e-mentor
DWUMIESIĘCZNIK SZKOŁY GŁÓWNEJ HANDLOWEJ W WARSZAWIE
WSPÓŁWYDAWCA: FUNDACJA PROMOCJI I AKREDYTACJI KIERUNKÓW EKONOMICZNYCH

2021, nr 2 (89)

Glenc, P. (2021). Narzędzia do automatycznego streszczania tekstów w języku polskim. Stan badań naukowych i prac wdrożeniowych. *e-mentor, 2*(89), 67–77. https://doi.org/10.15219/em89.1513

Piotr Glenc

Narzędzia do automatycznego streszczania tekstów w języku polskim. Stan badań naukowych i prac wdrożeniowych

Tools for automatic summarization of texts in Polish. State of the research and implementation works

Abstract

The goal of the publication is to present the state of research and works carried out in Poland on the issue of automatic text summarization. The author describes principal theoretical and methodological issues related to automatic summary generation followed by the outline of the selected works on the automatic abstracting of Polish texts. The author also provides three examples of IT tools that generate summaries of texts in Polish (Summarize, Resoomer, and NICOLAS) and their characteristics derived from the conducted experiment, which included quality assessment of generated summaries using ROUGE-N metrics. The results of both actions showed a deficiency of tools allowing to automatically create summaries of Polish texts, especially in the abstractive approach. Most of the proposed solutions are based on the extractive method, which uses parts of the original text to create its abstract. There is also a shortage of tools generating one common summary of many text documents and specialized tools generating summaries of documents related to specific subject areas. Moreover, it is necessary to intensify works on creating the corpora of Polish-language text summaries, which the computer scientists could apply to evaluate their newly developed tools.

Keywords: text summarization, Natural Language Processing, text documents, Polish language processing, automation of knowledge acquisition

Wprowadzenie

Automatyczne streszczanie tekstów jest jednym z zagadnień rozpatrywanych na gruncie przetwarzania języka naturalnego, podobnie jak np. rozpoznawanie mowy, automatyczne tłumaczenie tekstów, analiza sentymentu czy tworzenie systemów konwersacyjnych (Maylawati i in., 2019). Wypracowane w tym obszarze techniki pozwalają na generowanie przez programy komputerowe takich streszczeń dłuższych tekstów (lub zbiorów tekstów), które zawierają ważne informacje z tekstów źródłowych, z zachowaniem ich ogólnego sensu (por. Kannaiya Raja i in., 2019). W konsekwencji ma to pozwolić na ograniczenie wysiłku ludzi oraz oszczędność czasu podczas analizy długiego dokumentu tekstowego, a także na podjęcie decyzji o ewentualnej potrzebie zapoznania się z całym dokumentem (Al-Saleh i Menai, 2018; Chetia i Hazarika, 2019). Automatyczne generowanie streszczeń ma również eliminować redundancję (powtarzalność) informacji występującą w tekstach (Rajasekaran i Varalakshmi, 2018). Wielu badaczy zauważa, że potrzeba stosowania technik automatycznego streszczania tekstów związana jest z dużym przyrostem ilości danych w postaci tekstowej, zwłaszcza w przestrzeni internetu (zob. np. Al Qassem i in., 2017; Swamy i Srinath, 2019).

Celem artykułu jest przedstawienie stanu badań i prac nad zagadnieniem automatycznego streszczania tekstów w języku polskim oraz prezentacja przykładowych narzędzi pozwalających na generowanie streszczeń polskojęzycznych dokumentów tekstowych.

Piotr Glenc, Uniwersytet Ekonomiczny w Katowicach, https://orcid.org/0000-0003-4147-4133

Opracowanie, oprócz niniejszego wprowadzenia i podsumowania, składa się z trzech części. W części pierwszej przedstawiono podstawowe zagadnienia teoretyczne i metodologiczne związane z automatycznym streszczaniem tekstów. Część druga stanowi opis wybranych prac przeprowadzonych dotychczas nad zagadnieniem automatycznego streszczania tekstów polskojęzycznych. W trzeciej części opisano wyniki eksperymentu porównawczego, polegającego na wygenerowaniu streszczeń tekstów w języku polskim z wykorzystaniem narzędzi: Summarize, Resoomer oraz NICOLAS oraz ocenie jakości wygenerowanych streszczeń.

Automatyczne streszczanie tekstów – podstawowe zagadnienia

Zadanie streszczania tekstów tradycyjnie było i jest do dziś wykonywane przez ludzi. Rozwój technologii informatycznych sprawił, że z czasem zaczęto poszukiwać możliwości automatyzacji procesu streszczania tekstów z wykorzystaniem wyspecjalizowanych narzędzi. Automatyczne streszczanie tekstów nie jest zagadnieniem nowym. Jako klasyczną z tego obszaru najczęściej wskazuje się publikację Luhna (1958). Zagadnienie to jest jednak wciąż aktualne, na co wskazuje liczba publikacji naukowych powstających w kolejnych latach. W tabeli 1 przedstawiono liczbę publikacji wydanych w latach 1994–2020, indeksowanych w bazie Scopus, zawierających frazę *automatic text summarization* w tytule, abstrakcie bądź wykazie słów kluczowych[1].

Rosnąca w kolejnych przedziałach czasowych liczba publikacji wskazuje na aktualność zagadnienia. Ponad połowa publikacji związanych z frazą *automatic text summarization*, indeksowanych w bazie Scopus, powstała w ciągu ostatnich pięciu lat.

Tabela 1

Liczba publikacji indeksowanych w bazie Scopus dotyczących automatycznego streszczania tekstów w kolejnych przedziałach czasowych (stan na 12.04.2021)

Przedział czasowy	Liczba publikacji
1994–1995	1
1996–2000	7
2001–2005	36
2006–2010	102
2011–2015	136
2016–2020	334
Łącznie	616

Źródło: opracowanie własne na podstawie danych z bazy Scopus (www.scopus.com, dostęp: 12.04.2021).

Badania literaturowe, przeprowadzone w ramach opracowywania niniejszego artykułu, objęły analizę 562 publikacji spośród 616, o których mowa powyżej[2]; 65 publikacji nie dotyczyło wprost zagadnienia streszczania tekstów. Spośród pozostałych w 382 autorzy przedstawili propozycje własnych metod automatycznego streszczania tekstów lub propozycje udoskonaleń metod już istniejących. Pozostałe 115 publikacji dotyczyło m.in. takich zagadnień jak: przegląd dotychczasowych osiągnięć w dziedzinie automatycznego streszczania tekstów, tworzenie korpusów użytecznych przy ewaluacji metod automatycznego streszczania tekstów, metody oceny jakości streszczeń.

