

Generowanie słów kluczowych z krótkich tekstów za pomocą modelu T5

Piotr Pęzik (UŁ/VoiceLab), Agnieszka Mikołajczyk, Adam Wawrzyński (VoiceLab),
Bartłomiej Nitoń, Maciej Ogrodniczuk (IPI PAN)

Streszczenie

Tematem wystąpienia jest ewaluacja zastosowań polskiej wersji modelu językowego T5 (pIT5) w wydobywaniu słów kluczowych z krótkich tekstów. Autorzy opisują trening i testy modelu na opracowanym w ramach projektu CURLICAT Otwartym Korpusie Metadanych Nauki Polskiej. Porównana zostanie jakość słów kluczowych generowanych czterema różnymi metodami: a) pIT5 jako model text-to-text, b) extremeText jako klasyfikator z nadzorem, c) TermoPL jako metoda ekstrakcji terminologii oraz d) KeyBERT jako alternatywny model transformerowy. Zilustrowana zostanie również nieoczekiwana przenośność wytrenowanego modelu pIT5kw do domen tematycznych niezwiązanych z korpusem streszczeń artykułów naukowych, takich jak teksty informacyjne, czy też transkrypcje dialogów telefonicznych.

Słowa kluczowe: CURLICAT, Otwarty Korpus Metadanych Nauki Polskiej, model językowy T5, ekstrakcja terminologii

Wygenerowane przez: pIT5kw

Słowa kluczowe

- Bardzo różne definicje, np.
 - Wyrazy istotnie częstsze, równomiernie występujące w danym tekście, korpusie, okresie
 - Deskryptory/metadane artykułów naukowych. Są to zazwyczaj wyrazy lub frazy, przeważnie rzeczownikowe, które zwięźle opisują treść artykułu
 - Ograniczone do 3-5 pozycji. Otwarty, oddolnie tworzony, słabo kontrolowany słownik
 - Można je identyfikować ekstrakcyjnie, tzn. na podstawie dosłownych wystąpień w tekście, lub abstrakcyjnie, tzn. są jego uogólnionym opisem
 - Mają poprawić indeksowanie artykułów w bibliotekach cyfrowych

Ocena plonowania kilku odmian pomidora w uprawie ekologicznej

AUTORZY A. Szafirowska

ABSTRAKT

POL **ENG**

W latach 2005-2006 w Instytucie Warzywnictwa w Skierniewicach prowadzono badania porównawcze plonowania pomidora uprawianego metodą ekologiczną i konwencjonalną. Oceniano polskie odmiany nowej generacji, samokończące, z przeznaczeniem dla przemysłu. Badano następujące odmiany: Awizo F1, Batory F1, Etna F1, Rumba, Sokal F1, Wiola F1. Czynnikiem najsilniej wpływającym na wysokość i strukturę plonu było występowanie zarazy ziemniaka. W warunkach łagodnego przebiegu choroby (2005 rok) plon pomidora z uprawy ekologicznej był o około 9% niższy niż z konwencjonalnej. Wyróżniającą odmianą pod względem wysokości i struktury plonu w obu metodach uprawy była Batory F1. W warunkach dużego nasilenia choroby (2006 rok) w uprawie ekologicznej uzyskano o 26,5% niższy plon ogólny, a ilość porażonych owoców była 2,5-krotnie większa niż w uprawie konwencjonalnej. W obu latach badań niezależnie od warunków i metody uprawy najmniej owoców z objawami zarazy ziemniaka stwierdzono u odmiany Awizo F1.

SŁOWA KLUCZOWE

POL **ENG**

plonowanie, pomidor, odmiana, uprawa ekologiczna

Badania enzymatycznego wydzielania nasion z pulpy przy osmotycznym zablokowaniu wchłaniania wody

AUTORZY [O. Domoradzka](#), [M. Domoradzki](#), [W. Weiner](#), [Z. Witek](#)

ABSTRAKT

POL ENG

W pracy przebadano możliwość osmotycznego blokowania wchłaniania wody przez nasiona w procesie enzymatycznego wydzielania nasion z pulpy owocowej. Określono parametry procesu, zapewniającego zachowanie jakości i zdolności kiełkowania nasion mimo długiego czasu ich przebywania w zawieszynie wodnej.

