Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej

...

Tomasz Stanisławek, 28 marzec 2022 r.
Seminarium IPI PAN

Krótko o mnie

I. Doktorat wdrożeniowy w firmie Applica.ai (lata 2017-2022)

1. Temat: Ekstrakcja informacji z dokumentów o bogatej strukturze graficznej
2. Promotor: dr hab. inż. Przemysław Biecek (laboratorium MI2)
3. Osiągnięcia: sześć opublikowanych publikacji naukowych powiązanych z tematem mojej rozprawy doktorskiej

II. Doświadczenie zawodowe

1. ~ 8 lat w firmie Applica.ai
2. ~ 1 rok w firmie WASAT
3. ~ 3 lata w Ośrodku Przetwarzania Informacji
4. ~ 3 lata w firmie Open Finance

Agenda

1. Wprowadzenie do problemu
2. Zarys historyczny
3. Kamienie milowe
    a. Analiza problemu
    b. Przygotowanie zbiorów danych do weryfikacji jakości modeli
    c. Nowy model do ekstrakcji informacji
    d. Benchmark dla dziedziny rozumienia dokumentów
4. Kierunki dalszego rozwoju dziedziny

Wprowadzenie do problemu - przykład 1

Urodził się 29 września 1881 w domu nr 13 przy ulicy Jagiellońskiej we Lwowie, w żydowskiej rodzinie Artura Edlera i Adeli (z domu Landau) von Misesów. Ojciec Ludwiga był absolwentem Politechniki w Zurychu. Potem pracował w austriackim Ministerstwie Kolei jako inżynier. Ludwig był najstarszym z trójki chłopców. Jeden z braci umarł w dzieciństwie, natomiast Richard został znanym matematykiem. W latach 1892–1900 Ludwig von Mises uczęszczał do prywatnej szkoły podstawowej we Lwowie.

Ekstrakcja informacji

Imie i Nazwisko	Data urodzin
Ludwig von Mises	29-09-1881
Adam Smith	16-06-1723
...	...

Ekstrakcja informacji

Ludwig Heinrich Edler von Mises

Data i miejsce urodzenia	29 września 1881 Lwów
Data i miejsce śmierci	10 października 1973 Nowy Jork
Miejsce spoczynku	Ferncliff Cemetery, Hrabstwo Westchester
Zawód, zajęcie	ekonomista, filozof

Przypadek 1: Zwykły tekst | **Baza danych** | **Przypadek 2: Dokument o bogatej strukturze graficznej**

https://pl.wikipedia.org/wiki/Ludwig_von_Mises

Wprowadzenie do problemu - inne przykłady

BLUE MEDIA S.A.

ul. Powstańców Warszawy 6
81-718 Sopot
tel. +48 58 760 48 22
fax +48 58 5551 318

B L U E
M E D I A

Faktura nr ;3/2021/PAYBM_PRE

Data wystawienia: **31-03-2021**
Data sprzedaży: **31-03-2021**
Miejsce wystawienia: **Sopot**

Sprzedawca:
Blue Media Spółka Akcyjna
ul. Powstańców Warszawy 6
81-718 Sopot
NIP: PL5851351185

Nabywca:

NIP:

Lp	Nazwa towaru lub usługi	PKWiU	Jm	Ilość	Cena netto	Stawka VAT	Wartość Netto	Wartość VAT	Wartość Brutto
1	Wynagrodzenie z tytułu świadczenia usług płatniczych dla serwisu https://www.s		usł.	1	27.57	zw	27.57	0.00	27.57

	Wartość Netto	Wartość VAT	Wartość Brutto
zw	27.57	0.00	27.57
Razem	27.57	0.00	27.57

Forma zapłaty: **zapłacono przelewem**
Termin zapłaty: **31-03-2021**
Uwagi: Usługi pośrednictwa w zakresie transakcji płatniczych zwolnione z podatku VAT na podstawie art.43 ust.1 pkt.40 ustawy o podatku od towarów i usług

Razem: **27.57 PLN**
Płatność otrzymana: **27.57 PLN**
Do zapłaty: **0.00 PLN**
Słownie do zapłaty: zero PLN 0/100

Krystyna Wawrzyniak

ratownik medyczny

Telefon 678678678

E-mail kwawrzyniak@email.com

Jestem ratownikiem medycznym z 5-letnim doświadczeniem, obecnie pracuję w Szpitalu XYZ. Odznaczam się życzliwym i empatycznym stosunkiem do pacjentów, a praca jest moją pasją. W Państwa firmie chciałbym rozwijać swoje kompetencje.

Doświadczenie

2016 - 2020 — **Ratownik medyczny**

Szpital XYZ

Zakres obowiązków:

- Prowadzenie podstawowej i zaawansowanej resuscytacji krążeniowo-oddechowej.
- Przywracanie drożności dróg oddechowych.
- Unieruchamianie złamań, zwichnięć, skręceń i kręgosłupa.
- Wykonywanie i ocena zapisu EKG.
- Monitorowanie czynności układu oddechowego.

2015 - 2016 — **Ratownik medyczny**

Przychodnia przy Szpitalu ABC

Zakres obowiązków:

- Podawanie leków drogą dożylną, domięśniową, podskórną, doustną, podjęzykową, wziewną.
- Opatrywanie ran i tamowanie krwawień zewnętrznych.
- Dbanie o bezpieczeństwo osób korzystających z kąpieliska.
- Stała obserwacja wyznaczonego obszaru wodnego.
- Kontrola stanu technicznego urządzeń i sprzętu ratowniczego.

Edukacja

2015 - 2016 — **Uniwersytet Jagielloński – Collegium Medicum**

Medyczne Centrum Kształcenia Podyplomowego
Medycyna ekstremalna i medycyna podróży

2012 - 2015 — **Niepubliczna Wyższa Szkoła Medyczna we Wrocławiu**

Ratownictwo medyczne

Umiejętności

- Praktyczna znajomość zasad korzystania z basenu ●●●●●
- Język angielski ●●●●○
- Umiejętność pracy z dziećmi ●●●●○
- Wysoka sprawność fizyczna ●●●●○
- Orientacja na pacjenta, empatia i wysoka kultura osobista ●●●●●

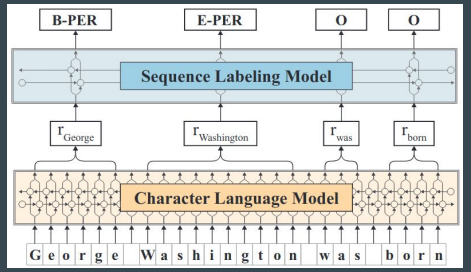
https://developers.bluemedia.pl/online/faq/faktury/czy-faktura-ktora-otrzymalem-jest-oplacona / https://interviewme.pl/blog-kategoria/cv-wzor

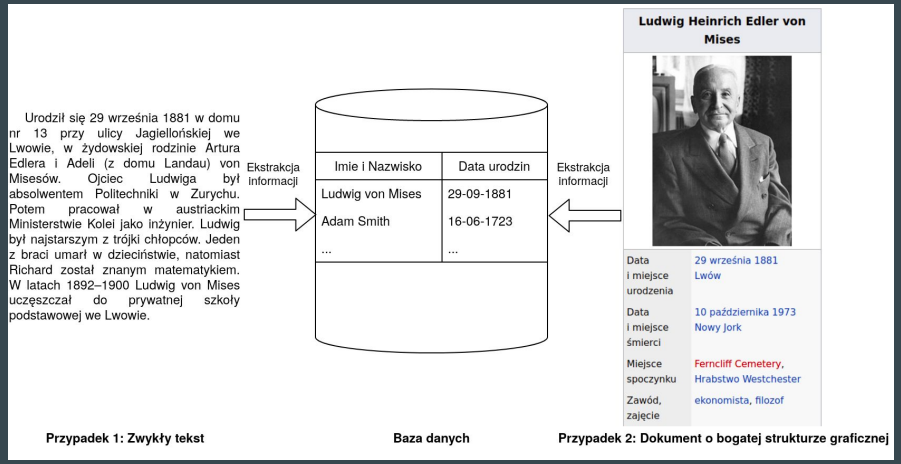
Wprowadzenie do problemu - znaczenie biznesowe

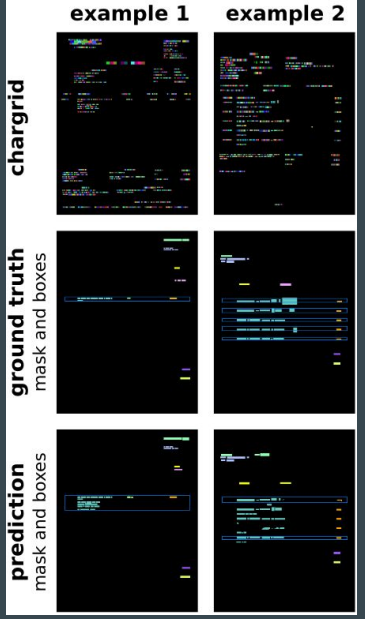
Ekstrakcja informacji jest bardzo ważnym aspektem zagadnienia robotyzacji procesów biznesowych (ang. Robotic Process Automation, RPA) związanych z przetwarzaniem dokumentów. Przewiduje się, że w 2022 roku cała branża związana z automatyzacją będzie warta ponad 600 miliardów dolarów (w 2020 roku warta była około 480 miliardów dolarów).

https://pl.freepik.com/premium-zdjecie/stos-dokumentow-papierowych-biurowych_6095110.htm
https://www.gartner.com/en/newsroom/press-releases/2021-04-28-gartner-forecasts-worldwide-hyperautomation-enabling-software-market-to-reach-nearly-600-billion-by-2022

Wprowadzenie do problemu - zadania







Przypadek 1: Zwykły tekst

Ekstrakcja informacji jako problem wykrywania jednostek nazewniczych (ang. *Named Entity Recognition, NER*), gdzie naszym zadaniem jest wykrycie nazw własnych (głównie osoby, organizacje i nazwy geograficzne).

W	latach	1892-1900	Ludwig	von	Misses	uczęszczał
O	O	O	I-PER	I-PER	I-PER	O

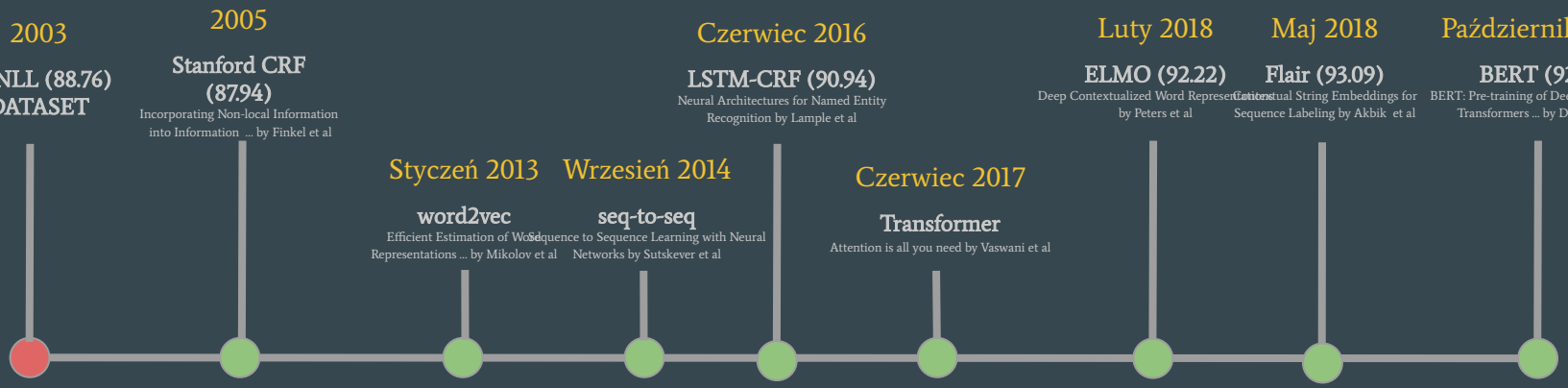
Przypadek 2: Dokument o bogatej strukturze graficznej

Ekstrakcja informacji jako problem ekstrakcji kluczowych informacji (ang. *Key Information Extraction, KIE*), gdzie naszym zadaniem jest wydobycie z dokumentu kluczowych pól.

W tym przypadku zazwyczaj mamy dostęp do anotacji na poziomie całego dokumentu, a nie na poziomie pojedynczego słowa/tokenu (jak to ma miejsce w przypadku zadania NER).

https://aclanthology.org/C18-1139.pdf / https://pl.wikipedia.org/wiki/Ludwig_von_Mises / https://arxiv.org/pdf/1809.08799.pdf

Zarys historyczny - ekstrakcja informacji z tekstu



Zarys historyczny - ekstrakcja informacji z dokumentów o bogatej strukturze graficznej

2003

CoNLL (88.76) DATASET

2005

Stanford CRF (87.94)

Incorporating Non-local Information into Information ... by Finkel et al

Styczeń 2013

word2vec

Efficient Estimation of Word Representations ... by Mikolov et al

Wrzesień 2014

seq-to-seq

Sequence to Sequence Learning with Neural Networks by Sutskever et al

Czerwiec 2016

LSTM-CRF (90.94)

Neural Architectures for Named Entity Recognition by Lample et al

Czerwiec 2017

Transformer

Attention is all you need by Vaswani et al

Luty 2018

ELMO (92.22)

Deep Contextualized Word Representations by Peters et al

Maj 2018

Flair (93.09)

Contextual String Embeddings for Sequence Labeling by Akbik et al

Październik 2018

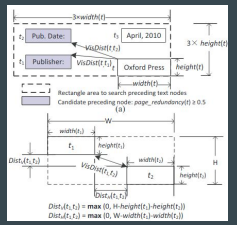
BERT (92.8)

BERT: Pre-training of Deep Bidirectional Transformers ... by Devlin et al

Czerwiec 2011

Web Data Extraction

From One Tree to a Forest: a Unified Solution for ... by Hao et al



Maj 2015

Digital Leafleting

Digital Leafleting: Extracting Structured Data from ... by Apostolova et al

Feature Name	Description
Font Size	The computed *font-size* attribute of the surrounding HTML DOM element, normalized to 7 basic sizes (*xx-small, x-small, small, medium, large, x-large, xx-large*).
Color	The computed *color* attribute of the surrounding HTML DOM element.