Kierunki prac i podstawy metodologii automatycznego streszczania tekstów

Wyróżnia się dwa kierunki prac nad automatycznym generowaniem streszczeń (Fejer i Omar, 2015; García-Hernández i Ledeneva, 2013; Zhu i Li, 2012):

- podejście ekstrakcyjne (ang. *extractive*) – wybór najważniejszych fragmentów z oryginalnego tekstu (zgodnie z przyjętymi kryteriami ważności) i tworzenie streszczenia (ekstraktu) z ich wykorzystaniem[3],
- podejście abstrakcyjne/abstraktowe (ang. *abstractive*) – tworzenie gramatycznie spójnych streszczeń (abstraktów) opisujących zawartość streszczanego dokumentu z wykorzystaniem zaawansowanych technik generowania języka naturalnego, bez wykorzystywania oryginalnych fragmentów tekstu.

Obok przywołanych podejść rzadko wyodrębnia się również podejście kompresyjne, w którym zdania z oryginalnego tekstu zostają skrócone przez usunięcie mało istotnych fragmentów (zob. np. Pontes i in., 2020). Większość badaczy jednak, prezentując klasyfikację podejść, nie wyszczególnia takiego podejścia, być może zaliczając kompresję do jednego z dwóch głównych nurtów, prawdopodobnie ekstrakcyjnego.

W większości dotychczasowych prac nad automatyzacją streszczania tekstów zaproponowano podejścia oparte na ekstrakcji (Xiang i in., 2018). Spostrzeżenie to potwierdzają wyniki badań literaturowych przeprowadzonych w toku tworzenia niniejszego opracowania (baza Scopus). W 88% publikacji[4] przedstawiono techniki wykorzystujące podejście ekstrakcyjne. Tylko 7% publikacji dotyczyło podejścia abstraktowego, choć można zaobserwować wzrost popularności tego podejścia w ostatnich latach. W 3% publikacji przed-

[1] Pierwsza indeksowana w bazie publikacja związana z analizowaną frazą pochodzi z roku 1994, stąd pierwszy przedział czasowy w tabeli obejmuje krótszy okres niż pozostałe.
[2] Mimo podjęcia prób nie udało się uzyskać dostępu do pozostałych 54 publikacji, co jednak nie wpływa na wnioski z badań literaturowych w takim zakresie, w jakim przedstawiono je w niniejszym opracowaniu.
[3] Świetlicka (2010) wskazuje, że – obok ekstrakcji tekstu – należy wyróżnić ekstrakcję faktów, czyli taką, gdzie zamiast segmentów tekstu z oryginalnego dokumentu wydobywa się konkretne informacje (np. daty, miejsca, nazwiska).
[4] Spośród 382, w których przedstawiono autorskie metody streszczania tekstów lub propozycje ich udoskonaleń.

stawiono podejście hybrydowe bądź propozycje kilku metod – zarówno w podejściu ekstrakcyjnym, jak i abstraktowym. W pozostałych publikacjach streszczenie nie było przedstawiane w postaci zwartego tekstu.

Podejścia do automatycznego streszczania tekstów różnicuje się nie tylko ze względu na oczekiwany wynik końcowy (ekstrakt lub abstrakt). Biorąc pod uwagę postać materiału źródłowego, na podstawie którego tworzone jest streszczenie, można zauważyć, że streszczenie może być generowane dla pojedynczego dokumentu (*single-document*) lub dla zbioru dokumentów (*multi-document*) (por. Dash i in., 2019; Kallimani i in., 2012; Kumar i Salim, 2012).

Dotychczas w obszarze automatycznego streszczania tekstów przedstawiono użyteczność wielu technik informatycznych. Klasyczne metody opierają się na podejściu statystycznym, gdzie głównym zadaniem jest wyznaczenie kluczowych słów (fraz) i identyfikacja ich wystąpień w zdaniach (zob. np. Kallimani i in., 2012). Do tej kategorii zaliczana jest praca Luhna (1958) bazująca na założeniu, że ważne słowa są powtarzane w tekście częściej niż pozostałe i zdania zawierające najwięcej takich słów powinny tworzyć streszczenie tekstu (por. Kumar i in., 2016). W automatycznym streszczaniu tekstów często wykorzystywane są techniki uczenia maszynowego[5], zarówno nienadzorowanego (zob. np. Alguliyev i in., 2019), jak i nadzorowanego (zob. np. Morid i in., 2016; Nandhini i Balasundaram, 2013), w tym aparat sztucznych sieci neuronowych (zob. np. Anand i Wagh, 2019; Zhang i in., 2019). W przypadku tworzenia narzędzi automatycznego streszczania tekstów dla konkretnego obszaru tematycznego często użyteczne jest wzbogacenie narzędzia o wiedzę dziedzinową, określającą najważniejsze pojęcia w danym obszarze (zob. np. Fell i in., 2019). Narzędzia takie często wspomagane są tworzonymi przez ekspertów ontologiami definiującymi kluczowe w danej dziedzinie terminy i powiązania między nimi (Kumar i Salim, 2012; Mohan i in., 2016).

Nieodłącznym elementem procesu automatycznego streszczania tekstów jest ocena jakości generowanych streszczeń (Rajasekaran i Varalakshmi, 2018). Dokonuje się jej zazwyczaj na podstawie porównania streszczeń generowanych przez narzędzia informatyczne z opracowanymi przez ludzi, wzorcowymi streszczeniami (tzw. streszczeniami *gold-standard*). Ocena jakości streszczeń jest procesem kosztochłonnym, jeśli jest dokonywana przez ludzi. Dlatego często poszukuje się podejść alternatywnych. Jednym z powszechnie stosowanych sposobów oceny jest wykorzystanie miar ROUGE (zob. Lin, 2004), których obliczanie może być automatyzowane. Miary ROUGE przyjmują wartości z zakresu 0–1. Wyższa wartość oznacza większy stopień podobieństwa wygenerowanego streszczenia do streszczenia wzorcowego (Dash i in., 2019). Wykorzystuje się następujące miary ROUGE[6] (Lin, 2004; Oufaida i in., 2014):

- ROUGE-N, wyznaczaną na podstawie porównania n-gramów[7] występujących w streszczeniu wygenerowanym i wzorcowym,
- ROUGE-L, wyznaczaną na podstawie porównania najdłuższych wspólnych sekwencji słów w zdaniach, występujących w streszczeniu wygenerowanym i wzorcowym,
- ROUGE-W, wyznaczaną, podobnie jak ROUGE-L, na podstawie najdłuższych wspólnych sekwencji słów, jednak wyżej oceniającą te streszczenia, w których występują sekwencje słów następujących bezpośrednio po sobie,
- ROUGE-S, wyznaczaną na tej samej zasadzie, co ROUGE-2, przy czym z wykorzystaniem skip-bigramów[8] zamiast bigramów,
- ROUGE-SU, będącą rozszerzeniem miary ROUGE-S o dodatkowe uwzględnienie współwystępowania unigramów.