SŁOWA KLUCZOWE

POL ENG

wydzielanie enzymatyczne, nasiona, pomidor, papryka, ogórek

W treści pełnego artykułu?

Cebula i kapusta

AUTORZY Aleksander Genis

ABSTRAKT

ENG POL

Biorąc za punkt wyjścia budowę tytułowych warzyw, jednego posiadającego rdzeń widoczny po oderwaniu wierzchnich warstw, drugiego zaś pozbawionego go, niniejszy artykuł jest próbą ponownego odczytania literatury rosyjskiej za pomocą dwóch, przeciwstawnych perspektyw: paradygmatu kapusty i paradygmatu cebuli. Pierwszy paradygmat oparty jest na metafizyce sowieckiej, drugi zaś wyrasta z głębszych filozoficznych lektur, wedle których znaczenie kształtuje się poprzez pustkę i chaos. Oba paradygmaty użyte są jako odrębne perspektywy interpretacyjne i zilustrowane adekwatnymi przykładami literackimi.

SŁOWA KLUCZOWE

POL ENG

Teoria literatury, Literatura rosyjska, Historia literatury, Postmodernizm

Termin abstrakcyjny

Co chcemy stworzyć?

- Rozwiązanie, które przypisuje słowa kluczowe do krótkich tekstów
- Mogą to być jedno lub wielowyrazowe frazy, o ile są kompletne składniowo/semantycznie
- Preferowane są nominalizacje, ale nie we wszystkich domenach, zob. Twitter, 'hasztagi', slogany typu 'SzczepimySię' itp.
- Mogą występować w tekście, dosłownie, w sposób nieciągły, częściowo w tekście, w różnym szyku, przypadku, formie itp.
- W niektórych przypadkach są konwencjonalnymi uogólnieniami treści, np. 'postmodernizm', 'teoria literatury' itp.
- Powinny być lematyzowane z zachowaniem uzgodnień, relacji między składnikami fraz, wielkości liter itp.
 - Tatrzański Park Narodowy, Instytut Pamięci Narodowej, Kwiaty Polskie
- Powinny być precyzyjne, w mniejszym stopniu kompletne

Zastosowania wykrywania słów kluczowych

- Pierwotne: indeksowanie tematyczne tekstów naukowych
- Transfer do innych domen?
 - Np. indeksowanie tematyczne krótkich tekstów z innych domen, gatunków, typów tekstów? Ew. indeksowanie długich tekstów poprzez agregację etykiet przypisywanych fragmentom długich tekstów
 - Indeksowanie informacji prasowych
 - Oznaczanie tematów w dialogach telefonicznych

CURLICAT – Curated Multilingual LRs for CEF AT

Program: Instrument CEF (Connecting Europe Facility)

Nr umowy: INEA/CEF/ICT/A2019/1926831

Konsorcjanci: BG, HR, HU, PL, RU, SI, SK

Czas realizacji: czerwiec 2020 – maj / listopad 2022



Po polsku: Zoptymalizowane wielojęzyczne zasoby językowe na potrzeby tłumaczenia maszynowego

Czyli chodzi o: Udostępnienie anotowanych lingwistycznie korpusów językowych (20 mln słów) na bazie danych „z korpusów narodowych i innych dostępnych korpusów”.

A konkretnie o: Pozyskanie nowych tekstów z Biblioteki Nauki (<https://bibliotekanauki.pl>), harmonizację metadanych, anotacja lingwistyczna ...

POSMAC

Table 1. Top 10 scientific domains represented in the POSMAC.