Sierpień 2017

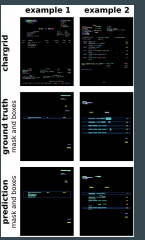
CloudScan

CloudScan - A configuration-free invoice analysis system ... by Palm et al

Wrzesień 2018

Chargrid

Chargrid: Towards Understanding 2D Documents by Katti et al



Zarys historyczny - początek prac nad doktoratem

2003

CoNLL (88.76) DATASET

2005

Stanford CRF (87.94)

Incorporating Non-local Information into Information ... by Finkel et al

2012

NKJP DATASET

Styczeń 2013

word2vec

Efficient Estimation of Word Representations ... by Mikolov et al

Wrzesień 2014

seq-to-seq

Sequence to Sequence Learning with Neural Networks by Sutskever et al

Czerwiec 2016

LSTM-CRF (90.94)

Neural Architectures for Named Entity Recognition by Lample et al

Czerwiec 2017

Transformer

Attention is all you need by Vaswani et al

Luty 2018

ELMO (92.22)

Deep Contextualized Word Representations by Peters et al

Maj 2018

Flair (93.09)

Contextual String Embeddings for Sequence Labeling by Akbik et al

Październik 2018

BERT (92.8)

BERT: Pre-training of Deep Bidirectional Transformers ... by Devlin et al

Wrzesień 2018

PolEval NER

Approaching nested named entity recognition ... by Borchmannet al

Wrzesień 2017

Początek doktoratu

Sierpień/Wrzesień 2018

Robotic Process Automation of Unstructured Data ... by Wróblewska et al

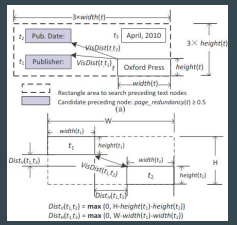
Listopad 2018

Zakończenie głównych prac nad system do automatyzacji procesów biznesowych

Czerwiec 2011

Web Data Extraction

From One Tree to a Forest: a Unified Solution for ... by Hao et al



Maj 2015

Digital Leafleting

Digital Leafleting: Extracting Structured Data from ... by Apostolova et al

Feature Name	Description
Font Size	The computed *font-size* attribute of the surrounding HTML DOM element, normalized to 7 basic sizes (*xx-small, x-small, small, medium, large, x-large, xx-large*).
Color	The computed *color* attribute of the surrounding HTML DOM element.

Sierpień 2017

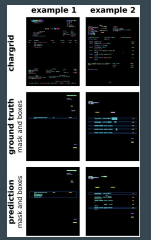
CloudScan

CloudScan - A configuration-free invoice analysis system ... by Palm et al

Wrzesień 2018

Chargrid

Chargrid: Towards Understanding 2D Documents by Katti et al



Zarys historyczny - pierwsza publikacja opisująca stworzony system

Robotic Process Automation of Unstructured Data with Machine Learning

Anna Wróblewska*,†, Tomasz Stanisławek*,†, Bartłomiej Prus-Zajączkowski*,†, Łukasz Garncarek†
*Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, Warszawa, Poland
†Applica.ai
ul. Wiślana 8, Warszawa, Poland
E-mail: {anna.wroblewska, tomasz.stanislawek, bartlomiej.prus, lukasz.garncarek}@applica.ai

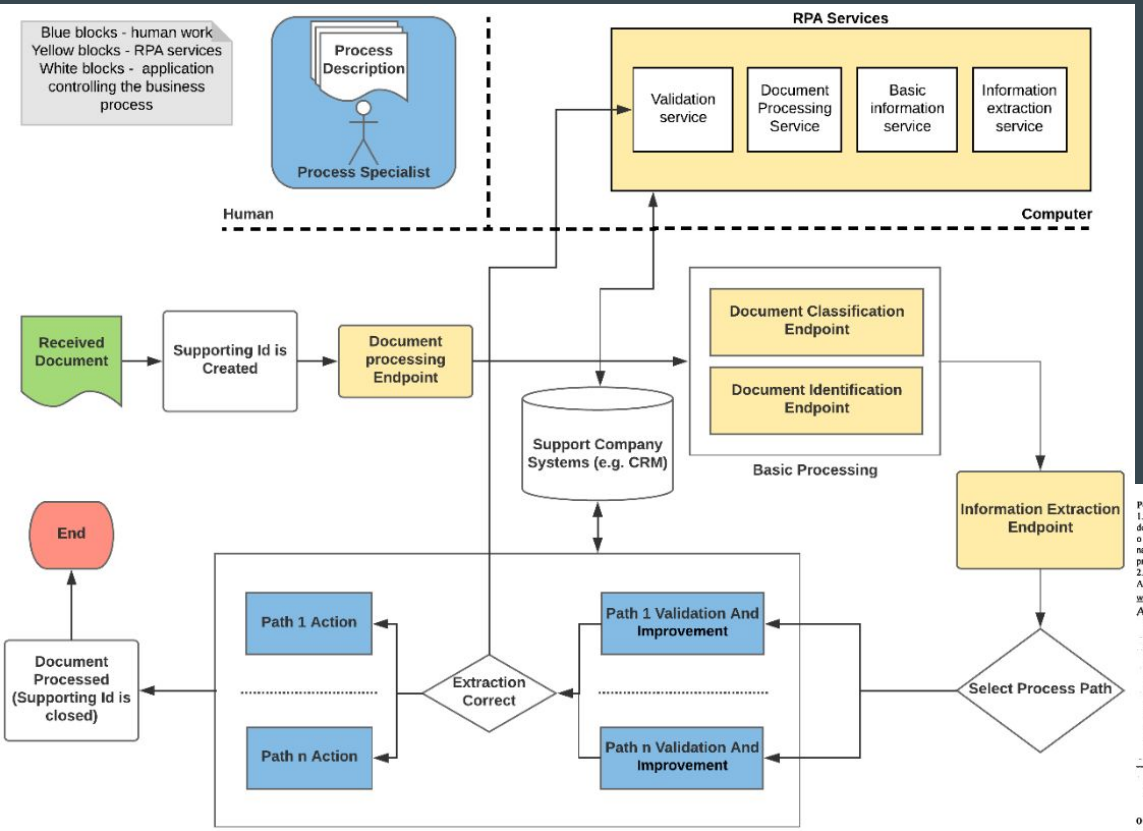
***Abstract*—In this paper we present our work in progress on building an artificial intelligence system dedicated to tasks regarding the processing of formal documents used in various kinds of business procedures. The main challenge is to build machine learning (ML) models to improve the quality and efficiency of business processes involving image processing, optical character recognition (OCR), text mining and information extraction. In the paper we introduce the research and application field, some common techniques used in this area and our preliminary results and conclusions.**

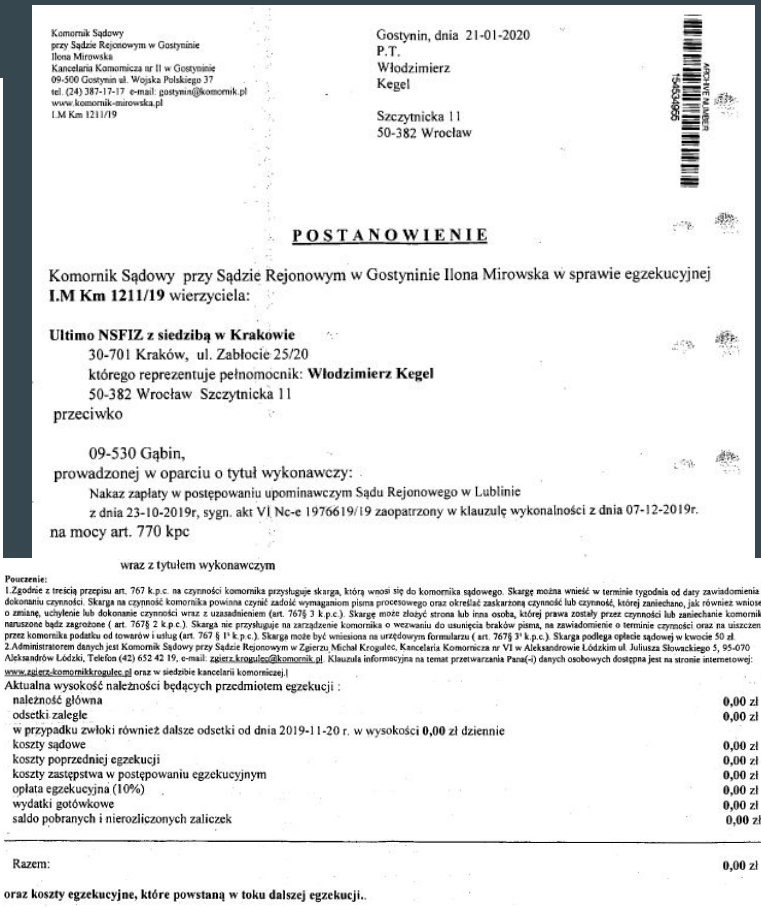
and sketch state-of-the-art ML methods that are used for particular parts of our system (section III). Then we show our approach to automation of processing formal documents, describing use cases and preliminary results (section IV). Finally, we conclude our work with future plans and important tips for RPA with ML methods (section V).

II. Robotic Process Automation of Unstructured Data

Robotic Process Automation of Unstructured Data with Machine Learning

Zarys historyczny





Komornik Sądowy
przy Sądzie Rejonowym w Gostyninie
Ilona Mirowska
Kancelaria Komornicza nr II w Gostyninie
09-500 Gostynin ul. Wojska Polskiego 37
tel. (24) 387-17-17 e-mail: gostynin@komornik.pl
www.komornik-mirowska.pl
I.M Km 1211/19

Gostynin, dnia 21-01-2020
P.T.
Włodzimierz
Kegel

Szczytnicka 11
50-382 Wrocław

POSTANOWIENIE

Komornik Sądowy przy Sądzie Rejonowym w Gostyninie Ilona Mirowska w sprawie egzekucyjnej **I.M Km 1211/19** wierzyciela:

Ultimo NSFIZ z siedzibą w Krakowie
30-701 Kraków, ul. Zabłocie 25/20
którego reprezentuje pełnomocnik: **Włodzimierz Kegel**
50-382 Wrocław Szczytnicka 11
przeciwko

09-530 Gąbin,
prowadzonej w oparciu o tytuł wykonawczy:
Nakaz zapłaty w postępowaniu upominawczym Sądu Rejonowego w Lublinie
z dnia 23-10-2019r, sygn. akt VI Nc-e 1976619/19 zaopatrzony w klauzulę wykonalności z dnia 07-12-2019r.
na mocy art. 770 kpc

wraz z tytułem wykonawczym

Pouczenie:
1.Zgodnie z treścią przepisu art. 767 k.p.c. na czynności komornika przysługuje skarga, którą wnosi się do komornika sądowego. Skargę można wnieść w terminie tygodnia od daty zawiadomienia o dokonaniu czynności. Skarga na czynność komornika powinna czynić zadość wymaganiom pisma procesowego oraz określać zaskarżoną czynność lub czynność, której zaniechano, jak również wniosek o zmianę, uchylenie lub dokonanie czynności wraz z uzasadnieniem (art. 767§ 3 k.p.c.). Skargę może złożyć strona lub inna osoba, której prawa zostały przez czynności lub zaniechanie komornika naruszone bądź zagrożone (art. 767§ 2 k.p.c.). Skarga nie przysługuje na zarządzenie komornika o wezwaniu do usunięcia braków pisma, na zawiadomienie o terminie czynności oraz na uiszczenie przez komornika podatku od towarów i usług (art. 767 § 1ª k.p.c.). Skarga może być wniesiona na urzędowym formularzu (art. 767§ 3ª k.p.c.). Skarga podlega opłacie sądowej w kwocie 50 zł.
2.Administratorem danych jest Komornik Sądowy przy Sądzie Rejonowym w Zgierzu Michał Krogulec, Kancelaria Komornicza nr VI w Aleksandrowie Łódzkim ul. Juliusza Słowackiego 5, 95-070 Aleksandrów Łódzki, Telefon (42) 652 42 19, e-mail: zgierz.krogulec@komornik.pl. Klauzula informacyjna na temat przetwarzania Pana(-i) danych osobowych dostępna jest na stronie internetowej: www.zgierz-komornikkrogulec.pl oraz w siedzibie kancelarii komorniczej.)

Aktualna wysokość należności będących przedmiotem egzekucji :

należność główna	**0,00 zł**
odsetki zaległe	**0,00 zł**
w przypadku zwłoki również dalsze odsetki od dnia 2019-11-20 r. w wysokości **0,00** zł dziennie	
koszty sądowe	**0,00 zł**
koszty poprzedniej egzekucji	**0,00 zł**
koszty zastępstwa w postępowaniu egzekucyjnym	**0,00 zł**
opłata egzekucyjna (10%)	**0,00 zł**
wydatki gotówkowe	**0,00 zł**
saldo pobranych i nierozliczonych zaliczek	**0,00 zł**
Razem:	**0,00 zł**

oraz koszty egzekucyjne, które powstaną w toku dalszej egzekucji..

