Vilniaus universitetas

Ataskaitinė informatikos krypties doktorantų konferencija 2024-10-04

Rolandas Gricius (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)

Preliminari darbo tema.

Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose.

Recognising the contents in digitised structured documents.

Darbo vadovas.

Prof. dr. Igoris Belovas.

Doktorantūros studijų laikotarpis.

2021 m. spalio mėn. 1 d. – 2025 m. rugsėjo mėn. 30 d.

Ataskaitinis laikotarpis.

2024 m. balandžio mėn. 1 d. – 2024 m. rugsėjo mėn. 30 d.

Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2021/2022)	2	3
II (2022/2023)	2	1
III (2023/2024)		
IV (2024/2025)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2021/2022)				1						
II (2022/2023)	1	1								
III (2023/2024)		2			1		Įteikta po recenz. pastabų			
IV (2024/2025)	1				1					
Iš viso:	2	3		1	2					

Ataskaitinio pusmečio darbo planas ir jo vykdymo suvestinė

Egzaminai 2023/2024 (II pusmetis)		
Planas	Įvykdyta	Būklė
-	-	-

Dalyvavimas konferencijose 2023/2024 (II pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
-	R. Gricius, I. Belovas. Using Large Language Models in Anti-Money Laundering Automation. 13th Counter Fraud, Cybercrime and Forensic Accounting Conference, 2024 birželio 12-13 d., Portsmutas, Didžioji Britanija	Tarptautinė
-	R. Gricius, I. Belovas. On Key Information Extraction from Business Documents. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)* 2024 rugsėjo 9-13 d., Vilniuje	Tarptautinė

Publikacijos 2023/2024 (II pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
IEEE Access	R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models. *IEEE Access*.	Ištaisytas atsižvelgus į recenzentų pastabas ir įteiktas antram recenzavimo ciklui: 2024-10-02	Žurnalas turi cituojamumo rodiklį (impact factor) CA WoS duomenų bazėje.

Informacija apie tarptautinius renginius ir publikacijas, kuriose pateikti pagrindiniai disertacijos rezultatai

Dalyvavimas tarptautinėse konferencijose	
	Aprašas
1.	R. Gricius, I. Belovas "Generation of Synthetic Invoices for the Training of Machine Learning Models". International Conference on Pattern Recognition Applications and Methods (ICPRAM) 2023, Lisabona, Portugalija, 2023-02-22 – 24 d.
2.	R. Gricius, I. Belovas. Using Large Language Models in Anti-Money Laundering Automation. 13th Counter Fraud, Cybercrime and Forensic Accounting Conference, 2024 birželio 12-13 d., Portsmutas, Didžioji Britanija
3.	R. Gricius, I. Belovas. On Key Information Extraction from Business Documents. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)* 2024 rugsėjo 9-13 d., Vilniuje

	Publikacijos (tik su citavimo rodikliu)	
	Bibliografinis aprašas	Būklė
1.	R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models. *IEEE Access*.	Ištaisytas atsižvelgus į recenzentų pastabas ir įteiktas antram recenzavimo ciklui: 2024-10-02