Domains	Documents	With keywords
Engineering and technical sciences	58 974	57 165
Social sciences	58 166	41 799
Agricultural sciences	29 811	15 492
Humanities	22 755	11 497
Exact and natural sciences	13 579	9 185
Humanities, Social sciences	12 809	7 063
Medical and health sciences	6 030	3 913
Medical and health sciences, Social sciences	828	571
Humanities, Medical and health sciences, Social sciences	601	455
Engineering and technical sciences, Humanities	312	312

Łącznie 216 214 abstraktów (czasem same tytuły)

Directory of Open Access Journals

- Baza danych agregująca ponad 16,500 czasopism naukowych o otwartym dostępie
- Metadane czasopism są dostępne na licencji [Creative Commons Attribution-ShareAlike 4.0 International \(CC BY-SA 4.0\) license](#)
- Metadane artykułów są dostępne na licencji [CC0 1.0 Universal \(CC0\) Public Domain Dedication](#)
- Zawiera metadane artykułów publikowanych w wielu językach
 - często streszczenie lub słowa kluczowe są w języku angielskim

POSMAC vs Directory of OA Journals

Zbiór	POSMAC	DOAJ
Dokumenty w języku polskim	216,214	17,319
Dokumentów ogółem	216,214	6,364,023
Liczba języków	2	81

FastText i extremeText

- Generowanie słów kluczowych jako klasyfikacja tekstów z nadzorem
- Używamy rozszerzenia FastText, tzn. extremeText, w którym zaimplementowano funkcję kosztu Probabilistic Labels Tree (Wydmuch et al. 2018), pozwalającą trenować modele na bardzo dużych taksonomiach
- Duży zbiór uczący, ponad 100 tys. streszczeń ze słowami kluczowymi
- ... ale ponad połowa występuje tylko raz
- Spodziewana niska przenośność do innych domen, zwłaszcza do danych konwersacyjnych, jako że model ten uczy się przypisywać etykiety na podstawie danych z domeny źródłowej

KeyBERT

- Adaptacja modelu transformerowego BERT do zadania identyfikacji słów kluczowych
- KeyBERT tworzy wektorowe reprezentacje dokumentu
- Wektory n-gramów o zadanej długości porównywane są z wektorami dokumentu
- Ranking kluczowych terminów liczony jest wg podobieństwa kosinusowego między wektorami n-gramów i dokumentu
- Czysto 'ekstrakcyjne' podejście, łatwo zidentyfikować terminy w tekście
- Arbitralna operacjonalizacja kluczowości, tylko n-gramy, brak normalizacji

TermoPL

- Narzędzie do wydobywania terminologii z *korpusów* (Marciniak et al. 2016)
- Identyfikuje terminy pasujące do schematów składniowych i tworzy ranking m.in. w oparciu o wariant miary C/NC-value (Frantzi et al. 2000)
- Użyliśmy do identyfikacji fraz z krótkich tekstów, aby sprawdzić maksymalne pokrycie tego czysto ekstrakcyjnego podejścia
- Planowane stworzenie listy rankingowej na zbiorze treningowym