Robotic Process Automation of Unstructured Data with Machine Learning

Analiza problemu

Named Entity Recognition - Is there a glass ceiling?

Tomasz Stanislawek[†,‡], **Anna Wróblewska**[†,‡], **Alicja Wójcika**[†,§], **Daniel Ziembicki**[†,§] **Przemyslaw Biecek**[‡,¶]

[†]Applica.ai, Warsaw, Poland
[¶]Samsung Research Poland, Warsaw, Poland
[‡]Faculty of Mathematics and Information Science, Warsaw University of Technology
[§]Department of Formal Linguistics, University of Warsaw

Abstract

Recent developments in Named Entity Recognition (NER) have resulted in better and better models. However, is there a glass ceiling? Do we know which types of errors are still hard or even impossible to correct? In this paper, we present a detailed analysis of the types of errors in state-of-the-art machine learning (ML) methods. Our study reveals the weak and strong points of the Stanford, CMU, FLAIR, ELMO and BERT models, as well as their shared limitations. We also introduce new techniques for improving annotation, for training processes and for checking a model's quality and stability.

Presented results are based on the CoNLL 2003 data set for the English language. A new enriched semantic annotation of errors for this data set and new diagnostic data sets are attached in the supplementary materials.

Of course, different models make different mistakes. Here, we have focused on models that constitute a kind of breakthrough in the NER domain. These models are: Stanford NER (Finkel et al., 2005), the model made by the NLP team from Carnegie Mellon University (CMU) (Lample et al., 2016), ELMO (Peters et al., 2018), FLAIR (Akbik et al., 2018) and BERT-Base (Devlin et al., 2018). In the Stanford model, Conditional Random Fields (CRF) with manually created features were tackled. Lample and the team (at CMU) used an LSTM deep neural network with an output with CRF for the first time. ELMO and FLAIR are new language modeling techniques as an encoder, and LSTM with a CRF layer as an output decoder. A team from Google used a fine-tuning approach with the BERT model in a NER problem for the first time, based on a Bi-diREctional Transformer language model (LM).

https://aclanthology.org/K19-1058.pdf

Analiza problemu

Błędy zbioru danych (DE)	Zależności na poziomie dokumentu (DL)
Błędy anotacji (DE-A) - oczywiste błędy anotacji	**Koreferencja na poziomie dokumentu (DL-CR)** - przypadki występowania obiektu w zdaniu, który występuje również w innym zdaniu w tym samym dokumencie
Literówki słowne (DE-WT) - literówki słowne w jednostce nazwanej	**Struktura dokumentu (DL-S)** - struktura dokumentu odgrywa znaczącą rolę, np. obiekty występują w tabelce
Zła segmentacja na poziomie słowa/zdania (DE-BS) - przypadki, gdzie słowo lub zdanie zostało źle podzielone na segmenty	**Kontekst dokumentu (DL-C)** - przypadki, gdzie potrzebny jest cały kontekst dokumentu, aby prawidłowo zaanotować jednostkę nazwaną
Zależności na poziomie zdania (SL)	**Ogólne właściwości (G)**
Struktura zdania (SL-S) - syntaktyczne właściwości lingwistyczne stanowiące mocną wskazówkę o danym obiekcie	**Dwuznaczność (G-A)** - przypadki, w których jednostki nazwane są użyte w innym znaczeniu, niż zazwyczaj stosowany
Kontekst zdania (SL-C) - przypadki, gdzie kontekst zdania jest wystarczający	**Niespójność anotacji (G-I)** - różna anotacja na poziomie zbioru dla tych samych jednostek nazwanych
	Trudne przypadki (G-HC) - różna możliwość interpretacji

Tablica 1: Taksonomia z kategoriami lingwistycznymi, przedstawiająca najbardziej prawdopodobne przyczyny błędów

Przykłady

1. **DE-A**: w zdaniu ”SOCCER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT” słowo ”CHINA” zostało zaanotowane jako typ PERSON

2. **DL-S**: spójrzmy na następujące trzy zdania, które w oczywisty sposób tworzą tabelę:
    a. ”Port Loading Waiting”
    b. ”Vancouver 5 7”
    c. ”Prince Rupert 1 3”

Niektóre modele mają problem z wykrywaniem wartości w tabelach przez co oznaczają nagłówki jako jednostki nazwane

https://aclanthology.org/K19-1058.pdf

Analiza problemu

Model	DE-WT	DE-BS	SL-S	SL-C	DL-CR	DL-S	DL-C	G-A	G-HC	G-I	Ogółem
Stanford	10	38	46	448	372	202	247	219	72	19	703
CMU	6	39	21	378	316	107	175	183	68	20	554
ELMO	9	33	13	250	198	97	144	98	65	21	395
Flair	8	33	16	223	184	100	146	101	59	20	370
BERT	10	40	11	300	263	117	170	94	65	20	472

Tablica 2: Liczba błędów dla wybranego modelu i kategorii lingwistycznej.

Najważniejsze wnioski

1. Modele języka rozwiązują bardzo dużo problemów związanych z niejednoznacznością (G-A)
2. Modele języka o prawie połowę zmniejszyły liczbę błędów w przypadkach, gdzie kontekst zdania jest wystarczający (SL-C)
3. Istnieją kategorie, w których nie ma znaczącej poprawy: DE-WT, DE-BS, G-HC, G-I. Związane jest to z błędami podczas tworzenia zbioru (zła segmentacja, brak spójności anotacji) oraz generalną złożonością języka naturalnego (czasem trudno jest określić, w jaki sposób zaanotować dane słowo)
4. Wszystkie przebadane rozwiązania w tamtym okresie bazowały wyłącznie na kontekście zdania. Natomiast wysoka liczba błędów popełnianych przez modele na poziomie zależności dokumentu (DL-CR, DL-C, DL-S) podpowiada, że żeby osiągnąć jeszcze lepsze wyniki, musimy budować rozwiązania bazujące na całym kontekście dokumentu.
5. Wysoka liczba błędów w kategorii związanej ze strukturą dokumentu (DL-S) informuje nas o kolejnym potencjalnym kierunku rozwoju dziedziny, jakim jest uwzględnienie tychże informacji (np. pozycji słów na stronie).

https://aclanthology.org/K19-1058.pdf

Ciekawostka - Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks

Dataset	Modality	Size	Model	Test Set Errors				
				CL guessed	MTurk checked	validated	estimated	% error
MNIST	image	10,000	2-conv CNN	100	100 (100%)	15	-	0.15
CIFAR-10	image	10,000	VGG	275	275 (100%)	54	-	0.54
CIFAR-100	image	10,000	VGG	2,235	2,235 (100%)	585	-	5.85
Caltech-256†	image	29,780	Wide ResNet-50-2	2,360	2,360 (100%)	458	-	1.54
ImageNet*	image	50,000	ResNet-50	5,440	5,440 (100%)	2,916	-	5.83
QuickDraw†	image	50,426,266	VGG	6,825,383	2,500 (0.04%)	1870	5,105,386	10.12
20news	text	7,532	TFIDF + SGD	93	93 (100%)	82	-	1.09
IMDB	text	25,000	FastText	1,310	1,310 (100%)	725	-	2.90
Amazon Reviews†	text	9,996,437	FastText	533,249	1,000 (0.2%)	732	390,338	3.90
AudioSet	audio	20,371	VGG	307	307 (100%)	275	-	1.35

*Because the ImageNet test set labels are not publicly available, the ILSVRC 2012 validation set is used.

†Because no explicit test set is provided, we study the entire dataset to ensure coverage of any train/test split.

https://arxiv.org/pdf/2103.14749.pdf / https://labelerrors.com/

Przygotowanie zbiorów danych (Kleister)

Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts

Tomasz Stanisławek[1,2], Filip Graliński[1,3], Anna Wróblewska[2], Dawid Lipiński[1], Agnieszka Kaliska[1,3], Paulina Rosalska[1], Bartosz Topolski[1], and Przemysław Biecek[2,4]

[1] Applica.ai, 15 Zajęcza, Warsaw, 00351 firstname.lastname@applica.ai
[2] Warsaw University of Technology, Koszykowa 75, Warsaw, Poland
firstname.lastname@pw.edu.pl
[3] Adam Mickiewicz University, 1 Wieniawskiego, Poznan, 61712, Poland
firstname.lastname@amu.edu.pl
[4] Samsung R&D Institute Poland, Plac Europejski 1, Warsaw, Poland
firstnameletter.lastname@samsung.com

Abstract. The relevance of the Key Information Extraction (KIE) task is increasingly important in natural language processing problems. But there are still only a few well-defined problems that serve as benchmarks for solutions in this area. To bridge this gap, we introduce two new datasets (*Kleister NDA* and *Kleister Charity*). They involve a mix of scanned and born-digital long formal English-language documents. In these datasets, an NLP system is expected to find or infer various types of entities by employing both textual and structural layout features. The Kleister Charity dataset consists of 2,788 annual financial reports of charity organizations, with 61,643 unique pages and 21,612 entities to extract. The Kleister NDA dataset has 540 Non-disclosure Agreements, with 3,229 unique pages and 2,160 entities to extract. We provide several state-of-the-art baseline systems from the KIE domain (Flair, BERT, RoBERTa, LayoutLM, LAMBERT), which show that our datasets pose a strong challenge to existing models. The best model achieved an 81.77 % and an 83.57 % F1-score on respectively the Kleister NDA and the Kleister Charity datasets. We share the datasets to encourage progress on more in-depth and complex information extraction tasks.

https://arxiv.org/pdf/2105.05796.pdf

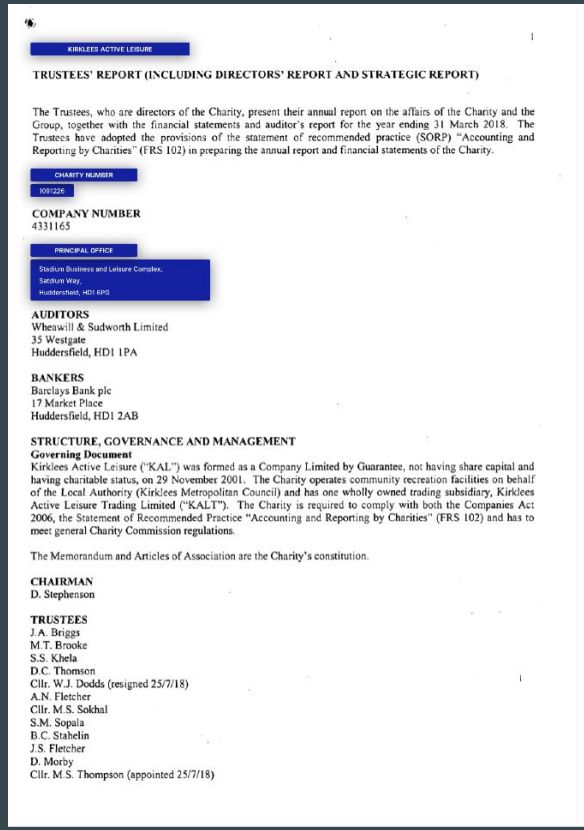
Przygotowanie zbiorów danych (Kleister) - porównanie

Dataset name	CoNLL 2003	WikiReading	FUNSD	SROIE	Kleister NDA	Kleister Charity
Source	Reuters news	Wikipedia	forms	receipts	EDGAR	UK Charity Com.
Documents	1,393	4.7M	199	973	540	2,778
Pages	—	—	199	973	3,229	61,643
Entities	35,089	18M	9,743	3,892	2,160	21,612
train docs	946	16.03M	149	626	254	1,729
dev docs	216	1.89M	—	—	83	440
test docs	231	0.95M	50	347	203	609
Input/Output on token level(*)	✓	✗	✓	✓	✗	✗
Long Document(*)	✗	✓	✗	✗	✓	✓
Complex layout(*)	✗	✗	✓	✓	✓	✓
OCR(*)	✗	✗	✓	✓	✗	✓

Table 1. Summary of the existing English datasets and the Kleister sets. (*) For detailed description see Section 3.3.

https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - przykłady



KIRKLEES ACTIVE LEISURE

TRUSTEES' REPORT (INCLUDING DIRECTORS' REPORT AND STRATEGIC REPORT)

The Trustees, who are directors of the Charity, present their annual report on the affairs of the Charity and the Group, together with the financial statements and auditor's report for the year ending 31 March 2018. The Trustees have adopted the provisions of the statement of recommended practice (SORP) "Accounting and Reporting by Charities" (FRS 102) in preparing the annual report and financial statements of the Charity.