Stan prac nad automatycznym streszczaniem tekstów w języku polskim

Prace nad automatycznym streszczaniem tekstów najczęściej prowadzone są z wykorzystaniem tekstów w języku angielskim. W analizowanym zbiorze (baza Scopus) ok. 67% publikacji opisywało techniki streszczania dostosowane do tekstów anglojęzycznych. Ponadto zaprezentowano metody streszczania teks-

[5] Choć tematyka uczenia maszynowego wykracza poza zakres niniejszego artykułu, to z uwagi na wykorzystywanie w jego treści terminologii związanej z uczeniem maszynowym, zasadne wydaje się przytoczenie w tym miejscu chociaż najbardziej uproszczonych definicji terminów z tego obszaru, z zastrzeżeniem, że nie wyczerpują one zagadnienia w pełni (dalsze definicje na podstawie: Liakos i in., 2018). **Uczenie maszynowe** polega na budowaniu przez komputery modeli na podstawie dostarczonych danych i późniejszym działaniu na podstawie tych modeli. **Uczenie nadzorowane** polega na tym, że dostarczane dane mają zdefiniowane zarówno *wejście* (np. zbiór cech klienta), jak i *wyjście* (np. informacja o tym, czy klient kupił czy nie kupił oferowanego produktu). W ten sposób komputer gromadzi „doświadczenie", na podstawie którego może zbudować model, który w przyszłości pozwoli mu przewidzieć oczekiwane *wyjście* tylko na podstawie dostarczonego *wejścia* (np. czy klient o określonych cechach kupi produkt, czy nie). W **uczeniu nienadzorowanym** do systemu dostarcza się dane bez zdefiniowanego wyjścia. Komputer przetwarza takie dane w celu odnalezienia w nich pewnych ukrytych zależności i wzorców. Pozwala to np. na wykrywanie anomalii. **Sztuczne sieci neuronowe** są systemami uczenia maszynowego, w których dane przetwarzane są z wykorzystaniem mechanizmów symulujących działanie ludzkiego mózgu (sieć powiązanych, współpracujących neuronów).

[6] Podano jedynie opis słowny poszczególnych miar. Wzory pozwalające na ich obliczenie przedstawia Lin (2004).

[7] N-gram rozumiany jest jako fraza złożona z *n* słów występujących obok siebie w tekstach. 1-gram (unigram) to pojedyncze słowo, 2-gram (bigram) to dwa słowa występujące obok siebie itd.

[8] Skip-n-gramy to frazy n-gramowe współwystępujące w tekście w określonych odstępach od siebie (rozdzielone innymi frazami).

tów m.in. w językach: chińskim (Zhuang i in., 2018), arabskim (Oufaida i in., 2014) i indonezyjskim (Slamet i in., 2018). Ani jedna publikacja w analizowanym zbiorze nie dotyczyła streszczania tekstów w języku polskim[9]. Nie oznacza to jednak, że badania w tym kierunku w ogóle nie zostały podjęte. Na gruncie polskiej nauki wysiłki w tym zakresie podjęli dotychczas nieliczni badacze.

Jednym z pierwszych systemów, zaprojektowanych z zamysłem automatycznego streszczania polskojęzycznych tekstów, był PolSumm (Suszczańska i Kulików, 2003), w uaktualnionej wersji przedstawiany jako PolSum2 (Ciura i in., 2004). Autorzy zaproponowali system realizujący zadanie streszczania w podejściu ekstrakcyjnym. Koncepcja PolSum2 zakłada generowanie streszczeń w następujących etapach (Ciura i in., 2004):

- analiza tekstu wejściowego z wykorzystaniem serwera analizy lingwistycznej (ang. Linguistic Analysis Server – LAS) (Kulików, 2003),
- wybór *n* kluczowych zdań z tekstu źródłowego[10] na podstawie wagi obliczonej dla każdego zdania, gdzie wartość *n* określana jest przez użytkownika,
- linearyzacja – generowanie streszczenia wynikowego.

Istotną cechą opisywanego systemu jest zdolność analizy niektórych relacji (zależności) anaforycznych. Polega ona na identyfikacji takich zależności i zastąpieniu anafor terminami, do których się odnoszą. Autorzy podają następujący przykład zastępowania anafor, który jednocześnie ilustruje użyteczność przeprowadzonej w pierwszym etapie analizy morfologicznej tekstu[11] (Ciura i in., 2004): „Ja jestem w rzędzie. On jest daleko. Mój przyjaciel idzie do niego". „Ja jestem w rzędzie. Rząd jest daleko. Mój przyjaciel idzie do rzędu".

W 2005 roku zaproponowano algorytm streszczający teksty w języku polskim bazujący na istniejących wówczas algorytmach zmodyfikowanych na potrzeby przetwarzania tekstów polskojęzycznych (Branny i Gajęcki, 2005). Wykorzystano podejście oparte na ekstrakcji zdań z oryginalnego tekstu. Ewaluacji działania algorytmu dokonano z wykorzystaniem tekstów prasowych. Zaproponowany algorytm streszczania obejmował trzy etapy:

- Podział tekstu źródłowego na zdania i akapity.
- Opracowanie list frekwencyjnych rzeczowników, liczebników i nazw własnych – dla całego tekstu i dla poszczególnych zdań.
- Przydzielenie punktacji poszczególnym zdaniom na podstawie określonych kryteriów i wybór do streszczenia zdań najwyżej punktowanych.

Kryteria, na podstawie których dokonano punktacji i rangowania poszczególnych zdań, obejmowały: obecność „słów tematycznych"[12], pozycję zdania w akapicie, obecność nazw własnych, obecność liczebników oraz – w niektórych przypadkach – również wynik zdań poprzedzających i następujących po danym zdaniu. W pracy przyjęto stopień kompresji[13] równy 30%, co – jak przyznają sami Autorzy – nie zawsze jest dobrym rozwiązaniem i powinno być raczej uzależnione od specyfiki streszczanego tekstu niż zakładane z góry.

Podejście, jakie zaproponowali Branny i Gajęcki (2005), zostało jedynie opisane, jednak nie udostępniono narzędzia (przynajmniej na moment powstawania publikacji), w którym byłoby ono zaimplementowane. Być może dlatego jeszcze w 2007 roku system PolSum2 był określany jako jedyne dotychczas zaproponowane narzędzie pozwalające na automatyczne streszczanie tekstów w języku polskim (zob. Dudczak, 2007). Dudczak (2007) zaproponował nowe podejście wykorzystujące ekstrakcję zdań z oryginalnego tekstu, zaimplementowane w narzędziu Lakon. W pracy opisane zostały metody wykorzystujące:

- informację o położeniu zdań – w pierwszej kolejności do streszczenia wybierane są pierwsze zdania z poszczególnych akapitów, następnie drugie itd. – do momentu osiągnięcia zadanej długości streszczenia,
- wagi zdań będące wynikiem sumowania wag występujących w nich słów kluczowych, wyznaczonych przez miary tf-idf oraz Okapi BM25[14],
- wagi zdań obliczone na podstawie wystąpień w nich słów należących do uprzednio wyznaczonych łańcuchów leksykalnych[15] (tworzonych na podstawie rzeczowników).

[9] Choć w niektórych publikacjach przedstawiono metody niezależne od języka, potencjalnie użyteczne przy generowaniu streszczeń w języku polskim (zob. np. Moen i in., 2016), to nie opisano eksperymentów z wykorzystaniem tekstów polskojęzycznych.

