).

Paragraph	Best guess	O	MAIN_ORG	AUTHOR	UDK	TITLE	SPEC	DEGREE	CITY_YEAR	WORK_ORG	SUPERVISOR	OPPONENTS
ЛЬВІВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ	MAIN_ORG		1									
	O	0.99										
Андрушко Руслана Петрівна	AUTHOR			1								
	O	1										
УДК 330 . 34 . 014 . 2 : 631	UDK				1							
	O	0.99										
ЕКОНОМІЧНИЙ І СОЦІАЛЬНИЙ РОЗВИТОК СЕЛА	TITLE					0.97						
В ПЕРЕХІДНИЙ ПЕРІОД ДО РИНКУ	TITLE					0.95						
	O	1										
Спеціальність 08 . 07 . 02 – економіка сільського господарства і АПК	SPEC						0.47					
	O	1										
АВТОРЕФЕРАТ	DEGREE							1				
дисертації на здобуття наукового ступеня	DEGREE							1				
кандидата економічних наук	DEGREE							0.95				
	O	1										
ЛЬВІВ — 2000	UDK											
	O	1										
Дисертацією є рукопис .	O	0.78										
	O	1										
Робота виконана на кафедрі менеджменту і права Львівського державного	WORK_ORG									0.81		
аграрного університету Міністерства аграрної політики України протягом 1996 - 1999 років .	TITLE											
	O	0.96										
Науковий керівник : доктор економічних наук , професор ,	SUPERVISOR										0.54	0.36
Галанець Василь Григорович ,	SUPERVISOR										0.82	
Львівський державний аграрний університет ,	SUPERVISOR										0.72	
завідувач кафедри менеджменту і права .	SUPERVISOR										0.6	0.2
	O	0.98										
Офіційні опоненти : доктор економічних наук ,	OPPONENTS										0.15	0.84
Костирко Ігор Григорович ,	OPPONENTS										0.17	0.66
Жидачівський целюлозно - паперовий комбінат ,	OPPONENTS										0.16	0.9
заступник директора з питань економіки ;	OPPONENTS										0.05	0.44
кандидат економічних наук ,	OPPONENTS										0.07	0.85

Fig. 3a. Output in HTML format of test corpus documents for visual comparison: marked up manually and using the model

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ :	PUBLICATIONS 0.24	0.91
1 . Біль Р . П . Розвиток соціальної сфери агропромислового комплексу Львівщини // Економічний вісник . - Львів : ЛДСГІ . - 1995 . - № 1 . - С . 75 - 76 . - 0 , 1 друк . арк .	PUBLICATIONS 0.14	0.98
2 . Андрушко Р . П . Напрями розвитку багатоукладної аграрної економіки // Вдосконалення виробничих відносин і управління сільськогосподарським виробництвом різних форм господарювання : Зб . наук . пр . - Львів : ЛДСГІ . - 1996 . - С . 50 - 53 . - 0 , 2 друк . арк .	PUBLICATIONS 0.08	0.98
3 . Андрушко Р . П . Соціально - економічна ситуація на селі // Вісник ЛДАУ : Економіка сільського господарства . - Львів : ЛДАУ . - 1997 . - № 3 . - С . 148 - 151 . - 0 , 2 друк . арк .	PUBLICATIONS 0.07	0.99
4 . Андрушко Р . П . Економічний і соціальний розвиток села у Львівській області // Вісник ЛДАУ : Економіка АПК . - Львів : ЛДАУ . - 1998 . - № 4 . - С . 194 - 199 . - 0 , 38 друк . арк .	PUBLICATIONS 0.06	0.99
5 . Андрушко Р . П . Соціальний розвиток села як пріоритетний напрямок аграрної реформи // Вісник ЛДАУ : Економіка АПК . - Львів : ЛДАУ . - 1999 . - № 5 . - С . 191 - 195 . - 0 , 22 друк . арк .	PUBLICATIONS 0.06	0.99
6 . Андрушко Р . П . Економічний і соціальний розвиток села в перехідний період до ринку // Вісник ЛДАУ : Економіка АПК . - Львів : ЛДАУ . - 1999 . - № 6 . - С . 86 - 89 . - 0 , 2 друк . арк .	PUBLICATIONS 0.06	0.99
7 . Андрушко Р . П . Комплексний підхід до проблеми економічного і соціального розвитку села Львівщини // Теорія і практика розвитку АПК : Тези доп . Міжнародної науково - практичної конференції , присвяченої пам ' яті професора Євгена Храпливого . - Львів : ЛДАУ . - 1999 . - С . 65 - 67 . - 0 , 13 друк . арк .	PUBLICATIONS 0.07	0.97
АНОТАЦІЇ	ABSTRACT_UK 0.33	0.78
Андрушко Р . П . Економічний і соціальний розвиток села в перехідний період до ринку . - Рукопис .	ABSTRACT_UK 0.28	0.89
Дисертація на здобуття наукового ступеня кандидата економічних наук за спеціальністю 08 . 07 . 02 – економіка сільського господарства і АПК . Львівський державний аграрний університет . Львівська область , Жовківський район , м . Дубляни , 2000 .	ABSTRACT_UK 0.15	0.89
На прикладі Львівської області , де аграрна реформа розпочалася раніше і проходить динамічніше , ніж в інших регіонах , дається оцінка ситуації в аграрному секторі економіки , економічного і соціального розвитку села в їх єдності і взаємозалежності .	ABSTRACT_UK 0.16	0.95