CHARITY NUMBER
1091226

COMPANY NUMBER
4331165

PRINCIPAL OFFICE
Stadium Business and Leisure Complex,
Satdium Way,
Huddersfield, HD1 6PG

AUDITORS
Whenwill & Sudworth Limited
35 Westgate
Huddersfield, HD1 1PA

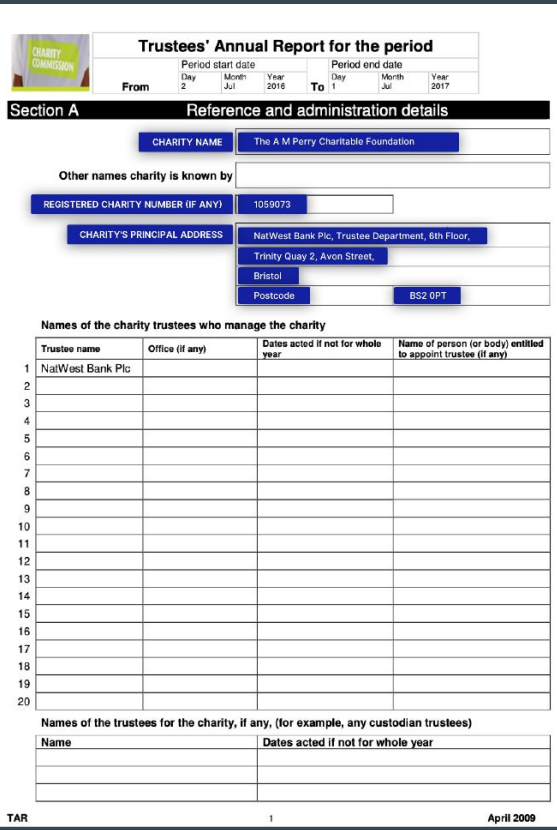
BANKERS
Barclays Bank plc
17 Market Place
Huddersfield, HD1 2AB

STRUCTURE, GOVERNANCE AND MANAGEMENT
Governing Document
Kirklees Active Leisure ("KAL") was formed as a Company Limited by Guarantee, not having share capital and having charitable status, on 29 November 2001. The Charity operates community recreation facilities on behalf of the Local Authority (Kirklees Metropolitan Council) and has one wholly owned trading subsidiary, Kirklees Active Leisure Trading Limited ("KALT"). The Charity is required to comply with both the Companies Act 2006, the Statement of Recommended Practice "Accounting and Reporting by Charities" (FRS 102) and has to meet general Charity Commission regulations.

The Memorandum and Articles of Association are the Charity's constitution.

CHAIRMAN
D. Stephenson

TRUSTEES
J.A. Briggs
M.T. Brooke
S.S. Khela
D.C. Thomson
Cllr. W.J. Dodds (resigned 25/7/18)
A.N. Fletcher
Cllr. M.S. Sokhal
S.M. Sopala
B.C. Stahelin
J.S. Fletcher
D. Morby
Cllr. M.S. Thompson (appointed 25/7/18)



CHARITY COMMISSION

Trustees' Annual Report for the period

	Period start date			Period end date		
	Day	Month	Year	Day	Month	Year
From	2	Jul	2016	To 1	Jul	2017

Section A Reference and administration details

CHARITY NAME: The A M Perry Charitable Foundation

Other names charity is known by:

REGISTERED CHARITY NUMBER (IF ANY): 1059073

CHARITY'S PRINCIPAL ADDRESS: NatWest Bank Plc, Trustee Department, 6th Floor, Trinity Quay 2, Avon Street, Bristol, Postcode BS2 0PT

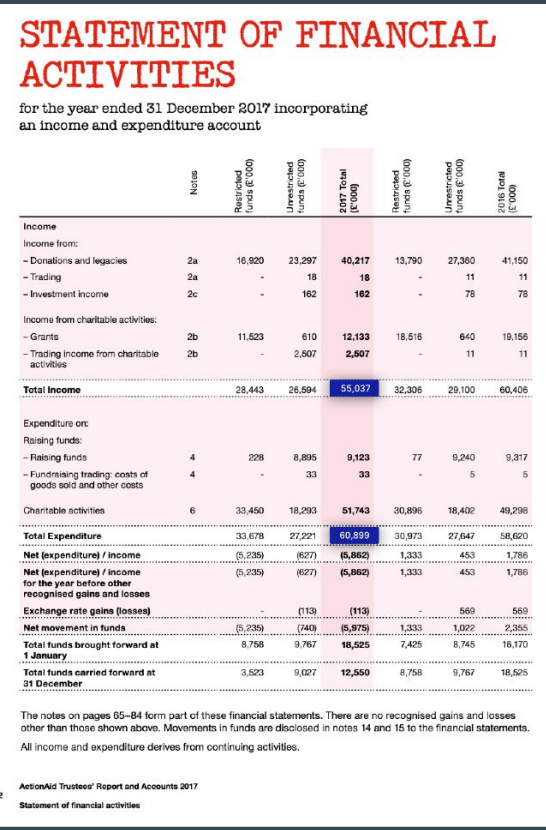
Names of the charity trustees who manage the charity

	Trustee name	Office (if any)	Dates acted if not for whole year	Name of person (or body) entitled to appoint trustee (if any)
1	NatWest Bank Plc			
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Names of the trustees for the charity, if any, (for example, any custodian trustees)

Name	Dates acted if not for whole year

TAR 1 April 2009



STATEMENT OF FINANCIAL ACTIVITIES

for the year ended 31 December 2017 incorporating an income and expenditure account

	Notes	Restricted funds (£'000)	Unrestricted funds (£'000)	2017 Total (£'000)	Restricted funds (£'000)	Unrestricted funds (£'000)	2016 Total (£'000)
Income							
Income from:							
– Donations and legacies	2a	16,920	23,297	**40,217**	13,790	27,360	41,150
– Trading	2a	-	18	**18**	-	11	11
– Investment income	2c	-	162	**162**	-	78	78
Income from charitable activities:							
– Grants	2b	11,523	610	**12,133**	18,516	640	19,156
– Trading income from charitable activities	2b	-	2,507	**2,507**	-	11	11
Total income		28,443	26,594	55,037	32,306	29,100	60,406
Expenditure on:							
Raising funds:							
– Raising funds	4	228	8,895	**9,123**	77	9,240	9,317
– Fundraising trading: costs of goods sold and other costs	4	-	33	**33**	-	5	5
Charitable activities	6	33,450	18,293	**51,743**	30,896	18,402	49,298
Total Expenditure		33,678	27,221	60,899	30,973	27,647	58,620
Net (expenditure) / income		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Net (expenditure) / income for the year before other recognised gains and losses		(5,235)	(627)	**(5,862)**	1,333	453	1,786
Exchange rate gains (losses)		-	(113)	**(113)**	-	569	569
Net movement in funds		(5,235)	(740)	**(5,975)**	1,333	1,022	2,355
Total funds brought forward at 1 January		8,758	9,767	**18,525**	7,425	8,745	16,170
Total funds carried forward at 31 December		3,523	9,027	**12,550**	8,758	9,767	18,525

The notes on pages 65–84 form part of these financial statements. There are no recognised gains and losses other than those shown above. Movements in funds are disclosed in notes 14 and 15 to the financial statements.

All income and expenditure derives from continuing activities.

62 ActionAid Trustees' Report and Accounts 2017
Statement of financial activities

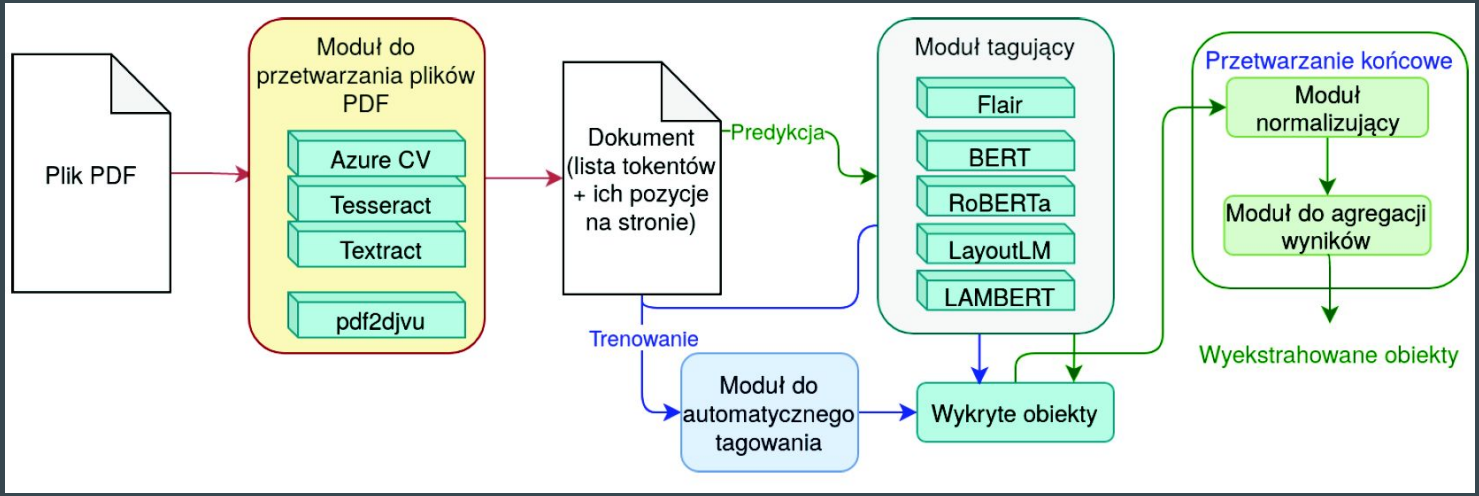
https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - wybrane statystyki

Zbiór	Nazwa obiektu	Typ obiektu	Liczba	Liczba unikalnych	Przykład obiektu
NDA	party	OGRANIZACJA/OSOBA	1,035	912	Ajinomoto Althea Inc.
	jurisdiction	LOKALIZACJA	531	37	New York
	effective_date	DATA	400	370	2005-07-03
	term	CZAS TRWANIA	194	22	P12M
Charity	post_town	ADRES	2,692	501	BURY
	postcode	ADRES	2,717	1,511	BL9 ONP
	street_line	ADRES	2,414	1,353	42-47 MINORIES
	charity_name	ORGANIZACJA	2,778	1,600	Mad Theatre Company
	charity_number	NUMER	2,763	1,514	1143209
	report_date	DATA	2,776	129	2016-09-30
	income	KWOTA	2,741	2,726	109370.00
	spending	KWOTA	2,731	2,712	90174.00

https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - wnioski



Anotacja na poziomie dokumentu

W związku z brakiem informacji o pozycji obiektów w tekście należy budować dodatkowe moduły wspomagające pracę tagerów sekwencyjnych lub przygotować rozwiązania działające w oparciu o model typu seq2seq.

https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - wnioski

Zbiór	Nazwa narzędzia	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**	Człowiek
NDA	Azure CV	$78.03_{\pm 0.12}$	$77.67_{\pm 0.18}$	$79.33_{\pm 0.68}$	$77.43_{\pm 0.29}$	$80.57_{\pm 0.25}$	97.86%
	pdf2djvu	$77.83_{\pm 0.26}$	$78.20_{\pm 0.17}$	$81.00_{\pm 0.05}$	$78.47_{\pm 0.76}$	$\mathbf{81.77_{\pm 0.09}}$	
	Tesseract	$76.57_{\pm 0.49}$	$76.60_{\pm 0.30}$	$77.81_{\pm 0.97}$	$77.70_{\pm 0.48}$	$81.03_{\pm 0.23}$	
	Textract	$77.37_{\pm 0.08}$	$74.83_{\pm 0.45}$	$79.49_{\pm 0.32}$	$77.40_{\pm 0.40}$	$77.37_{\pm 0.08}$	
Charity (*)	Azure CV	$81.17_{\pm 0.12}$	$78.33_{\pm 0.08}$	$81.50_{\pm 0.23}$	$81.53_{\pm 0.23}$	$\mathbf{83.57_{\pm 0.29}}$	97.45%
	Tesseract	$72.87_{\pm 0.81}$	$71.37_{\pm 1.25}$	$76.23_{\pm 0.15}$	$77.53_{\pm 0.20}$	$81.50_{\pm 0.07}$	
	Textract	$78.03_{\pm 0.12}$	$73.30_{\pm 0.43}$	$80.08_{\pm 0.15}$	$80.23_{\pm 0.41}$	$82.97_{\pm 0.21}$	

Bogata struktura graficzna

Modele działające wyłącznie w oparciu o tekst (Flair, BERT oraz RoBERTa) działają wyraźnie słabiej od modeli uwzględniających również informacje o strukturze dokumentu (LayaoutLM bazujący na modelu BERT oraz LAMBERT bazujący na modelu RoBERTa). Szczególnie widoczne jest to dla obiektów, które znajdują się w tabelkach lub formularzach.

https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - wnioski

Zbiór	Nazwa narzędzia	**Flair**	**BERT**	**RoBERTa**	**LayoutLM**	**LAMBERT**	Człowiek
NDA	Azure CV	$78.03_{\pm 0.12}$	$77.67_{\pm 0.18}$	$79.33_{\pm 0.68}$	$77.43_{\pm 0.29}$	$80.57_{\pm 0.25}$	97.86%
	pdf2djvu	$77.83_{\pm 0.26}$	$78.20_{\pm 0.17}$	$81.00_{\pm 0.05}$	$78.47_{\pm 0.76}$	$\mathbf{81.77_{\pm 0.09}}$	
	Tesseract	$76.57_{\pm 0.49}$	$76.60_{\pm 0.30}$	$77.81_{\pm 0.97}$	$77.70_{\pm 0.48}$	$81.03_{\pm 0.23}$	
	Textract	$77.37_{\pm 0.08}$	$74.83_{\pm 0.45}$	$79.49_{\pm 0.32}$	$77.40_{\pm 0.40}$	$77.37_{\pm 0.08}$	
Charity (*)	Azure CV	$81.17_{\pm 0.12}$	$78.33_{\pm 0.08}$	$81.50_{\pm 0.23}$	$81.53_{\pm 0.23}$	$\mathbf{83.57_{\pm 0.29}}$	97.45%
	Tesseract	$72.87_{\pm 0.81}$	$71.37_{\pm 1.25}$	$76.23_{\pm 0.15}$	$77.53_{\pm 0.20}$	$81.50_{\pm 0.07}$	
	Textract	$78.03_{\pm 0.12}$	$73.30_{\pm 0.43}$	$80.08_{\pm 0.15}$	$80.23_{\pm 0.41}$	$82.97_{\pm 0.21}$	