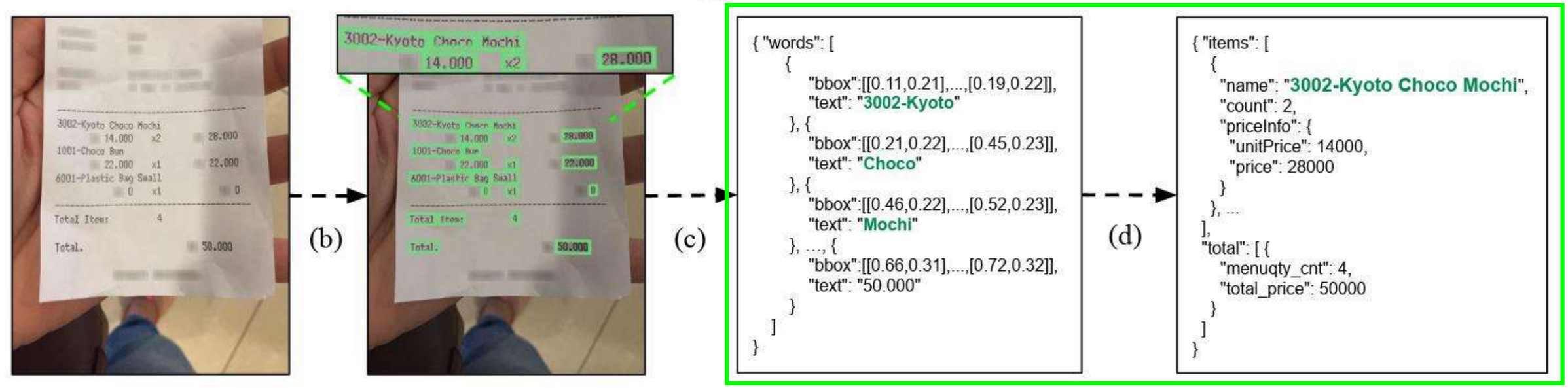
Kita mokslinė ir akademinė veikla

- **Laboratorinių darbų vedimas MIF bakalauro studentams, kursas "Informacinių sistemų saugos pagrindai", 2023/2024 rudens semestre**
- **Laboratorinių darbų vedimas MIF bakalauro studentams, kursas "Finansų inžinerija ir modeliavimas", 2024/2025 rudens semestre**
- **Konsultavimas bakalauro studento rašto darbo ruošimo klausimais, tema "*Recognising the contents in digitised financial documents*"**

Tyrimo objektas, tikslas ir uždaviniai

- **Tyrimo objektas – teksto esybių atpažinimas pagal tekstą ir jo išdėstymą**
- **Tikslas – naudojant natūralios kalbos apdorojimo metodus, atpažinti ir ištraukti tolesniam apdorojimui sąskaitos duomenis, reikšmingus:**
  - **teisėtumui – privalomus pagal teisės aktus duomenis**
  - **apskaitai – data, pirkėjo ir pardavėjo duomenys, sandorio ir mokesčių sumos**
  - **sandorio vykdymui – pristatymo duomenys, apmokėjimo detalės**
- **Uždaviniai – sudaryti duomenų rinkinį tyrimui, atlikti teorinį tyrimą identifikuojant metodus, empirinį tyrimą palyginant jų veikimą ir modifikuoti pritaikant Lietuvos specifikai ir surinktiems duomenims**

Klasikinė darbų seka



Tekstas + išdėstymas ----> dokumento informacija

End-to-End apdorojimas: Vaizdas ----> dokumento informacija

Vilniaus universitetas

G. Kim et al. OCR-free Document Understanding Transformer. https://doi.org/10.48550/arXiv.2111.15664

Multimodalinių daugiakalbių DKM tikslumas ištraukiant duomenis

Multimodal offline LLMs

- *MiniCPM-Llama3-V* - the newest model in the MiniCPM-V series of small and mid-sized models designed for vision-language understanding
- *QWen-VL-Chat* - the visual multimodal version of the large model series, Qwen, proposed by Alibaba Cloud
- *TextMonkey* - TextMonkey, a large multimodal model (LMM) tailored for text-centric tasks

Data field	MiniCPM	QWen-VL	TextMonkey
doctype	90%	74%	86%
date	76%	44%	62%
issuer	68%	32%	56%
receiver	22%	10%	22%
address	2%	2%	2%

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- **Apmokytas ir išbandytas klasikinis CRF klasifikatorius iš Stanford NLP NER paketo.**
- **Apmokytas ir išbandytas klasikinis BiLSTM-CRF iš Stanza paketo, pasiekti geresni rezultatai nei vien CRF klasifikatoriaus naudojimo.**
- **Apmokytas ir išbandytas Microsoft modelis LayoutXLM, pasiekti geresni rezultatai nei klasikinių modelių.**
- **Palyginti tarpusavyje daugiakalbiai multimodaliniai parsisiunčiami didieji kalbos modeliai MiniCPM-V, Qwen-VL ir TextMonkey.**
- **Atsižvelgus į recenzentų pastabas ištaisytas ir įteiktas antram recenzavimo ciklui straipsnis Web of Science reitinguojamame žurnale *IEEE Access*. R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models.**

Kito pusmečio darbo planas

1. **Tęsiamas empirinis tyrimas**
    - **Modifikuotų algoritmų eksperimentinis tyrimas analizuojant jų efektyvumą.**
2. **Gautų rezultatų analizė ir apibendrinimas.**
3. **Mokslinių tyrimų disertacijos tema analitinės apžvalgos pildymas naujai atsirandančiais straipsniais**
4. **Parengti publikaciją apie duomenų ištraukimo (Key Information Extraction) sprendimą Web of Science reitinguojamame leidinyje.**

Kito pusmečio darbo planas

1. **Tęsiamas empirinis tyrimas**
    - **Modifikuotų algoritmų eksperimentinis tyrimas analizuojant jų efektyvumą.**
2. **Gautų rezultatų analizė ir apibendrinimas.**
3. **Mokslinių tyrimų disertacijos tema analitinės apžvalgos pildymas naujai atsirandančiais straipsniais**
4. **Parengti publikaciją apie duomenų ištraukimo (Key Information Extraction) sprendimą Web of Science reitinguojamame leidinyje.**

Vilniaus universitetas

Ačiū už dėmesį

Rolandas Gricius
VU DMSTI doktorantas
rolandas.gricius@mif.stud.vu.lt