Fig. 3b. Output in HTML format of test corpus documents for visual comparison: marked up manually and using the model

Results evaluation

The test dataset of another 15 manually labeled documents is used to make the final evaluation of the model (Fig. 4). It is executed independently after the model parameters are adjusted to improve results for the test dataset. Standard precision, recall, and F1 measures are used for evaluation. The strict check is made for the whole section of a document to be labeled correctly, i.e. partial overlap of correct labeling only for some paragraphs in the section is considered wrong. Scores are calculated over all documents in the dataset.

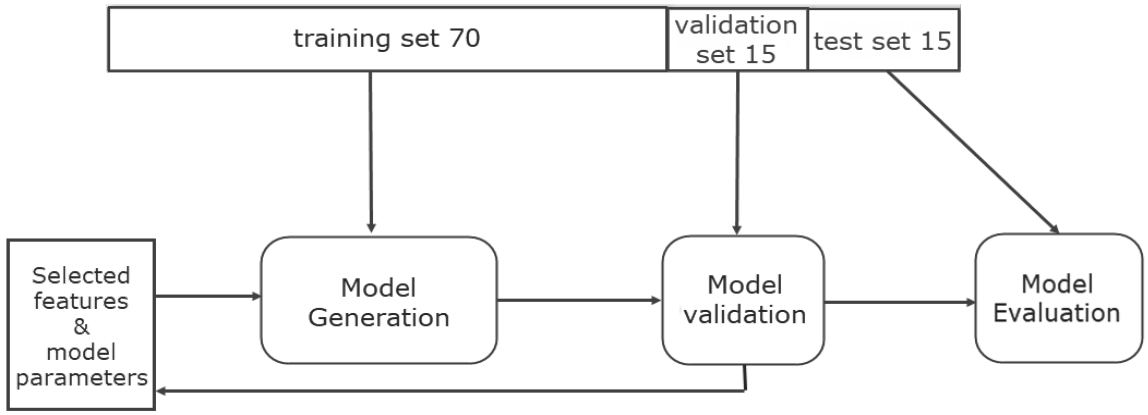

Fig. 4. The train and validation datasets are used to build the model, while the test dataset is used to evaluate it

The overall F1 score averaged over all labels is 84. Detailed results can be found in Table 4. All values are multiplied by 100 for convenience.

Table 4. Trained model results evaluation. Precision, recall and F1-score per label. All numbers multiplied by 100 for convenience

Label	Precision	Recall	F1
MAIN_ORG	81	87	84
AUTHOR	79	73	76
UDK	100	100	100
TITLE	79	73	76

continuation tab. 4.