Optyczne rozpoznawanie znaków

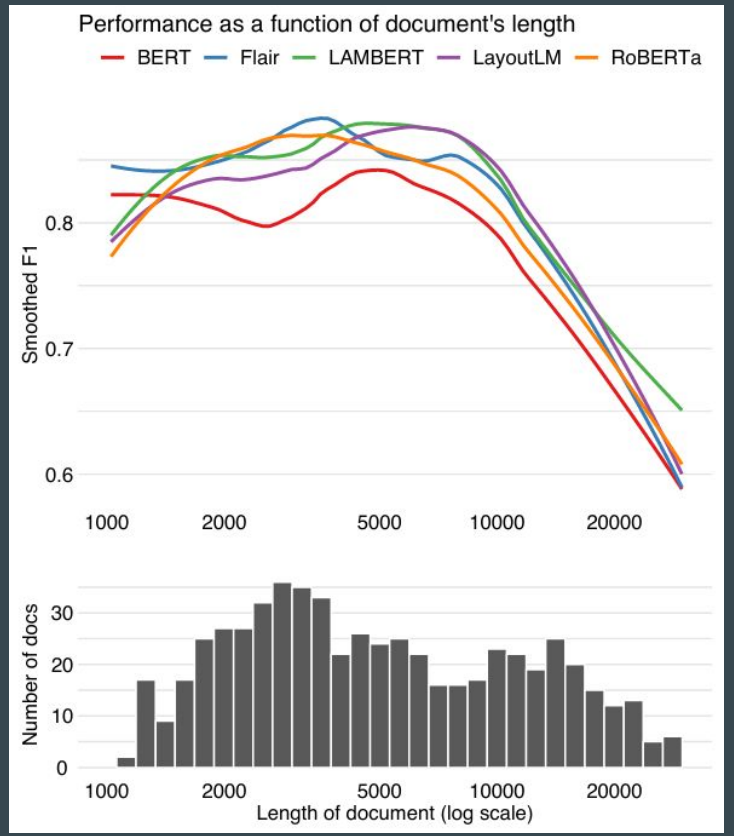
Wybór narzędzia do przetwarzania plików PDF był bardzo istotny w kontekście jakości ekstrakcji. Okazuje się, że w niektórych przypadkach był on bardziej istotny niż sam wybór modelu. Najlepszym narzędziem z przebadanych okazał się Azure CV.

https://arxiv.org/pdf/2105.05796.pdf

Przygotowanie zbiorów danych (Kleister) - wnioski

Długie dokumenty

Na bardzo długich dokumentach zaobserwowano znaczące pogorszenie się skuteczności modelu.



https://arxiv.org/pdf/2105.05796.pdf

LAMBERT - nowy model języka

LAMBERT: Layout-Aware Language Modeling for Information Extraction

Łukasz Garncarek[1⋆], Rafał Powalski[1], Tomasz Stanisławek[1,2], Bartosz Topolski[1], Piotr Halama[1], Michał Turski[1,3], and Filip Graliński[1,3]

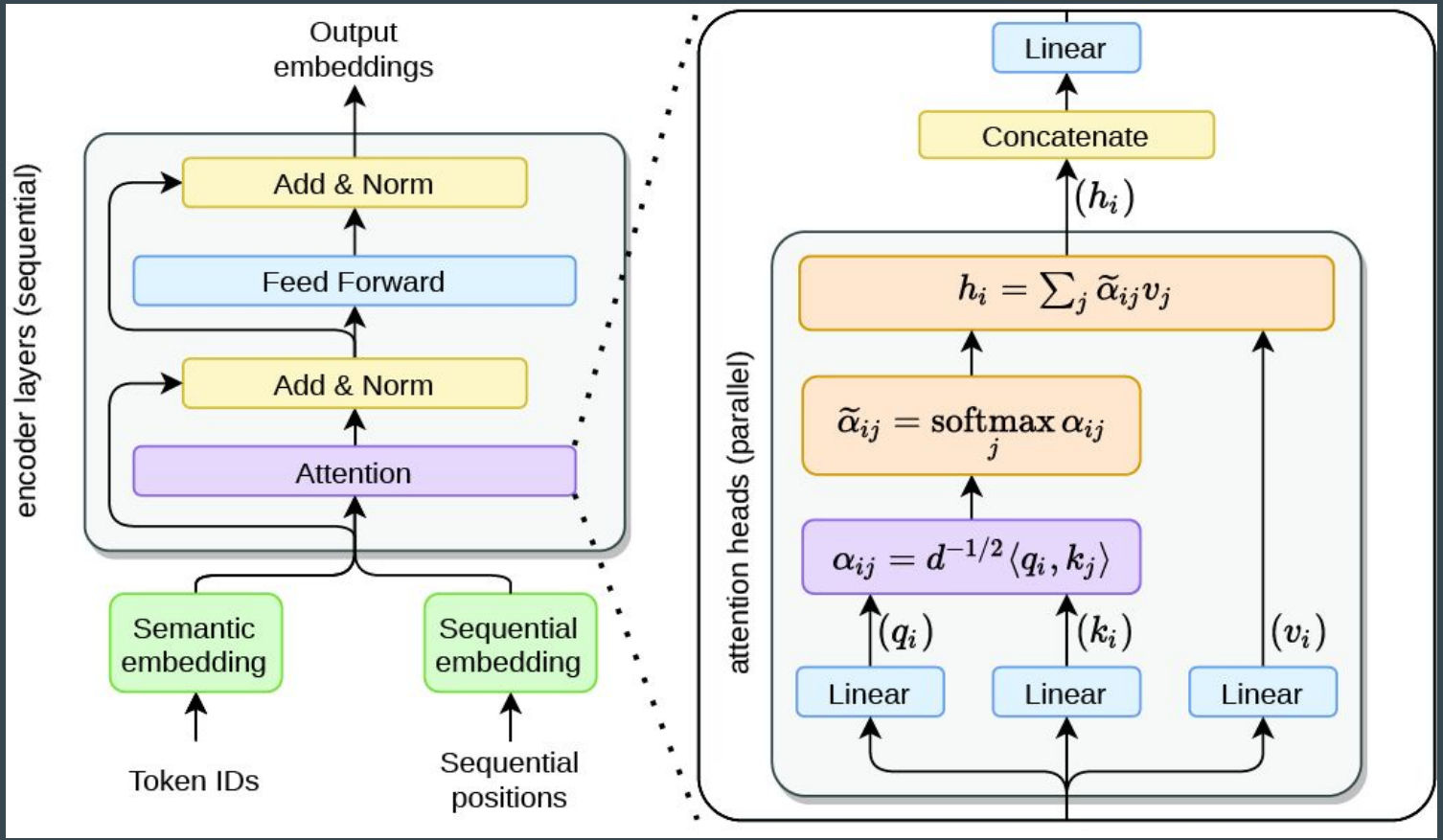
[1] Applica.ai, Zajęcza 15, 00-351 Warszawa, Poland
firstname.lastname@applica.ai
[2] Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland
firstname.lastname@pw.edu.pl
[3] Adam Mickiewicz University, 1 Wieniawskiego, 61-712 Poznań, Poland
firstname.lastname@amu.edu.pl

Abstract. We introduce a simple new approach to the problem of understanding documents where non-trivial layout influences the local semantics. To this end, we modify the Transformer encoder architecture in a way that allows it to use layout features obtained from an OCR system, without the need to re-learn language semantics from scratch. We only augment the input of the model with the coordinates of token bounding boxes, avoiding, in this way, the use of raw images. This leads to a layout-aware language model which can then be fine-tuned on downstream tasks.

The model is evaluated on an end-to-end information extraction task using four publicly available datasets: Kleister NDA, Kleister Charity, SROIE and CORD. We show that our model achieves superior performance on datasets consisting of visually rich documents, while also outperforming the baseline RoBERTa on documents with flat layout (NDA F_1 increase from 78.50 to 80.42). Our solution ranked first on the public leaderboard for the Key Information Extraction from the SROIE dataset, improving the SOTA F_1-score from 97.81 to 98.17.

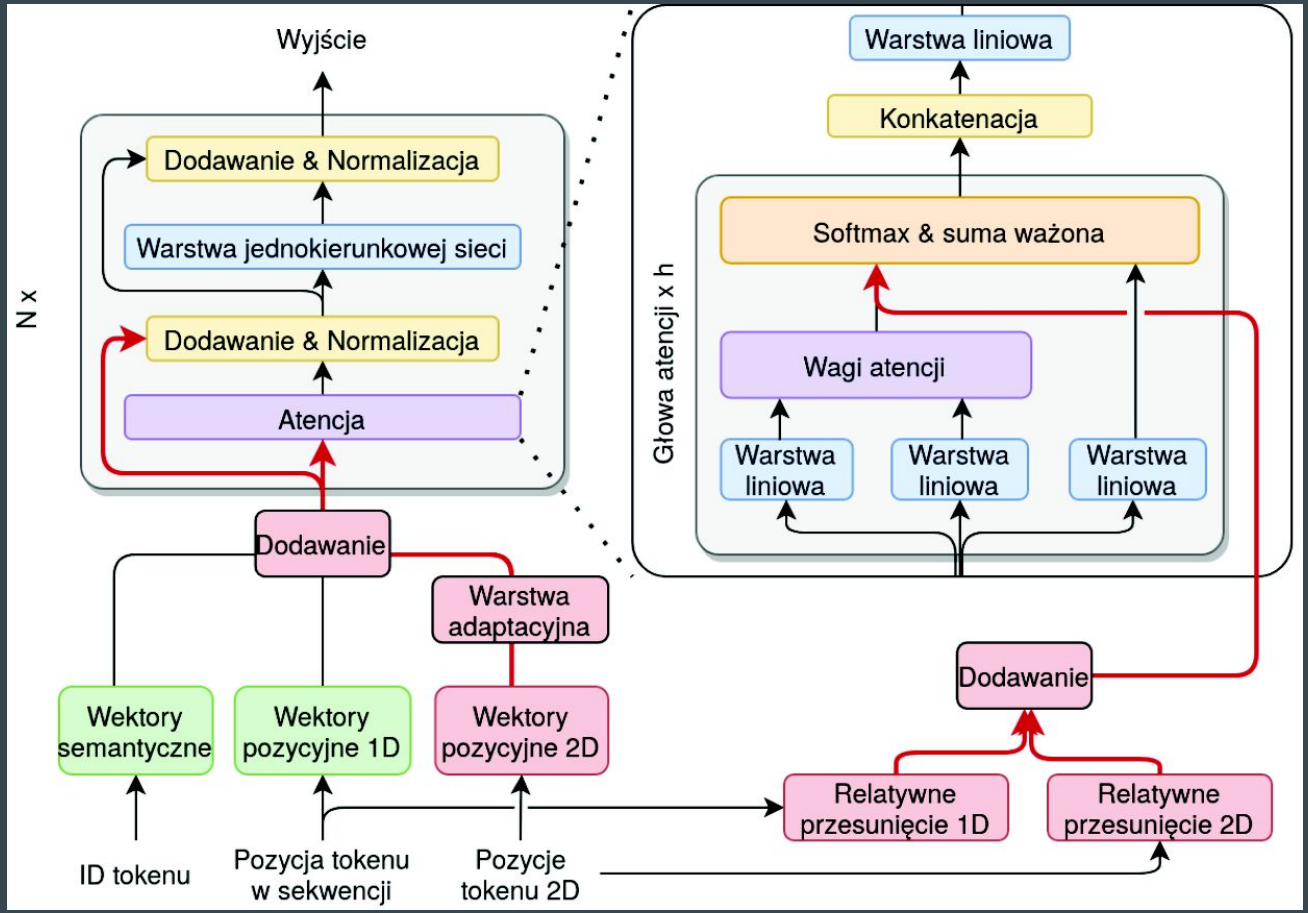
https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - wprowadzenie