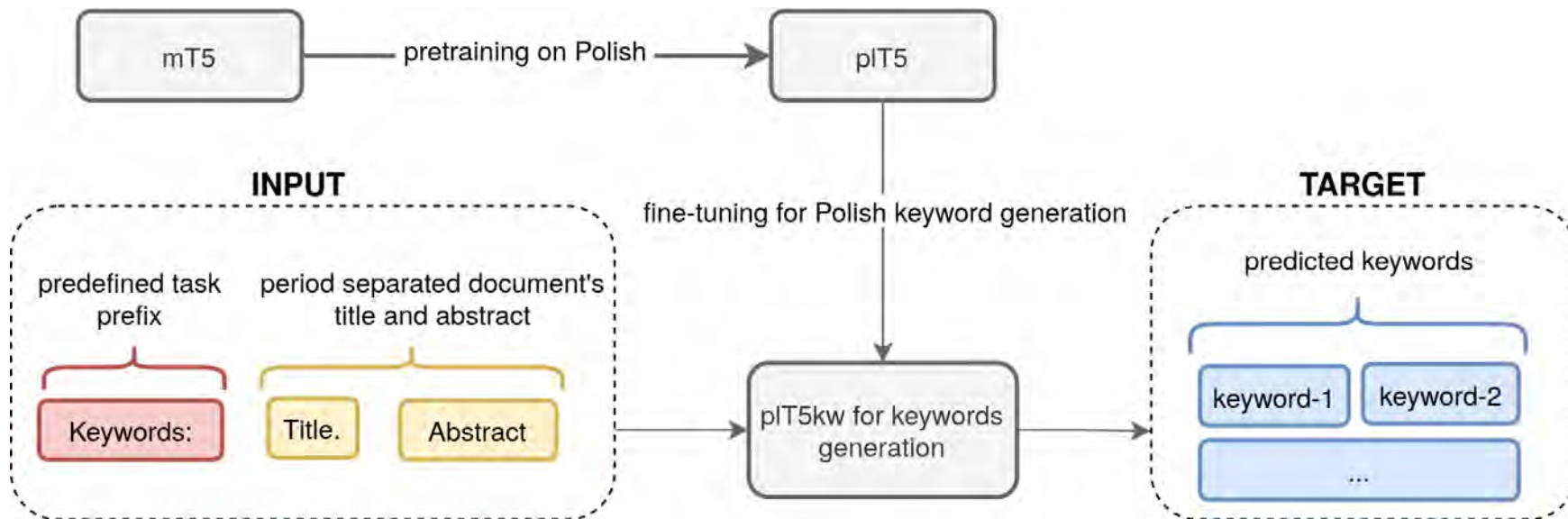
Text-to-Text Transfer Transformer – T5

- Model w architekturze koder-dekoder wykorzystujący bloki Transformer zaprezentowany przez Google (<https://huggingface.co/t5-base>)
- Wielozadaniowy
 - Wejściem modelu jest tekst poprzedzony prefiksem, a wyjściem jest docelowy tekst. Prefiks definiuje typ zadania: np. "Przetłumacz na z polskiego na angielski:"
- **Zalety:**
 - Taka sama funkcja celu do wszystkich zadań, od analizy sentymentu, przez tłumaczenie i odpowiadanie na pytania, ale też klasyfikacja czy nawet zadania regresyjne
 - Wyniki SOTA na wielu zadaniach
 - Umożliwia wykorzystanie jednego modelu do wielu zadań, jednocześnie często uzyskując na nich wyższe wyniki niż trenowane osobno
- **Wada:**
 - wymaga dostosowania zadania do problemu generowania tekstu
 - 2-krotnie większa liczba parametrów w stosunku do modeli zawierających jedynie koder (BERT) lub dekodery (GPT)

Polska wersja T5 – plT5

- Wielojęzyczny model mT5 dotrenowany na danych polskich
 - Dotrenowany na sześciu korpusach
 - Ewaluowany do zadań streszczania tekstu, Q&A, tłumaczenia maszynowego i inne
 - Opracowany i udostępniony przez zespół Allegro i IPI PAN
- Dostępny w wersjach small/base/large (huggingface.co/allegro/plT5-base)
- 50k tokenów zamiast 250k (źródło: prezentacja Allegro na NLP Day)
- Lepsze wyniki na zadaniach z KLEJ niż często używany Herbert i mT5

Model do generacji słów kluczowych – pIT5kw



Model do generacji słów kluczowych – pIT5kw

- Dotrenowany model pIT5 base (huggingface.co/allegro/plt5-base)
- Model wytrenowany na parach tytuł + streszczenie i słowa kluczowe
(dane wejściowe) (docelowe wyjście)
- Trenowany z nowym prefiksem “Słowa kluczowe: “
- Charakterystyka
 - Wysoka precyzja, trafne słowa kluczowe
 - Lepsze wyniki zero-shot na języku angielskim niż istniejące metody trenowane na języku polskim i testowane na polskim

Parametry treningu: 10 epok, batch_size=8, lr=3e-5, multiplicative scheduler=0.7, and 100 warm-up steps linearly increasing the learning rate from zero to a target lr

Ewaluacja wewnętrzna

Table 2. Results of evaluation on the full set of POSMAC keywords.