[10] Konkretna metoda wyboru zdań nie została opisana. System ma służyć jako framework, w którym mogą być zaimplementowane różne metody streszczania tekstów.

[11] Stosowane są tutaj odmiany formy bazowej „rząd" odnoszącej się do szeregu miejsc, a nie organu władzy państwowej.

[12] Rzeczowników znajdujących się w pierwszym akapicie tekstu i w pierwszym zdaniu drugiego akapitu.

[13] Wskaźnik określający proporcję długości streszczenia do długości oryginalnego tekstu.

[14] Miara tf-idf określa istotność danego słowa (termu) w zależności od tego, jak często pojawia się ono w danym dokumencie i w jak wielu dokumentach z analizowanego zbioru ono występuje. Przyjmuje ona wyższe wartości dla słów pojawiających się często w danym dokumencie i rzadko w innych dokumentach. Problemem przy ocenianiu istotności dokumentów z wykorzystaniem tej miary jest fakt, że dłuższe dokumenty uzyskują wyższe wyniki tf-idf, co ma równoważyć miara Okapi BM25 uwzględniająca m.in. średnią długość dokumentów w analizowanym zbiorze (zob. Dudczak i in., 2008).

[15] Więcej informacji na temat procedury tworzenia łańcuchów leksykalnych można znaleźć np. w pracy Dudczaka i in. (2008).

Dla celów porównawczych w narzędziu Lakon zaimplementowano też dodatkowe metody:

- losowy wybór zdań do streszczenia,
- wybór *n* pierwszych zdań z tekstu źródłowego.

Zaimplementowane metody uzyskały od 42% do 53% zgodności ze streszczeniem wzorcowym utworzonym na bazie streszczeń opracowanych przez grupę wolontariuszy.

Inną próbę opracowania narzędzia streszczającego teksty w języku polskim opisała Świetlicka (2010). Na potrzeby eksperymentów w ramach pracy przygotowany został korpus zawierający polskojęzyczne teksty prasowe i ich streszczenia. Autorka, stosując podejście oparte na ekstrakcji zdań, wykorzystała szereg algorytmów uczenia maszynowego. Do opisu poszczególnych zdań wykorzystano 26 charakterystyk, m.in.: centralność zdania[16], tf-idf, odsetek słów zaczynających się od wielkiej litery, długość zdania, pozycję w tekście, podobieństwo do tytułu. W efekcie badań opracowane zostało narzędzie Summarizer, które w wielu późniejszych pracach było uznawane za wysoce skuteczne i wykorzystywane jako punkt odniesienia dla nowo proponowanych narzędzi (zob. np. Kopeć, 2015; Kopeć, 2018; Ozimek, 2020).

Prace w kolejnych latach ukierunkowane były na doskonalenie i poszukiwanie nowych metod streszczania tekstów w języku polskim. Aby umożliwić ewaluację nowo tworzonych metod, Ogrodniczuk i Kopeć (2014) opracowali korpus streszczeń polskojęzycznych tekstów prasowych (Polski Korpus Streszczeń). Korpus składa się z 569 artykułów prasowych oraz ich streszczeń stworzonych przez 11 osób, zarówno w podejściu ekstrakcyjnym (dla każdego artykułu), jak i abstraktowym (dla 154 artykułów). Streszczenia tworzone były w trzech wariantach zawierających odpowiednio: 20%, 10% i 5% liczby słów w oryginalnym tekście. Streszczenia poszczególnych artykułów były opracowywane przez pięć różnych osób. Takie założenia spowodowały, że ostatecznie powstał korpus składający się z 10 845 streszczeń (8535 ekstraktów i 2310 abstraktów). Korpus został udostępniony do pobrania w internecie (http://zil.ipipan.waw.pl/PolishSummariesCorpus).

Jassem i Pawluczuk (2015) do streszczania tekstów zastosowali sztuczne sieci neuronowe, przeprowadzając szereg eksperymentów z wykorzystaniem różnych cech charakteryzujących poszczególne zdania tekstu. Oprócz cech wskazanych już we wcześniejszych pracach (takich jak np. tf-idf, centralność zdania, długość zdania) Autorzy zaproponowali również zestaw innych cech, wcześniej niewykorzystywanych, zwłaszcza wynikających z rozpoznawania jednostek nazwanych[17] w zdaniach, m.in. liczbę jednostek nazwanych odnoszących się do osób, organizacji, miejsc czy dat.

Kolejne narzędzie pozwalające na automatyczne streszczanie polskojęzycznych tekstów zaproponował Kopeć (2015). Narzędzie nazwane EMILY wykorzystywało podejście ekstrakcyjne, a jednostki tekstu mogły być wybierane do streszczenia na dwa sposoby:

- wybór pełnych zdań (EMILY-S),
- wybór mniejszych jednostek tekstu, z których każda zawiera czasownik[18] (EMILY-C).

Istotną cechą zastosowanego podejścia było wykorzystanie analizy koreferencji, czyli odwołań do tego samego obiektu w różnych wyrażeniach w tekście[19]. Testów narzędzia dokonano wykorzystując Polski Korpus Streszczeń (Ogrodniczuk i Kopeć, 2014). Jak przyznaje sam Autor, EMILY nie osiągnęła wyników znacząco lepszych od większości wcześniej opracowanych narzędzi. Wariant EMILY-S pozwolił uzyskać nieznacznie lepsze wyniki niż EMILY-C. Testy wskazały na wysoką (na tle innych testowanych rozwiązań) jakość streszczeń generowanych przez narzędzie, jakie zaproponowała Świetlicka (2010). Co ciekawe – wysoce skuteczne okazało się jedno z najprostszych możliwych rozwiązań, czyli wybór początkowego fragmentu tekstu o długości zależnej od oczekiwanego stopnia kompresji. Należy mieć jednak na uwadze specyfikę tekstów wykorzystywanych do eksperymentów – artykułów prasowych. Te najczęściej są zbudowane właśnie tak, że najistotniejsze informacje podaje się na początku, a następnie rozwija bardziej szczegółowo w dalszych fragmentach tekstu. W 2018 roku Autor zaproponował nowe, doskonalsze narzędzie, NICOLAS (Kopeć, 2018), oparte – podobnie jak EMILY – na analizie koreferencji. NICOLAS wykorzystuje algorytmy uczenia maszynowego, opisując poszczególne zdania – oprócz cech wynikających z identyfikacji zależności koreferencyjnych – również cechami takimi jak np. znak, jakim kończy się zdanie (czy kończy się kropką lub znakiem zapytania), pozycja zdania w tekście, długość zdania. Narzędzie NICOLAS zostało udostępnione w internecie (http://zil.ipipan.waw.pl/Nicolas).

Pierwsze próby opracowania narzędzia streszczającego teksty polskojęzyczne w podejściu abstraktowym opisano dopiero w roku 2020[20] (Ozimek, 2020). Do streszczania tekstów zastosowano metody głębokiego uczenia (sztuczne sieci neuronowe), z wykorzystaniem

[16] Wskaźnik określający podobieństwo danego zdania do innych zdań w tekście.