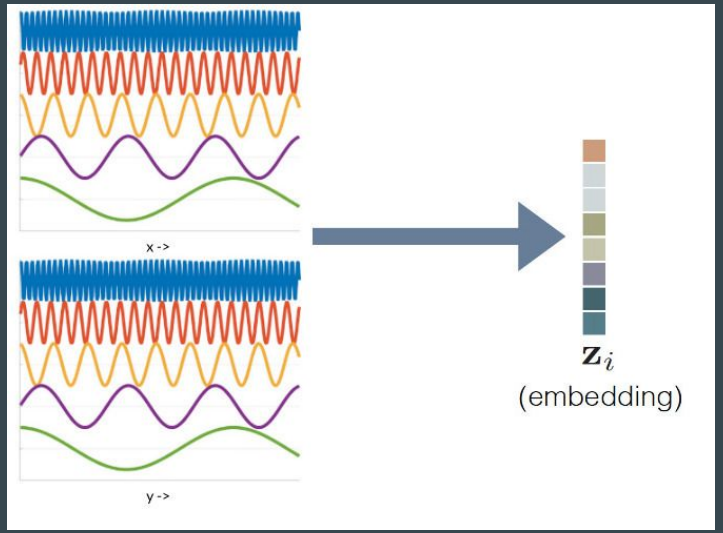
https://arxiv.org/pdf/2002.08087.pdf

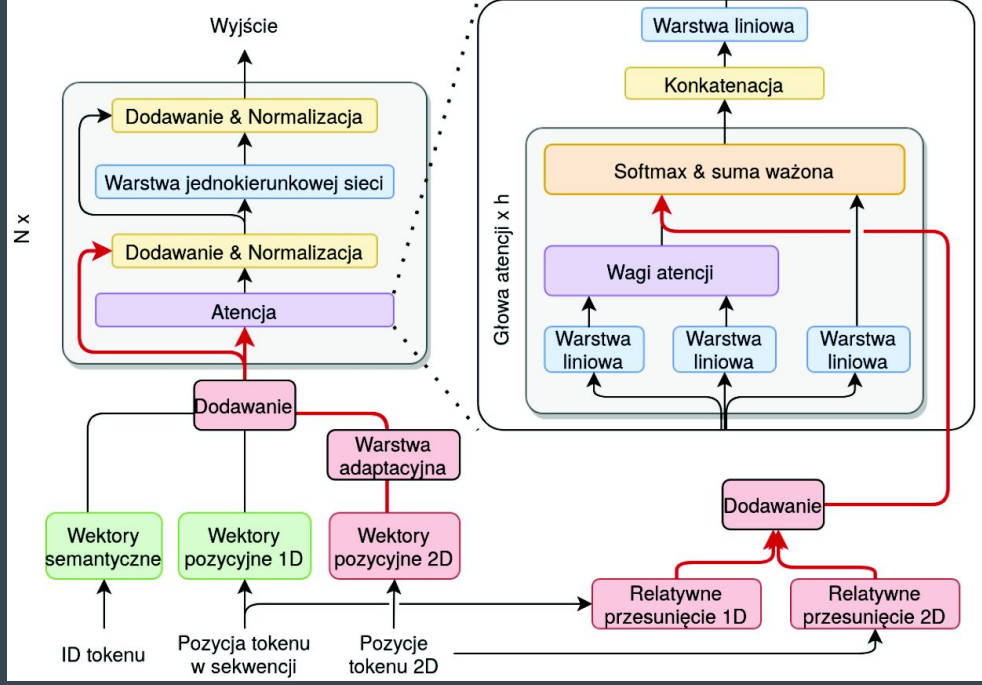
LAMBERT - architektura



https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - architektura



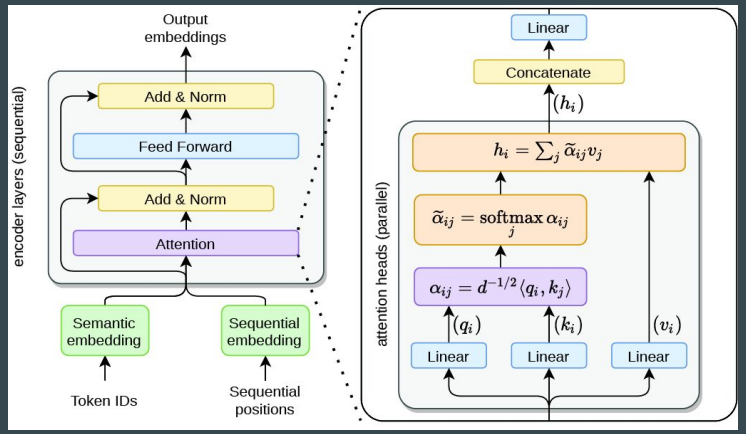


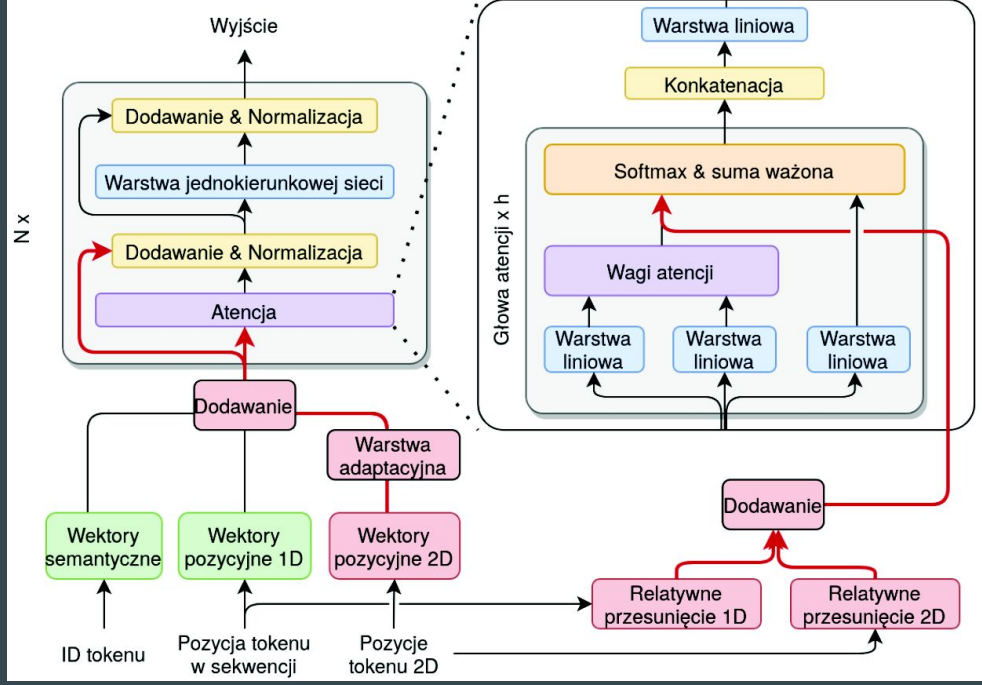
Zmiany wprowadzone względem oryginalnego modelu RoBERTa:

1. Modyfikacja wektora wejściowego x_i o informację o strukturze dokumentu

https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - architektura





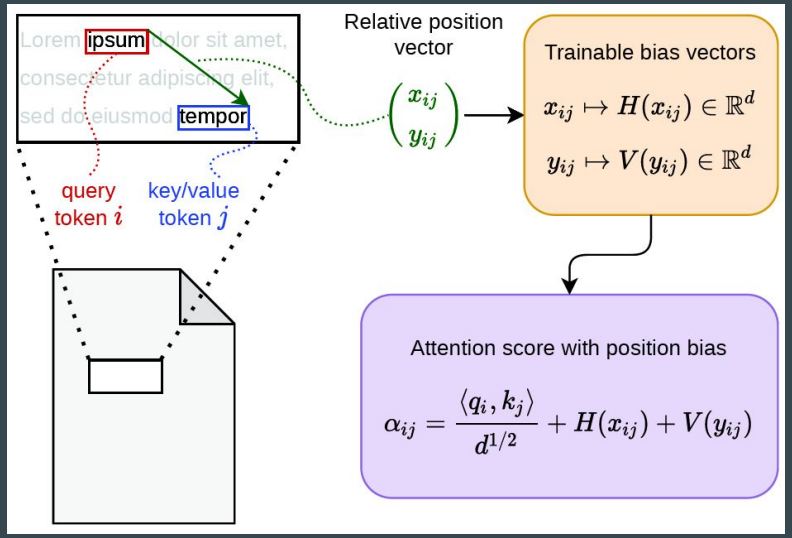
2. Modyfikacja wag atencji poprzez dodanie używanego w modelu T5 relatywnego kodowania pozycji

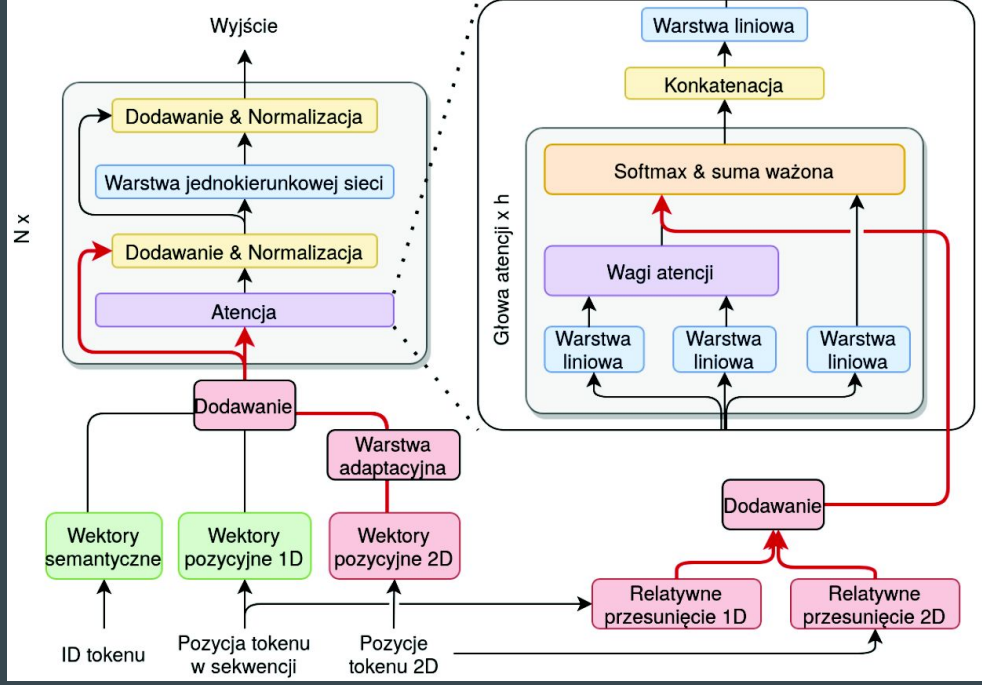
$$\alpha'_{ij} = \alpha_{ij} + \beta_{ij}, \qquad (2)$$

gdzie parametr $\beta_{ij} = W(i-j)$ jest wyuczalnym wektorem wag uzależnionym od relatywnej pozycji tokenów i oraz j,

https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - architektura





3. Rozszerzenie relatywnego kodowania pozycji o dwa dodatkowe parametry, związane z relatywną pozycją dwóch tokenów względem odległości od siebie w poziomie oraz w pionie strony

$$\beta_{ij} = W(i-j) + H(\lfloor \xi_i - \xi_j \rfloor) + V(\lfloor \eta_i - \eta_j \rfloor),$$

gdzie $(\xi_i, \eta_i) = (Cx_1, C(y_1 + y_2)/2)$ są współrzędnymi $i - tego$ tokenu na stronie

https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - badania ablacyjne

Train epochs and pages	Embeddings dimension	Inputs: sequential	Inputs: seq. bias	Inputs: layout	Inputs: 2D bias	Datasets: NDA	Datasets: Charity	Datasets: SROIE*	Datasets: CORD
		•				$78.50_{\pm 1.16}$	$77.88_{\pm 0.48}$	$94.28_{\pm 0.42}$	$91.98_{\pm 0.62}$
			•			$\mathbf{1.94}_{\pm 0.46}$	$-0.82_{\pm 0.74}$	$0.33_{\pm 0.22}$	$-0.15_{\pm 0.49}$
		•	•			$\mathbf{2.42}_{\pm 0.61}$	$0.52_{\pm 0.64}$	$0.79_{\pm 0.17}$	$0.03_{\pm 0.57}$
				•		$1.25_{\pm 0.59}$	$\mathbf{2.62}_{\pm 0.80}$	$\mathbf{1.86}_{\pm 0.15}$	$0.89_{\pm 0.83}$
					•	$-0.49_{\pm 0.62}$	$2.02_{\pm 0.48}$	$0.53_{\pm 0.28}$	$-0.17_{\pm 0.62}$
8×2M	128			•	•	$0.88_{\pm 0.50}$	$\mathbf{3.00}_{\pm 0.37}$	$\mathbf{1.94}_{\pm 0.16}$	$0.68_{\pm 0.62}$
		•		•		$1.74_{\pm 0.67}$	$0.06_{\pm 0.93}$	$\mathbf{1.94}_{\pm 0.18}$	$\mathbf{1.42}_{\pm 0.53}$
		•			•	$1.73_{\pm 0.60}$	$2.02_{\pm 0.53}$	$\mathbf{2.09}_{\pm 0.22}$	$\mathbf{1.93}_{\pm 0.71}$
		•		•	•	$0.54_{\pm 0.85}$	$1.84_{\pm 0.42}$	$\mathbf{2.08}_{\pm 0.38}$	$\mathbf{2.15}_{\pm 0.65}$
		•	•	•		$1.66_{\pm 0.76}$	$0.32_{\pm 1.35}$	$\mathbf{1.75}_{\pm 0.35}$	$1.06_{\pm 0.54}$
		•	•		•	$0.85_{\pm 0.91}$	$1.84_{\pm 0.27}$	$\mathbf{2.01}_{\pm 0.24}$	$\mathbf{1.95}_{\pm 0.46}$
		•	•	•	•	$\mathbf{1.81}_{\pm 0.60}$	$\mathbf{2.06}_{\pm 0.69}$	$\mathbf{1.96}_{\pm 0.16}$	$\mathbf{1.77}_{\pm 0.46}$
	128	•		•		$1.74_{\pm 0.67}$	$0.06_{\pm 0.93}$	$\mathbf{1.94}_{\pm 0.18}$	$\mathbf{1.42}_{\pm 0.53}$
	384	•		•		$0.90_{\pm 0.54}$	$0.70_{\pm 0.40}$	$\mathbf{1.86}_{\pm 0.22}$	$\mathbf{1.51}_{\pm 0.60}$
8×2M	768	•		•		$0.71_{\pm 1.04}$	$0.50_{\pm 0.85}$	$\mathbf{2.18}_{\pm 0.25}$	$\mathbf{1.54}_{\pm 0.51}$
	768[b]	•		•		$0.77_{\pm 0.58}$	$\mathbf{2.30}_{\pm 0.20}$	$0.37_{\pm 0.15}$	$\mathbf{1.58}_{\pm 0.52}$
8×2M		•	•	•	•	$\mathbf{1.81}_{\pm 0.60}$	$2.06_{\pm 0.26}$	$1.96_{\pm 0.18}$	$1.77_{\pm 0.46}$
8×2M[a]	128	•	•	•	•	$\mathbf{1.86}_{\pm 0.66}$	$1.92_{\pm 0.19}$	$\mathbf{2.60}_{\pm 0.18}$	$1.59_{\pm 0.61}$
25×3M[a]		•	•	•	•	$\mathbf{1.92}_{\pm 0.50}$	$\mathbf{3.46}_{\pm 0.21}$	$\mathbf{2.65}_{\pm 0.13}$	$\mathbf{2.43}_{\pm 0.19}$

https://arxiv.org/pdf/2002.08087.pdf

LAMBERT - wyniki

Model	Liczba parametrów	Nasze eksperymenty				Wyniki zewnętrzne
		NDA	Charity	SROIE*	CORD	SROIE
RoBERTa	125M	77.91	76.36	94.05	91.57	92.39[b]
RoBERTa (16M)	125M	78.50	77.88	94.28	91.98	93.03[b]
LayoutLM	113M	77.50	77.20	94.00	93.82	94.38[a]
	343M	79.14	77.13	96.48	93.62	97.09[b]
LAMBERT (16M)	125M	80.31	79.94	96.24	93.75	—
LAMBERT (75M)	125M	**80.42**	**81.34**	**96.93**	**94.41**	**98.17**[b]

https://arxiv.org/pdf/2002.08087.pdf

Due Benchmark

DUE: End-to-End Document Understanding Benchmark

Łukasz Borchmann[*ℵ†] **Michał Pietruszka**[*ℵ‡] **Tomasz Stanisławek**[*ℵ§]