Method	Rank	P	Micro R	F ₁	P	Macro R	F ₁
extremeText	1	0.175	0.038	0.063	0.007	0.004	0.005
	3	0.117	0.077	0.093	0.011	0.011	0.011
	5	0.090	0.099	0.094	0.013	0.016	0.015
	10	0.060	0.131	0.082	0.015	0.025	0.019
plT5kw	1	0.345	0.076	0.124	0.054	0.047	0.050
	3	0.328	0.212	0.257	0.133	0.127	0.129
	5	0.318	0.237	0.271	0.143	0.140	0.141
KeyBERT	1	0.030	0.007	0.011	0.004	0.003	0.003
	3	0.015	0.010	0.012	0.006	0.004	0.005
	5	0.011	0.012	0.011	0.006	0.005	0.005
TermoPL	1	0.118	0.026	0.043	0.004	0.003	0.003
	3	0.070	0.046	0.056	0.006	0.005	0.006
	5	0.051	0.056	0.053	0.007	0.007	0.007
	ALL	0.025	0.339	0.047	0.017	0.030	0.022

Ewaluacja wewnętrzna cd.

Table 3. Results of evaluation on keywords occurring at least 10 times in the POSMAC dataset.

Method	Rank	Micro			Macro		
		P	R	F ₁	P	R	F ₁
extremeText	1	0.210	0.077	0.112	0.037	0.017	0.023
	3	0.139	0.152	0.145	0.045	0.042	0.043
	5	0.107	0.196	0.139	0.049	0.063	0.055
	10	0.072	0.262	0.112	0.041	0.098	0.058
plT5kw	1	0.377	0.138	0.202	0.119	0.071	0.089
	3	0.361	0.301	0.328	0.185	0.147	0.164
	5	0.357	0.316	0.335	0.188	0.153	0.169
KeyBERT	1	0.018	0.007	0.010	0.003	0.001	0.001
	3	0.009	0.010	0.009	0.004	0.001	0.002
	5	0.007	0.012	0.009	0.004	0.001	0.002
TermoPL	1	0.076	0.028	0.041	0.002	0.001	0.001
	3	0.046	0.051	0.048	0.003	0.001	0.002
	5	0.033	0.061	0.043	0.003	0.001	0.002
	ALL	0.021	0.457	0.040	0.004	0.008	0.005

Ewaluacja wewnętrzna cd.

Table 4. Evaluation of plT5kw on the set of keywords found in the training set.

Method	Rank	Micro			Macro		
		P	R	F ₁	P	R	F ₁
plT5kw	1	0.425	0.093	0.153	0.086	0.074	0.080
	3	0.415	0.212	0.281	0.165	0.158	0.161
	5	0.412	0.227	0.293	0.172	0.167	0.169

plT5kw generuje regularnie trafne słowa kluczowe, które nie występowały w zbiorze treningowym

Ewaluacja na innych zbiorach

- Czego uczy się model T5 na tym zadaniu?
- Hipoteza: T5 uczy się nie tylko przypisywania etykiet z domeny streszczeń naukowych, ale też identyfikacji i formatowania (lematyzacji, uzgadniania, pisowni itp.) potencjalnie nieciągłych fraz rzeczownikowych.
- “The results are only as good as the corpus.” (Sinclair 1991:13)
 - Czy wyniki mogą być lepsze niż korpus?
- Analiza jakościowa na:
 - Tekstach informacyjnych
 - Transkrypcjach dialogów telefonicznych

Table 5. Comparison of keywords generated for 5 news stories.