[17] Ang. Named Entities, czyli konkretne obiekty, pojawiające się w tekście, takie jak np. osoby, organizacje, miejsca (Glenc, 2020).

[18] Rozdzielonych w ramach zdania spójnikiem: *i*, *albo*, *lub* lub jednym ze znaków: przecinek, średnik, dwukropek, nawias, dywiz, półpauza, cudzysłów.

[19] Koreferencja występuje np. w tekście: „Ola poszła do sklepu. Dziewczynka wzięła ze sobą torbę na zakupy". Słowa „Ola" i „dziewczynka" odnoszą się do tego samego obiektu.

[20] Zgodnie z deklaracją Autora. Przeprowadzone w toku opracowywania niniejszej publikacji badania literaturowe również nie wykazały istnienia wcześniejszych publikacji opisujących wykorzystanie podejścia abstraktowego do streszczania polskojęzycznych tekstów.

modelu Seq2Seq[21] z mechanizmem atencji[22]. Podejście to było inspirowane podobnymi, wcześniej opisanymi próbami automatycznego streszczania tekstów w języku angielskim. W ramach pracy przygotowany został także autorski zbiór polskojęzycznych tekstów – artykułów zaczerpniętych z różnych stron internetowych wraz z ich streszczeniami[23]. Przeprowadzona ewaluacja wykazała niższą skuteczność zaproponowanego narzędzia na tle wcześniej opracowanych narzędzi streszczających teksty polskojęzyczne. Należy jednak mieć na uwadze, że dokonano tu porównania narzędzi wykorzystujących podejście ekstrakcyjne z pionierskim narzędziem wykorzystującym podejście abstraktowe, co mimo wszystko pozwala uznać uzyskane wyniki za obiecujące, a zaproponowane narzędzie za dobrą podstawę do dalszych prac.

W tabeli 2 dokonano syntetycznego opisu podejmowanych na przestrzeni kolejnych lat prac nad narzędziami streszczającymi w sposób automatyczny teksty polskojęzyczne.

Istnieją również ogólnodostępne narzędzia pozwalające na automatyczne streszczanie tekstów polskojęzycznych, które dotychczas nie zostały opisane w literaturze naukowej. Do tej grupy można zaliczyć internetowe narzędzia Summarize (https://ws.clarin-pl.eu/summarize.shtml) i Resoomer (https://resoomer.com/pl/). Zostały one wykorzystane w eksperymencie opisanym w dalszej części niniejszego artykułu, co pozwoliło na przynajmniej częściowe poznanie ich specyfiki i funkcjonalności.

Analizując postęp prac w obszarze automatycznego streszczania tekstów w języku polskim można zauważyć następujące ograniczenia:

- Do ewaluacji proponowanych rozwiązań wykorzystywano artykuły prasowe[24].
- Zaproponowane rozwiązania pozwalają głównie na streszczanie pojedynczych dokumentów. Niektóre rozwiązania umożliwiają co prawda generowanie streszczeń wielu dokumentów, jednak nie zostały wystarczająco przetestowane pod tym kątem.
- Zaproponowane rozwiązania nie uwzględniają tematyki dokumentów, zostały opracowane z zamysłem streszczania tekstów ze wszystkich dziedzin.
- Dopiero niedawno rozpoczęto prace nad generowaniem streszczeń w podejściu abstraktowym i są one jeszcze w początkowej fazie.

Wskazane ograniczenia wytyczają potencjalne kierunki dalszych prac. Wysiłki powinny być podjęte zwłaszcza na gruncie informatyki, przy zaangażowaniu ekspertów z obszaru lingwistyki. Należy jednak zauważyć, że prace nad zagadnieniem automatycznego streszczania tekstów nie muszą zamykać się jedynie w kręgu badaczy reprezentujących wspomniane dyscypliny. Zasadne wydaje się bowiem tworzenie

Tabela 2

Wybrane publikacje dotyczące automatycznego streszczania tekstów w j. polskim

Nazwa narzędzia/systemu	Praca	Podejście	Metoda/technika
PolSumm; PolSum2	Suszczańska i Kulików, 2003; Ciura i in., 2004	Ekstrakcja	Nie została opisana konkretna metoda streszczania.
	Branny i Gajęcki, 2005	Ekstrakcja	Rangowanie zdań na podstawie przyjętych kryteriów.
Lakon	Dudczak, 2007	Ekstrakcja	Różne metody.
Summarizer	Świetlicka, 2010	Ekstrakcja	Uczenie maszynowe.
	Gramacki i Gramacki, 2011	Ekstrakcja	Przekształcenia macierzy *term-sentence*.
	Jassem i Pawluczuk, 2015)	Ekstrakcja	Sztuczne sieci neuronowe.
EMILY	Kopeć, 2015	Ekstrakcja	Uczenie maszynowe, analiza koreferencji.
NICOLAS	Kopeć, 2018	Ekstrakcja	Uczenie maszynowe, analiza koreferencji.
NLPer	Ozimek, 2020	Abstrakcja	Głębokie uczenie z wykorzystaniem modelu Seq2Seq i mechanizmu atencji.

Źródło: opracowanie własne.

[21] Model polegający na przekształceniu wejściowej sekwencji (w opisywanym przypadku – sekwencji słów) w sekwencję wyjściową wykorzystywany również m.in. do automatycznego generowania tłumaczeń tekstów.

[22] Mechanizm użyteczny przy dłuższych sekwencjach. Pozwala za pomocą tzw. wag atencji określić istotność poszczególnych elementów sekwencji na kolejnych etapach generowania streszczenia.

[23] Za streszczenie uznany został tytuł artykułu oraz jego nagłówek (część wprowadzająca). Nie angażowano ludzi w proces tworzenia streszczeń.

[24] Jest to z jednej strony zrozumiałe z uwagi na fakt, że tego typu teksty są najczęściej pisane poprawnym językiem, usystematyzowane i łatwo dostępne. Rodzi to jednak pytania o skuteczność proponowanych technik przy streszczaniu innego rodzaju tekstów.

narzędzi wyspecjalizowanych, dostosowanych do streszczania tekstów dotyczących konkretnej tematyki czy konkretnego rodzaju dokumentów powstających np. w organizacjach. Rola ekspertów z określonych dziedzin może być tutaj kluczowa – zarówno na etapie tworzenia nowych rozwiązań, jak i przy ich ocenie.

Narzędzia generujące streszczenia tekstów w języku polskim – opis i ocena jakości

W niniejszej części opracowania przedstawiono wyniki eksperymentu przeprowadzonego z wykorzystaniem trzech narzędzi pozwalających na automatyczne streszczanie tekstów w języku polskim: Summarize, Resoomer i NICOLAS. Dwa pierwsze nie zostały opisane w literaturze naukowej[25], stąd zasadne jest ich wykorzystanie w eksperymencie i dokonanie opisu na tej podstawie.