Dawid Jurkiewicz[ℵ¶] **Michał Turski**[ℵ¶] **Karolina Szyndler**[ℵ] **Filip Graliński**[ℵ¶]

[ℵ]Applica.ai
firstname.lastname@applica.ai

[†]Poznan University of Technology [‡]Jagiellonian University, Krakow

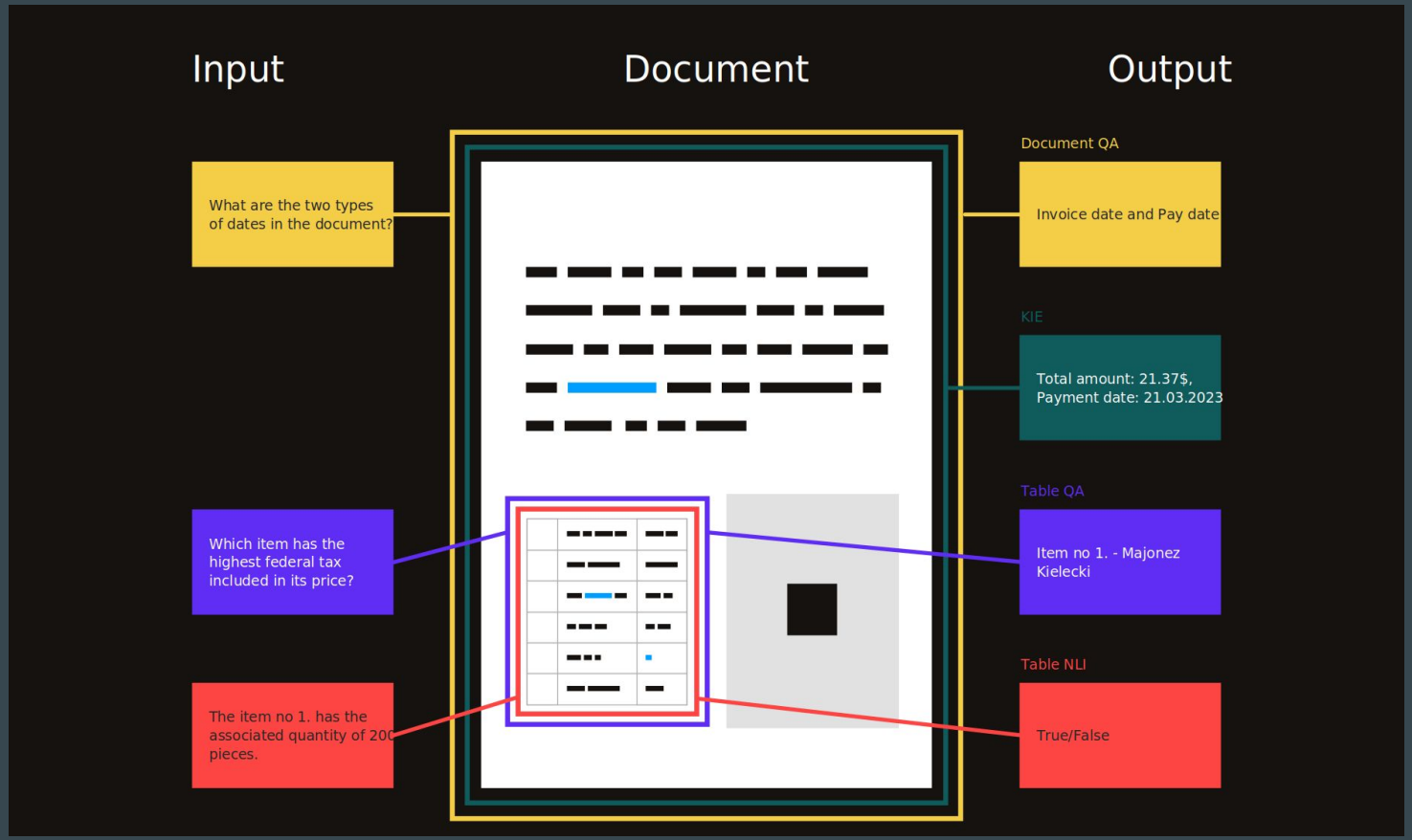
[§]Warsaw University of Technology [¶]Adam Mickiewicz University, Poznan

Abstract

Understanding documents with rich layouts plays a vital role in digitization and hyper-automation but remains a challenging topic in the NLP research community. Additionally, the lack of a commonly accepted benchmark made it difficult to quantify progress in the domain. To empower research in this field, we introduce the Document Understanding Evaluation (DUE) benchmark consisting of both available and reformulated datasets to measure the end-to-end capabilities of systems in real-world scenarios. The benchmark includes Visual Question Answering, Key Information Extraction, and Machine Reading Comprehension tasks over various document domains and layouts featuring tables, graphs, lists, and infographics. In addition, the current study reports systematic baselines and analyzes challenges in currently available datasets using recent advances in layout-aware language modeling. We open both the benchmarks and reference implementations and make them available at https://duebenchmark.com and https://github.com/due-benchmark.

https://openreview.net/pdf?id=rNs2FvIGDK

Due Benchmark - wprowadzenie



https://openreview.net/pdf?id=rNs2FvIGDK

Due Benchmark - wybrane zbiory danych

Dataset	Type	Size (thousands)			Selection criteria				Input	Domain	Comment
		Train	Dev	Test	End-to-end	Quality	Difficulty	Licensing			
Kleister Charity [45]	KIE	1.73	.44	.61	+	+	+	+	Doc.	Finances	
PWC [26]	KIE	.2	.06	.12	+	+	+	+	Doc.	Scientific	
DeepForm [47]	KIE	.7	.1	.3	+	+	+	+	Doc.	Finances	
DocVQA [33]	Visual QA	10.2	1.3	1.3	+	+	+	+	Doc.	Business	
InfographicsVQA [32]	Visual QA	4.4	.5	.6	+	+	+	+	Doc.	Open	
TabFact [7]	Table NLI	13.2	1.7	1.7	+	+	+	+	Exc.	Open	
WTQ [39]	Table QA	1.4	.3	.4	+	+	+	+	Exc.	Open	
Kleister NDA [45]	KIE	.25	.08	.2	+	+	–	+	Doc.	Legal	Dominated by extraction from free text
SROIE [20]	KIE	.63	-	.35	+	+	–	+	Doc.	Finances	No room for improvement
CORD [38]	KIE	.8	.1	.1	+	+	–	+	Doc.	Finances	No room for improvement
Wildreceipt [46]	KIE	1.27	-	.47	+	+	–	+	Doc.	Finances	No room for improvement
WebSRC [5]	KIE	4.55	.9	1.0	+	–	+	+	Doc.	Open	Templated input data
FUNSD [24]	KIE	.15	-	.05	+	–	+	+	Doc.	Finances	Known disadvantages [51]
DocVQA [32]	Visual QA	4.4	.5	.6	–	+	+	+	Col.	Open	Document Collection Question Answering
TextbookQA [28]	Visual QA	.67	.2	.21	+	–	+	+	Doc.	Educational	Source files are not available
MultiModalQA [48]	Visual QA	23.82	2.44	3.66	+	–	+	+	Doc.	Open	Automatically generated questions
VisualMRC [49]	Visual MRC	7	1	2	+	+	–	+	Doc.	Open	Human performance reached
RVL-CDIP [17]	Classification	320	40	40	+	+	–	+	Doc.	Finances	No room for improvement
DocFigure [25]	Classification	19.8	-	13.1	+	+	–	+	Doc.	Scientific	No room for improvement
EURLEX57K [3]	Classification	45	6	6	+	+	–	+	Doc.	Legal	Dominated by extraction from free text
MELINDA [54]	Classification	4.34	.45	.58	+	–	+	+	Doc.	Scientific	Semi-supervised annotation
S2-VL [44]	DLA	1.3	-	-	–	+	+	+	Doc.	Scientific	Cross-validation for training and testing
DocBank [30]	DLA	398	50	50	–	–	+	+	Doc.	Scientific	Automatic annotation
Publaynet [61]	DLA	340.4	11.9	12	–	–	+	+	Doc.	Scientific	Automatic annotation
FinTabNet [60]	DLA	61.8	7.19	7.01	–	+	+	+	Doc.	Finances	Different styles in comparison to sci./gov. docs
PlotQA [34]	Figure QA	157	33.7	33.7	+	–	+	+	Exc.	Open	Synthetic
Leaf-QA [4]	Figure QA	200	40	8.15	+	–	+	+	Exc.	Open	Templated questions
TAT-QA [62]	Table QA	2.2	.28	.28	+	–	+	+	Exc.	Finances	Source files are not available
WikiOPS [9]	Table QA	17.28	2.47	4.67	+	+	–	+	Exc.	Open	No room for improvement
FeTaQA [35]	Table QA	7.33	1.0	2.0	+	–	+	+	Exc.	Open	Answers as a free-form text
HybridQA [8]	Table QA	62.68	3.47	3.46	–	+	+	+	Col.	Open	Multihop Question Answering
OTT-QA [6]	Table QA	41.46	2.24	2.16	–	+	+	+	Col.	Open	Multihop Question Answering
INFOTABS [14]	Table NLI	1.74	.2	.6	+	+	+	+	Col.	Open	TabFact is very similar

https://openreview.net/pdf?id=rNs2FvIGDK

Due Benchmark - wybrane zbiory danych

Ranking Table

Description Paper Source Code

Date	Method	Recall	Precision	Hmean
2021-11-24	StrucTexT	98.70%	98.70%	98.70%
2022-03-18	GraphDoc	98.13%	98.77%	98.45%
2022-01-21	Textmind + ERNIE-Layout	97.26%	99.48%	98.36%
2021-04-19	IE	97.05%	99.56%	98.29%
2021-07-20	Linklogis_BeeAI	97.05%	99.34%	98.18%
2021-01-02	Applica.ai Lambert 2.0 + Excluding OCR Errors + Fixing total entity	96.83%	99.56%	98.17%
2021-06-02	Multimodal Transformer for Information Extraction	96.76%	99.56%	98.14%
2021-02-16	Applica.ai TILT + Excluding OCR Errors + Fixing total entity	96.83%	99.41%	98.10%
2020-12-24	LayoutLM 2.0 (single model)	96.61%	99.04%	97.81%
2021-01-01	Applica.ai Lambert 2.0 + Excluding OCR Mismatch	96.40%	99.11%	97.74%
2020-12-07	Tencent Youtu	96.47%	98.89%	97.67%
2020-12-28	IE method	96.33%	98.53%	97.41%
2020-05-07	HIK_OCR_Exclude_ocr_mismatch	96.33%	98.38%	97.34%
2020-04-18	LayoutLM + Excluding OCR Mismatch	96.04%	98.16%	97.09%
2021-10-25	Character-Aware CNN + Highway + BiLSTM	96.18%	97.45%	96.81%
2020-04-15	PICK-PAPCIC & XZMU	95.46%	96.79%	96.12%
2020-04-16	LayoutLM	96.04%	96.04%	96.04%
2020-03-26	Applica.ai roberta-base-2D	95.39%	95.80%	95.60%
2021-12-08	1208-cblm	94.81%	95.22%	95.02%
2021-12-08	1208-cblm	94.24%	94.65%	94.44%
2022-02-28	CHL	94.24%	94.65%	94.44%
2020-06-05	great	94.24%	94.24%	94.24%
2019-08-14	PATech_AICenter	94.02%	94.02%	94.02%
2021-06-11	IE + GCN+ OCR	92.44%	94.90%	93.65%
2021-02-21	RoBERTa-base finetuned on business documents	92.80%	93.27%	93.03%

method: Applica.ai Lambert 2.0 + Excluding OCR Errors + Fixing total entity 2021-01-02

Authors: **Applica.ai research team**

Affiliation: **Applica.ai**

Description: **Following the same evaluation rules as others, the OCR mismatch errors are excluded in the submission.**

Additionally, we have manually fixed annotation discrepancies in "total" entity in the test set.