#	Short title	plT5kw	KeyBERT	TermoPL	ExtremeText
1	Polish PM meets new German Chancellor	Niemcy, Polska, Unia Europejska, Nord Stream 2, bezpieczeństwo, kryzys migracyjny, polityka	premiera morawieckiego, migracyjne energetyczne, powiedzieli czego, będzie bardzo, serii spotkań	nowy kanclerz Niemiec, instrument szantażu, kwestia migracyjna, kanclerz Niemiec, Olaf Scholza, ... (237 in total)	rosja, polska, unia_europejska
2	Gas pipe explosion on Sicily	gaz, wybuch, Włochy	wybuchu gazociągu, czterech osób, miejscowości ravanusa, jak poinformowała, pomocze przeszukiwaniu	dziesiątek osób, miejsce wybuchu, wybuch gazociągu, włoska wyspa, ... (71 in total)	paleontologia, skamieniałości, fauna_kopalna
3	Person wounded in Kielce knife attack	Kielce, policja, rana, rynek	policji kielcach, rynku 31, wyjaśniają okoliczności, narzędziem przez, letni mężczyzna	kielecki rynek, poszukiwanie napastnika, ostre narzędzie, ... (44 in total)	bezpieczeństwo, kontrola, policja
4	Omicron coronavirus variant in UK	Covid, Omikron, Wielka Brytania, koronawirus, hospitalizacja	przypadki hospitalizacji, powiedzmy 50, minister edukacji, brytania pierwsze, liczba potwierdzonych	nowy wariant, wariant Omikron, wielka Brytania, ... (126 in total)	migracja, bezpieczeństwo, transport_kolejowy
5	Tatry National Park landmark vandalized	Tatry, wandalizm, szlak turystyczny	pomalowali granitowy, narodowy film, tatrzański park, morskiego oka, sprawę gdzie	morskie oko, tatrzański park narodowy, granitowy głaz, ... (88 in total)	historia, ochrona, bezpieczeństwo

[Link](#)[Link](#)[Link](#)[Link](#)[Link](#)

Ewaluacja na DiaBiz

- **Agent:** Dzień dobry, Karolina Damczyk, **Energetyka** SA, w czym mogę pomóc?
- **Klient:** Dzień dobry. Ja się nazywam Małgorzata Porowska i dzwonię, żeby poprosić o **wyłączenie prądu**, ponieważ miałam odcięty prąd. Natomiast już dzisiaj rano uregulowałam wszystkie opłaty i chciałam się dowiedzieć, kiedy mogę się spodziewać ponownego włączenia.
- **Agent:** Dobrze, zaraz to sprawdzimy. Poproszę o imię i nazwisko właściciela umowy, i numer klienta lub numer PESEL, żebym mogła to zweryfikować. Małgorzata Porowska. Tak. Porow... Porowska, tak?
- **plT5kw:** **energia elektryczna, odłączenie prądu, wyłączenie prądu**

Ewaluacja na DiaBiz

- **A:** W czym mogę pomóc?
- **K:** Dzień dobry. Ja niedawno dzwoniłam i zgłaszałam zepsucie lodówki. Tak? Przyjęliście do realizacji i ma przyjechać kurier, który ma zabrać starą lodówkę, ale chciałabym się... chciałabym się dowiedzieć, czy dostanę zastępczy sprzęt.
- **A:** Mhm. Czy po prostu wymienimy to na jakąś inną lodówkę, tak? Dokładnie, dokładnie tak. Dobrze, to ja już to teraz sprawdzę, tylko żebym mogła to zrobić, to proszę, żeby podała mi pani numer polisy, o której rozmawiamy.
- **pIT5kw:** lodówka, polisa, sprzęt zastępczy, zgłoszenie zepsucia