Cel eksperymentu i pytania badawcze

Celem eksperymentu było zbadanie funkcjonalności wymienionych narzędzi oraz ocena jakości generowanych przez nie streszczeń. Postawiono następujące pytania badawcze:

- Czy narzędzia generują takie same streszczenia dla tych samych tekstów?
- Które z podejść (ekstrakcja/abstrakcja) jest wykorzystywane w narzędziach Summarize i Resoomer?
- Jaka jest jakość streszczeń generowanych przez narzędzia Summarize i Resoomer dla tekstów z Polskiego Korpusu Streszczeń?
- Jaka jest jakość streszczeń generowanych przez narzędzia Summarize, Resoomer i NICOLAS dla zapisów tekstowych wybranych wypowiedzi z debaty sejmowej?

Charakterystyka narzędzi

- **Summarize** – narzędzie internetowe pozwalające na generowanie streszczeń pojedynczych dokumentów tekstowych lub ich zbiorów. Jedynym parametrem, który może określić użytkownik przed wygenerowaniem streszczenia jest narzędzie, jakie ma być zastosowane do analizy morfologicznej tekstu, której wyniki są wykorzystywane w procesie generowania streszczenia (Morfeusz 1 lub Morfeusz 2).
- **Resoomer** – narzędzie internetowe pozwalające na generowanie streszczeń pojedynczych tekstów w różnych językach, w tym w języku polskim. Narzędzie dostosowane jest do streszczania tekstów argumentacyjnych. Użytkownik ma możliwość wyboru rodzaju streszczenia: automatyczne (dobierana optymalna długość streszczenia), ręczne (długość streszczenia określana przez użytkownika), zoptymalizowane (w streszczeniu zachowane są słowa kluczowe i tematy wskazane przez użytkownika). Możliwe jest również wyświetlenie wyników analizy tekstu, gdzie zdania uznane za ważne zostają podkreślone.
- **NICOLAS** – narzędzie zostało opisane w niniejszym tekście. W ramach eksperymentu wykorzystano wersję .jar narzędzia, która uruchamiana jest z poziomu wiersza poleceń.

Organizacja eksperymentu

Realizację eksperymentu podzielono na dwa etapy. W ramach pierwszego etapu wykorzystano wybrane teksty z Polskiego Korpusu Streszczeń i wygenerowano ich streszczenia z wykorzystaniem narzędzi Summarize i Resoomer[26]. W ramach drugiego etapu wykorzystano zapisy wypowiedzi wygłoszonych podczas jednego z posiedzeń sejmu. Przygotowano dla nich streszczenia wzorcowe i porównano je ze streszczeniami wygenerowanymi przez narzędzia: Summarize, Resoomer i NICOLAS.

Wykorzystane teksty źródłowe i streszczenia wzorcowe

W pierwszym etapie eksperymentu tekstami źródłowymi było 14 artykułów prasowych z Polskiego Korpusu Streszczeń. W tabeli 3 przedstawiono charakterystykę wykorzystanych tekstów.

Ponieważ jedno z narzędzi wykorzystywanych w eksperymencie (Summarize) nie pozwala na określenie oczekiwanej długości streszczenia, konieczne było ustalenie, w jaki sposób zapewniona zostanie porównywalna długość streszczeń generowanych przez poszczególne narzędzia. Zauważono, że długość streszczeń generowanych przez Summarize często zawiera się w przedziale od 30% do 40% długości tekstu źródłowego. Zdecydowano, że do eksperymentów zostanie założony 40-procentowy stopień kompresji, tj. zostaną wykorzystane takie teksty z korpusu, dla których długość streszczeń generowanych przez Summarize jest możliwie najbliższa tej wartości, zaś w narzędziach Resoomer i NICOLAS taki stopień kompresji zostanie zadany. Założenie to pociągnęło jednak za sobą konieczność utworzenia nowych streszczeń wzorcowych będących punktem odniesienia w procesie oceny jakości generowanych streszczeń. Polski Korpus Streszczeń nie zawiera bowiem streszczeń o stopniu kompresji wynoszącym 40%. Na potrzeby eksperymentu streszczenia takie zostały opracowane z wykorzystaniem 20-procento-

[25] Z informacji pozyskanych od przedstawicieli konsorcjum CLARIN-PL udostępniającego narzędzie Summarize wynika, że narzędzie to korzysta z mechanizmów platformy MEAD (Radev i in., 2004) pozwalającej na tworzenie rozwiązań streszczających teksty w różnych językach, jednak samo narzędzie Summarize nie zostało dotychczas opisane w żadnej publikacji.

[26] Na tym etapie nie zastosowano narzędzia NICOLAS, gdyż teksty z Polskiego Korpusu Streszczeń były wykorzystywane przy jego uczeniu, przez co uzyskane wyniki mogłyby być zawyżone.

Tabela 3
Charakterystyka tekstów wykorzystanych w eksperymencie (I etap)

Lp.	Identyfikator w korpusie	Obszar tematyczny (sekcja)	Liczba słów w oryginalnym tekście
1.	199704210011	Nauka i Technika	1007
2.	199704220018	Kultura	1250
3.	199704300031	Prawo	3340
4.	199801020079	Prawo	2520
5.	199801030148	Publicystyka, Opinie	1504
6.	199801200106	Ekonomia	1116
7.	199801260047	Sport	1198
8.	199901230088	Sport	1367
9.	199911200030	Kraj	1110
10.	200001030029	Kultura	1304
11.	200001060053	Publicystyka, Opinie	1035
12.	200012130100	Publicystyka, Opinie	1987
13.	200108180109	Kraj	1290
14.	200202210054	Kultura	1379

Źródło: opracowanie własne na podstawie danych w Polskim Korpusie Streszczeń. „The Polish Summaries Corpus", M. Ogrodniczuk i M. Kopeć, 2014. W N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, i S. Piperidis, (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, *LREC 2014* (s. 3712–3715). Rejkiawik, Islandia. European Language Resources Association (ELRA)

wych streszczeń z korpusu w następujący sposób: do streszczenia wzorcowego wybierane były fragmenty 20-procentowych streszczeń w kolejności ich umieszczenia w korpusie – najpierw fragmenty wskazane przez pierwszą osobę, później dodatkowe fragmenty wskazane przez drugą osobę itd., aż do uzyskania streszczenia o założonej długości[27].

W drugim etapie eksperymentu jako teksty źródłowe wykorzystano zapisy 10 wypowiedzi wygłoszonych 15 kwietnia 2021 podczas posiedzenia sejmu. Teksty zostały oczyszczone z fragmentów niestanowiących części wypowiedzi (takich jak np. głosy z sali). W tabeli 4 przedstawiono charakterystykę wykorzystanych tekstów.

Tabela 4
Charakterystyka tekstów wykorzystanych w eksperymencie (II etap)

Lp.	Autor wypowiedzi	Liczba słów w wypowiedzi
1.	M. Wąsik	249
2.	S. Kaleta	634
3.	W. Kraska	430
4.	K. Bosak	280
5.	A. Niedzielski	2601
6.	C. Grabarczyk	401
7.	S. Krajewski	467
8.	M. Falej	574
9.	J. Mucha	635
10.	I. M. Kozłowska	778

Źródło: opracowanie własne na podstawie tekstów pobranych z Systemu Informacyjnego Sejmu (www.sejm.gov.pl, pobrano 30.04.2021).