SPEC	87	93	90
DEGREE	93	100	97
CITY_YEAR	100	100	100
WORK_ORG	87	87	87
SUPERVISOR	38	92	53
OPPONENTS	80	80	80
LEAD_ORG	93	93	93
DEFENSE	100	100	100
LIBRARY	80	92	86
SENT	100	100	100
SECRETARY	87	87	87
PUBLICATIONS	31	33	32
ABSTRACT_UK	80	80	80
ABSTRACT_EN	100	100	100
ABSTRACT_RU	80	80	80
Average	**83**	**87**	**84**

Interpretation

The trained model shows best results on short document sections with consistently strong statistical text features. The long sections that include heterogeneous paragraphs are predicted the worst, e.g. publications section consists of section title followed by list items, and while the latter are detected pretty good on paragraph level, the section title often is mispredicted as not having a label, and thus the whole section is considered incorrect. In general, scores are high enough for practical applications.

Conclusions

A method of extracting high-level sections from weakly structured text documents is built. The method is based on an artificial neural network and thus requires a training dataset. The dataset is manually labeled to build, validate and evaluate the model. The model performs well and proves that machine learning can be successfully applied to the problem of extracting logical structure from the text documents. It is also simpler than rule-based methods that require special skills to set up the algorithm.

Future research goal is to improve scores, especially for long document sections, by modifying neural network architecture.

References

1. KUDIM K.A., PROSKUDINA G.YU. (2019). Methods and tools for extracting personal data from theses abstracts Problems in programming. [online – pp.isofts.kiev.ua] (2). P. 38–46. (in Russian). Available from: http://pp.isofts.kiev.ua/ojs1/article/view/359 [Accessed 04/08/2022].
2. KUDIM K.A., PROSKUDINA G.YU. (2020). A method for extracting data from semistructured documents Problems in programming. [online – pp.isofts.kiev.ua] (1). P. 25–32. (in Russian). Available from: http://pp.isofts.kiev.ua/ojs1/article/view/388 [Accessed 04/08/2022].
3. YI HE. (2017) Extracting Document Structure of a Text with Visual and Textual Cues. University of Twente. Elsevier. 78 p. (in English). Available from: https://essay.utwente.nl/72979/1/Yi He - master thesis - final version.pdf [Accessed 05/08/2022]
4. STEFFEN NISSEN. (2005). Neural Networks Made Simple. Software 2.0. [online – software20.org] (2). P. 14–19. Available from: http://fann.sourceforge.net/fann_en.pdf [Accessed 05/08/2022].
5. MARTIN RIEDMILLER, HEINRICH BRAUN. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm – Neural Networks. IEEE International Conference. P.586-591. Available from: https://paginas.fe.up.pt/~ee02162/dissertacao/RPROP paper.pdf

Received 11.08.2022

About authors:

Kudim Kuzma Alekseevich,
junior researcher of Institute of Software Systems NAS of Ukraine.
Publications in Ukrainian journals – 19.
Publications in foreign journals – 2. 1
http://orcid.org/0000-0001-9483-5495,

Proskudina Galyna Yurievna
researcher of Institute of Software Systems NAS of Ukraine.
Publications in Ukrainian journals – 32.
Publications in foreign journals – 15.
http://orcid.org/0000-0001-9094-1565.

Place of work:

Institute of Software Systems NAS of Ukraine,
03187, Kyiv-187,
Academician Glushkov Avenue, 40, build 5.
Phone: +38(050) 368 49 27.
E-mail: kuzmaka@gmail.com,
guproskudina@gmail.com

Прізвища та ініціали авторів і назва доповіді англійською мовою:
Kudim K.A., Proskudina G.Yu.
Extracting structure from text documents based on machine learning

Прізвища та ініціали авторів і назва доповіді українською мовою:
Кудім К.О., Проскудіна Г.Ю.
Витяг структури з текстових документів на основі машинного навчання