Ewaluacja na DiaBiz

- **K:** No to jest po prostu... Nie ma życia teraz bez prądu. Ani pralka, ani nic. No nie mam tutaj też... No nie bardzo nawet mam od kogo pieniądze pożyczyć. Jak to można by właśnie rozłożyć? W jaki sposób?
- **A:** Dobrze. Dobrze. Ja rozumiem. Oczywiście bardzo jest mi przykro, że panią taka sytuacja spotkała. Najpierw jednak po prostu musimy załatwić formalności, więc czy ja mogłabym jeszcze raz poprosić o pani imię i nazwisko?
- **pIT5kw:** *doradztwo finansowe, pożyczka*

Formuliczne fragmenty dialogów

- **A:** Pozdrawiam. Do usłyszenia.
 - **pIT5kw:** after-affence, children
 - Po dobraniu parametru num_beams: do usłyszenia, pozdrowienie
-
- **K:** Dziękuję bardzo. Pozdrawiam serdecznie.
 - **A:** Pozdrawiam również. Do usłyszenia, miłego dnia.
 - **pIT5kw:** after-affence, children
 - Po dobraniu parametru num_beams: **do usłyszenia, miłosierdzie**

- Po 10 rozmów ze scenariuszy z 6 dziedzin tematycznych
- Zwykle zliczenie najczęstszych słów
- Następnie agregacja po wektorach FastText, np. ubezpieczenie/ubezpieczenia → ubezpieczenie
- Obiecująca metoda generowania skupień tematycznych z trafnymi, choć niekompletnymi etykietami

Table 6. Frequent keywords generated by plT5kw for phone dialogues in different customer support domains and their English translations.

Domain	Keywords PL	Keywords EN
medical	gastroskopia (4), diagnostyka medyczna (3), numer PESEL (2), diagnostyka (2), lekarz POZ (1)	gastroscopy (4), medical diagnosis (3), National Identification Number (2), diagnosis (2), GP (1)
tourism	hotel (9), turystyka (6), Afryka (2), Turcja (2), atrakcje (2)	hotel (9), tourism (6), Africa (2), Turkey (2), attractions (2)
insurance	naprawa (6), uszkodzenie (5), ekspres do kawy (3), reklamacja (3), ubezpieczenie (2)	repair (6), damage (5), coffee machine (3), complaints (3), insurance (2)
telecom	rejestracja (4), numer telefonu (3), umowa (2), abonent (2), Internet (2)	registration (4), phone number (3), contract (2), customer (2), Internet (2)
banking	identyfikacja (6), bankowość internetowa (4), logistyka (3), Bank Narodowy S.A. (3), logowanie (2)	identification (6), online banking (4), logistics (3), National Bank S.A. (3), login (2)
energy	licznik (6), zerwanie plomby (4), plomba na liczniku (8), PESEL (2), weryfikacja tożsamości (1)	meter (6), seal break (4), seal on meter (8), National Identification Number (2), identity verification (1)

Podsumowanie

- Testowany model T5 dotrenowany na streszczeniach tekstów naukowych przypisuje z reguły trafne, choć niekompletne słowa kluczowe
- Dzięki trybowi text-to-text model potrafi poprawnie formatować słowa kluczowe (lematyzować, uzgadniać, zachowywać pisownię wielką literą itp.)
- Składa nieciągłe frazy rzeczownikowe, przywraca częstszy szyk
- Zachowuje część wiedzy domenowej, ale działa zarówno ekstrakcyjnie jak też abstrakcyjnie
- Nie wymaga definiowania reguł lematyzacji itp., ale trudno też kontrolować jego działanie

Dalsze prace, ewaluacja i zastosowania

- Niewielki, ale dedykowany do tego zadania korpus esPress opracowywany przez PWr w ramach CLARIN-BIZ, zob. <http://hdl.handle.net/11321/589>
- Anotacja ekstrakcyjna, ale z wytycznymi i pomiarami zgodności, co ma znaczenie przy ocenie pokrycia. Co ciekawe, najlepszy wynik do tej pory to 0,77
- Model wielojęzyczny na korpusie [DOAJ](#)
- Testy na innych domenach