[27] W przypadku jednego z tekstów (200001060053) wykorzystanie wszystkich streszczeń udostępnionych w korpusie nie wystarczyło, aby uzyskać 40-procentowy stopień kompresji, dlatego konieczne było wskazanie dodatkowych fragmentów, aby utworzyć wzorcowe streszczenie.

Podobnie jak w pierwszym etapie eksperymentu, założono 40-procentowy stopień kompresji. Streszczenia wzorcowe zostały przygotowane przez autora niniejszego opracowania z wykorzystaniem fragmentów oryginalnych wypowiedzi.

Zastosowane miary jakości

Do oceny jakości poszczególnych streszczeń wykorzystano miary: ROUGE-1, ROUGE-2 oraz ROUGE-3[28].

Wyniki i wnioski

W tabeli 5 przedstawiono średnie wartości poszczególnych miar ROUGE-N wyznaczonych w ramach pierwszego etapu eksperymentu (teksty z Polskiego Korpusu Streszczeń).

Wyższe wartości poszczególnych miar ROUGE-N uzyskały streszczenia wygenerowane w narzędziu Summarize. Należy jednak zauważyć, że narzędzie Resoomer dostosowane jest do streszczania tekstów argumentacyjnych, a nie wszystkie artykuły wykorzystane do ewaluacji miały taki charakter. Uzyskane wyniki nie pozwoliły zatem na jednoznaczne wskazanie, które z narzędzi generuje streszczenia wyższej jakości. Wyniki należało raczej uznać za porównywalne, zaś dla dokonania bardziej precyzyjnej oceny konieczne było wykorzystanie tekstów argumentacyjnych. Wymaganie to zostało spełnione w drugim etapie eksperymentu.

W tabeli 6 przedstawiono średnie wartości poszczególnych miar ROUGE-N wyznaczonych w ramach drugiego etapu eksperymentu (wypowiedzi z debaty sejmowej).

Najwyższe wartości poszczególnych miar ROUGE-N uzyskały streszczenia wygenerowane w narzędziu Summarize, choć można zauważyć, że żadne z narzędzi nie generowało streszczeń o jakości wyraźnie niższej niż pozostałe. Wartości ROUGE-N były niższe niż w pierwszym etapie eksperymentu, co wskazuje na to, że rodzaj streszczanych tekstów może mieć wpływ na jakość generowanych streszczeń i zagadnienie to warto uczynić przedmiotem dalszych badań. Pozostałe wnioski z przeprowadzonego eksperymentu są następujące:

- Wszystkie analizowane narzędzia stosują podejście oparte na ekstrakcji fragmentów z oryginalnego tekstu.
- Narzędzia nie generują takich samych streszczeń dla tych samych tekstów źródłowych.
- Narzędzie Resoomer posiada mechanizmy pozwalające na kompresję zdań (w niektórych tekstach ze zdań usuwane były dopiski ujęte w nawiasy).

Jako ograniczenia opisanego eksperymentu należy uznać: niedużą liczbę wykorzystanych tekstów oraz proces tworzenia streszczeń wzorcowych. W pierwszym etapie streszczenia dla stopnia kompresji wynoszącego 40% utworzono na bazie kilku streszczeń przygotowanych z założeniem stopnia kompresji wynoszącego 20%. W drugim etapie streszczenia wzorcowe były tworzone tylko przez jedną osobę. Ograniczenia te wynikają z braku profesjonalnych korpusów streszczeń polskojęzycznych tekstów (poza Polskim Korpusem Streszczeń) i wyraźnie wskazują na potrzebę tworzenia takich korpusów, zwłaszcza przy zaangażowaniu ekspertów z dziedzin, których dotyczą streszczane teksty.

Podsumowanie

Automatyczne streszczanie tekstów jest zagadnieniem aktualnym, godnym uwagi nie tylko w obszarze nauk technicznych. Dotychczas stosunkowo niewielką uwagę poświęcono rozwiązaniom tego typu dostosowanym do przetwarzania tekstów w języku polskim. Niski stopień zaawansowania prac dotyczy zwłaszcza wykorzystania podejścia abstraktowego do streszczania tekstów.

Tabela 5

Średnie wartości poszczególnych miar ROUGE-N (I etap eksperymentu)

	ROUGE-1	ROUGE-2	ROUGE-3
Summarize	0,607	0,435	0,392
Resoomer	0,605	0,428	0,387

Źródło: opracowanie własne.

Tabela 6

Średnie wartości poszczególnych miar ROUGE-N (II etap eksperymentu)

	ROUGE-1	ROUGE-2	ROUGE-3
Summarize	0,588	0,431	0,376
Resoomer	0,581	0,418	0,365
NICOLAS	0,576	0,414	0,355

Źródło: opracowanie własne.

[28] Obliczone z wykorzystaniem biblioteki *rouge-score* języka Python (https://pypi.org/project/rouge-score/, dostęp: 28.04.2021).

Przeprowadzony eksperyment z wykorzystaniem tekstów w języku polskim i narzędzi Summarize, Resoomer i NICOLAS wykazał, że każde z tych narzędzi wykorzystuje do tworzenia streszczeń mechanizm ekstrakcji zdań z oryginalnego tekstu. Dla takich samych tekstów wejściowych narzędzia generowały różne wyniki. Pod względem miar ROUGE-N streszczenia najwyższej jakości był generowane przez narzędzie Summarize. Ograniczeniem związanym z wykorzystaniem tego narzędzia do ewentualnej dalszej ewaluacji jest brak możliwości wyboru oczekiwanej długości generowanego streszczenia.

Opisane w niniejszym artykule badania pozwoliły na identyfikację takich obszarów związanych z automatycznym streszczaniem tekstów polskojęzycznych, w których widoczne są pewne niedoskonałości lub niewielki postęp prac. Za główne wyzwania można uznać: opracowanie metod streszczania tekstów w podejściu abstraktowym oraz metod dostosowanych do streszczania tekstów dotyczących określonej tematyki, a także tworzenie nowych korpusów tekstów i ich streszczeń zawierających teksty inne niż artykuły prasowe, które mogłyby służyć do ewaluacji nowo tworzonych metod.

Bibliografia

Al Qassem, L. M., Wang, D., Al Mahmoud, Z., Barada, H., Al-Rubaie, A. i Almoosa, N. I. (2017). Automatic Arabic summarization: A survey of methodologies and systems. *Procedia Computer Science*, *117*, 10–18. https://doi.org/10.1016/j.procs.2017.10.088

Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A. i Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. *Expert Systems*, *36*(1), e12340. https://doi.org/10.1111/exsy.12340

Al-Saleh, A. i Menai, M. E. B. (2018). Solving multi-document summarization as an orienteering problem. *Algorithms*, *11*(7), 96. https://doi.org/10.3390/a11070096

Anand, D. i Wagh, R. (2019). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.11.015

Branny, E. i Gajęcki, M. (2005). Text summarizing in Polish. *Computer Science*, *7*, 31–48.