Note:
1. We submitted the best solution out of 100 fine-tuned models
2. In this task there is an annotation discrepancy in "total" entity which caused unfair comparison between models (In train/test sets "total" entity was randomly prefixed by "RM"). Number of errors in the top solutions caused by this kind of annotation error:

Applica.ai Lambert 2.0 + Excluding OCR Errors + Fixing total entity = 0
LayoutLM 2.0 (single model) = 3 (example: 275)
Applica.ai Lambert 2.0 + Excluding OCR Mismatch = 8 (example: 77)
Tencent Youtu = 8 (example: 120)
VIE = 0
HIK_OCR_Exclude_ocr_mismatch = 0
LayoutLM + Excluding OCR Mismatch = 9 (example: 121)

Garncarek, et al. "LAMBERT: Layout-Aware (Language) Modeling using BERT for information extraction" arxiv preprint

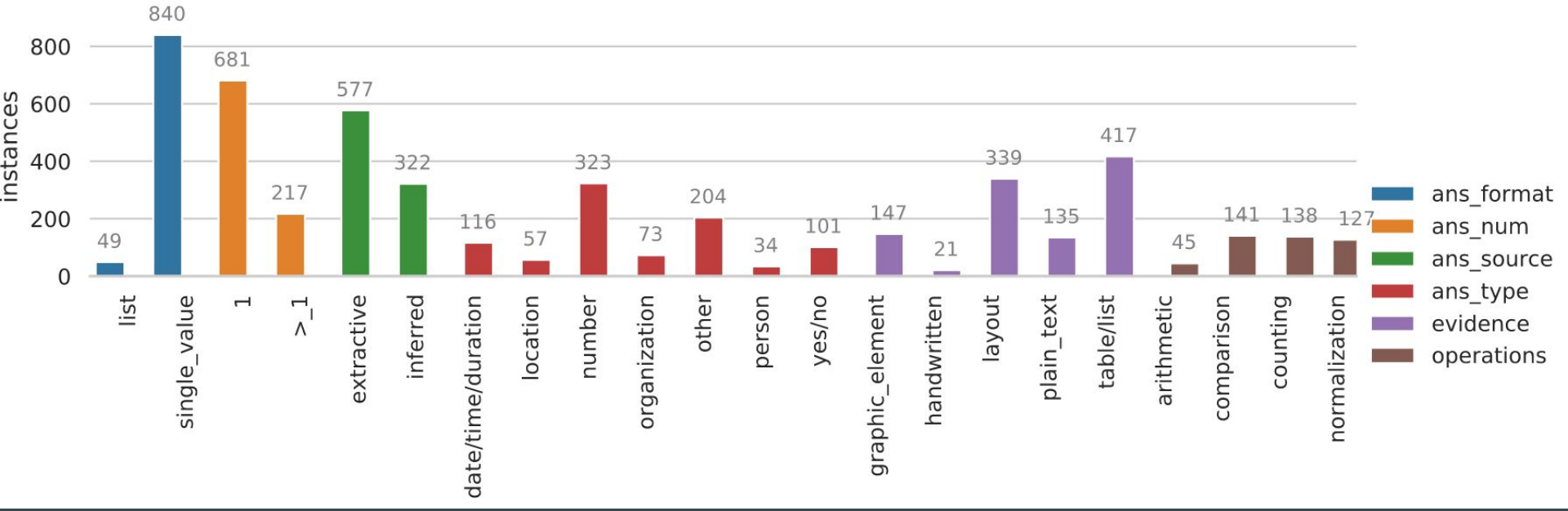
https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=3

Due Benchmark - wybrane zbiory danych

Nazwa zbioru	Liczność (tys.)			Wkład do zbiorów						Domena	Typ zadania
	Zbiór trenujący	Zbiór walidujący	Zbiór testowy	Obliczenie trudności	Poprawa zbioru	Przeformułowanie	Poprawa podziału	Ujednolicenie formatu	Zbiór diagnostyczny		
Kleister Charity [43]	1.73	0.44	0.61	−	−	−	−	+	+	Finansowa	Ekstrakcja informacji z długich dokumentów
PWC [17]	0.2	0.06	0.12	+	+	+	+	+	+	Naukowa	Ekstrakcja informacji z publikacji naukowych
DeepForm [47]	0.7	0.1	0.3	+	+	−	+	+	+	Finansowa	Ekstrakcja informacji z dokumentów biznesowych
DocVQA [27]	10.2	1.3	1.3	−	−	−	−	+	+	Biznesowa	Zadawanie pytań na dokumentach biznesowych
InfographicsVQA [26]	4.40	0.50	0.60	−	−	−	−	+	+	Ogólna	Zadawanie pytań na infografikach
TabFact [5]	13.2	1.7	1.7	−	−	+	−	+	+	Ogólna	Problem *Natural Language Inference* na tabelkach
WTQ [31]	1.4	0.3	0.4	+	−	+	+	+	+	Ogólna	Zadawanie pytań na tabelkach

https://openreview.net/pdf?id=rNs2FvIGDK

Due Benchmark - zbiory diagnostyczne



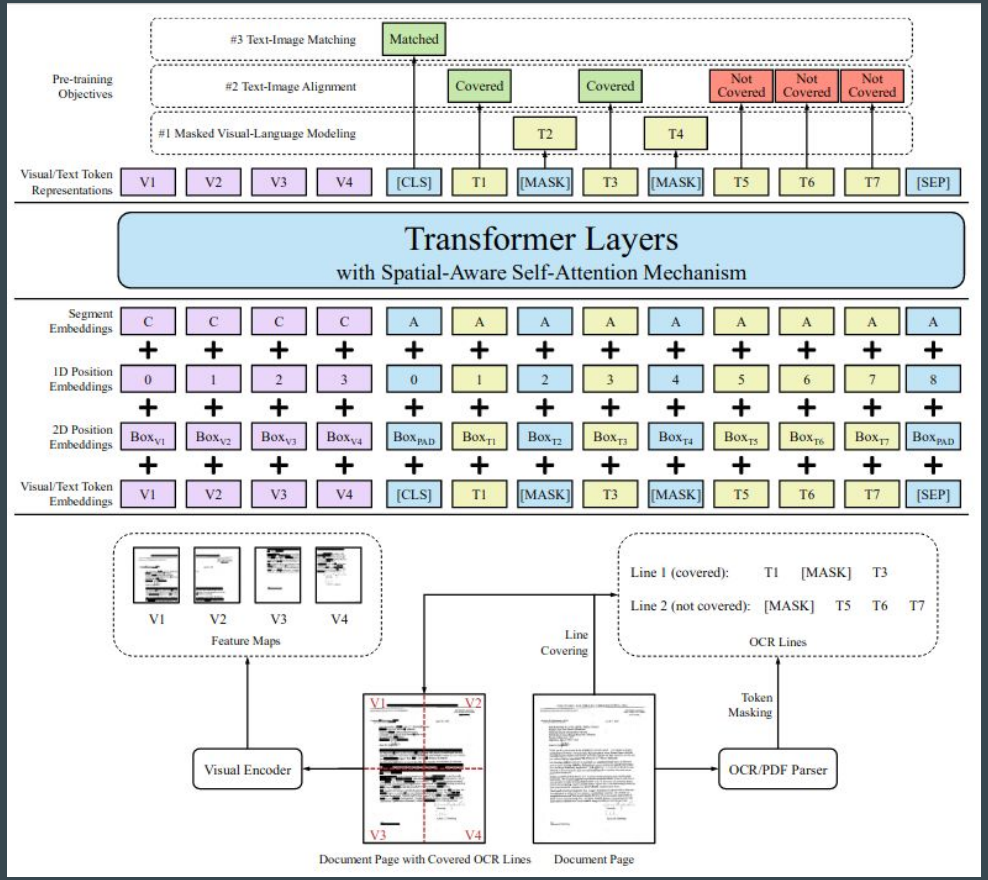
https://openreview.net/pdf?id=rNs2FvIGDK

Due Benchmark - wyniki

Dataset / Task type	Score (task-specific metric)					
	T5	T5+2D	T5+U	T5+2D+U	External best	Human
DocVQA	$70.4_{\pm 2.1}$	$69.8_{\pm 0.7}$	$76.3_{\pm 0.3}$	$81.0_{\pm 0.2}$	87.1 [40]	98.1
InfographicsVQA	$36.7_{\pm 0.6}$	$39.2_{\pm 1.0}$	$37.1_{\pm 0.2}$	$46.1_{\pm 0.1}$	61.2 [40]	98.0
Kleister Charity	$74.3_{\pm 0.3}$	$72.6_{\pm 1.1}$	$76.0_{\pm 0.1}$	$75.9_{\pm 0.7}$	83.6 [63]	97.5
PWC★	$25.3_{\pm 3.3}$	$25.7_{\pm 1.0}$	$27.6_{\pm 0.6}$	$26.8_{\pm 1.8}$	—	69.3
DeepForm★	$74.4_{\pm 0.6}$	$74.0_{\pm 0.7}$	$82.9_{\pm 0.9}$	$83.3_{\pm 0.3}$	—	98.5
WikiTableQuestions★	$33.3_{\pm 0.7}$	$30.8_{\pm 1.9}$	$38.1_{\pm 0.1}$	$43.3_{\pm 0.4}$	—	76.7
TabFact★	$58.9_{\pm 0.5}$	$58.0_{\pm 0.3}$	$76.0_{\pm 0.1}$	$78.6_{\pm 0.1}$	—	92.1
Visual QA	53.6	54.5	56.7	63.5	n/a	98.1
KIE	69.1	67.7	74.8	76.4	n/a	88.4
Table QA/NLI	29.4	29.0	38.0	39.3	n/a	84.4
Overall	50.7	50.4	56.5	59.8	n/a	90.3

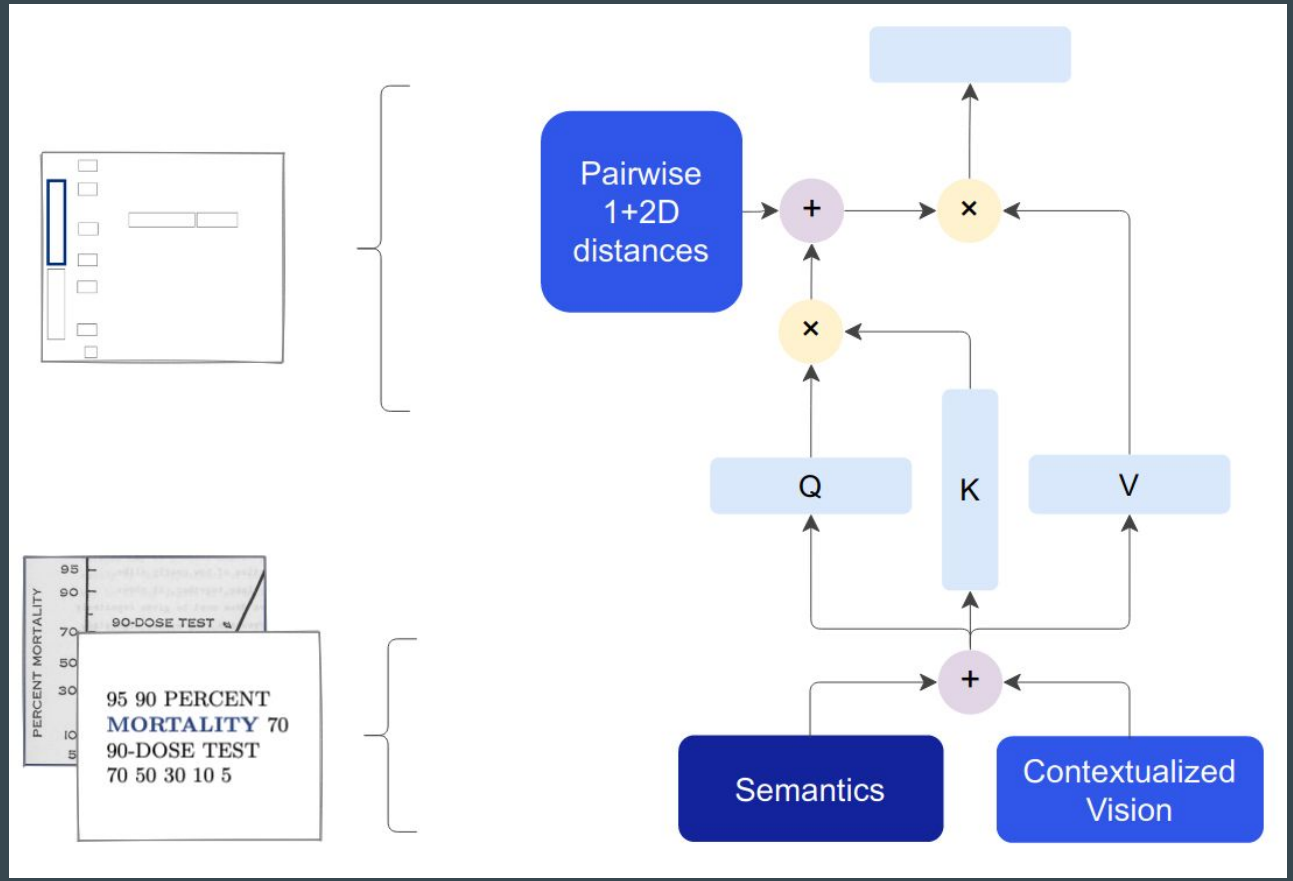
https://openreview.net/pdf?id=rNs2FvIGDK

Rozwoju dziedziny - LayoutLMv2 (modalność wizualna)



https://arxiv.org/abs/2012.14740

Rozwoju dziedziny - TILT (modalność wizualna + seq2seq)



https://arxiv.org/pdf/2102.09550.pdf

Pytania / Komentarze