Chetia, G. i Hazarika, G. C. (2019). Single document text summarization of a resource-poor language using an unsupervised technique. *International Journal of Engineering and Advanced Technology*, *9*(1), 6278–6281. https://doi.org/10.35940/ijeat.a2250.109119

Ciura, M., Grund, D., Kulików, S., Suszczańska, N. i Okatan, A. (2004). A system to adapt techniques of text summarizing to Polish. In A. Ocatan (red.), *Computational Intelligence* (s. 117–120). Proceedings of the International Conference on Computational Intelligence. 17–19 grudnia, Istambuł, Turcja.

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S. i Chakraborty, A. (2019). Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–28. https://doi.org/10.1145/3359274

Dudczak, A. (2007). *Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego* (praca magisterska). Politechnika Poznańska. http://www.cs.put.poznan.pl/dweiss/research/lakon/publications/thesis.pdf

Dudczak, A., Stefanowski, J. i Weiss, D. (2008). *Automatyczna selekcja zdań dla tekstów prasowych w języku polskim*. Instytut Informatyki Politechniki Poznańskiej, Raport Techniczny RA-03/08. http://www.cs.put.poznan.pl/dweiss/research/lakon/publications/techreport.pdf

Fejer, H. N. i Omar, N. (2015). Automatic multi-document Arabic text summarization using clustering and keyphrase extraction. *Journal of Artificial Intelligence*, *8*(1), 1–9. https://doi.org/10.3923/JAI.2015.1.9

Fell, M., Cabrio, E., Gandon, F. i Giboin, A. (2019). Song lyrics summarization inspired by audio thumbnailing. *Proceedings of International Conference Recent Advances in Natural Language Processing, RANLP* (s. 328–337), 2–4 sierpnia, Warna, Bułgaria. https://doi.org/10.26615/978-954-452-056-4_038

García-Hernández, R. A. i Ledeneva, Y. (2013). Single extractive text summarization based on a genetic algorithm. W J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. S. Rodríguez i G. S. di Baja (Eds.), *Pattern recognition* (s. 374–383). 5th Mexican Conference, MCPR 2013. 26–29 czerwca, Berlin, Niemcy. Springer. https://doi.org/10.1007/978-3-642-38989-4_38

Glenc, P. (2020). Automatyzacja analizy cyfrowej komunikacji organizacji, W B. Filipczyk i J. Gołuchowski (red.), *Cyfrowa komunikacja organizacji* (s. 108–125). Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Gramacki, J. i Gramacki, A. (2011). Automatyczne tworzenie podsumowań tekstów metodami algebraicznymi. *Pomiary Automatyka Kontrola*, *57*(7), 751–755.

Jassem, K. i Pawluczuk, Ł. (2015). Automatic summarization of Polish news articles by sentence selection. W M. Ganzha, L. Maciaszek i M. Paprzycki (red.), *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)* (s. 337–341). 13–16 września, Łódź, Polska. https://doi.org/10.15439/2015f186

Kallimani, J. S., Srinivasa, K. G. i Reddy, B. E. (2012). Summarizing news paper articles: Experiments with ontology-based, customized, extractive text summary and word scoring. *Cybernetics and Information Technologies*, *12*(2), 34–50. https://doi.org/10.2478/cait-2012-0011

Kannaiya Raja, N., Bakala, N. i Suresh, S. (2019). NLP: Text summarization by frequency and sentence position methods. *International Journal of Recent Technology and Engineering*, *8*(3), 3869–3872. https://doi.org/10.35940/ijrte.c5088.098319

Kopeć, M. (2015). Coreference-based content selection for automatic summarization of Polish news. W *Selected problems in information technologies* (s. 23–46). Information Technologies: Research and their Interdisciplinary Applications 2015. 22–24 października, Warszawa, Polska. ITRIA 2015. Conference Proceedings.

Kopeć, M. (2018). *Summarization of Polish press articles using coreference* (praca doktorska). Instytut Podstaw Informatyki Polskiej Akademii Nauk. http://zil.ipipan.waw.pl/MateuszKopec?action=AttachFile&do=view&target=m.kopec-phd-thesis.pdf

Kulików, S. (2003). Implementacja serwera analizy lingwistycznej dla systemu Theos – translatora tekstu na język migowy. *Studia Informatica*, *24*(3), 171–178.

Kumar, Y. J. i Salim, N. (2012). Automatic multi document summarization approaches. *Journal of Computer Science*, *8*(1), 133–140. https://doi.org/10.3844/JCSSP.2012.133.140

Kumar, Y. J., Goh, O. S., Basiron, H., Choon, N. H. i Suppiah, P. C. (2016). A review on automatic text summarization approaches. *Journal of Computer Science*, *12*(4), 178–190. https://doi.org/10.3844/jcssp.2016.178.190

Liakos, K. G., Busato, P., Moshou, D., Pearson, S. i Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Lin, C. (2004). ROUGE: A package for automatic evaluation of summaries. W M. Moens i S. Szpakowicz (red.), *Text summarization branches out: Proceedings of the ACL-04Workshop* (s. 74–81). 25–26 lipca, Barcelona, Hiszpania. https://www.aclweb.org/anthology/W04-1013.pdf

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. https://doi.org/10.1147/RD.22.0159

Maylawati, D. S., Kumar, Y. J., Kasmin, F. B. i Ramdhani, M. A. (2019). An idea based on sequential pattern mining and deep learning for text summarization. *Journal of Physics: Conference Series*, *1402*(7), 077013. IOP Publishing. https://doi.org/10.1088/1742-6596/1402/7/077013

Moen, H., Peltonen, L. M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T. i Salanterä, S. (2016). Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, *67*, 25–37. https://doi.org/10.1016/j.artmed.2016.01.003

Mohan, M. J., Sunitha, C., Ganesh, A. i Jaya, A. (2016). A study on ontology based abstractive summarization. *Procedia Computer Science*, *87*, 32–37. https://doi.org/10.1016/J.PROCS.2016.05.122

Morid, M. A., Fiszman, M., Raja, K., Jonnalagadda, S. R. i Del Fiol, G. (2016). Classification of clinically useful sentences in clinical evidence resources. *Journal of Biomedical Informatics*, *60*, 14–22. https://doi.org/10.1016/j.jbi.2016.01.003

Nandhini, K. i Balasundaram, S. R. (2013). Improving readability through extractive summarization for learners with reading difficulties. *Egyptian Informatics Journal*, *14*(3), 195–204. https://doi.org/10.1016/J.EIJ.2013.09.001

Ogrodniczuk, M. i Kopeć, M. (2014). The Polish Summaries Corpus. W N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, i S. Piperidis, (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, *LREC 2014* (s. 3712–3715). Rejkiawik, Islandia. European Language Resources Association (ELRA).

Oufaida, H., Nouali, O. i Blache, P. (2014). Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. *Journal of King Saud University – Computer and Information Sciences*, *26*(4), 450–461. https://doi.org/10.1016/j.jksuci.2014.06.008

Ozimek, W. (2020). *Automatic summary of texts in Polish* (praca magisterska). Uniwersytet Jagielloński w Krakowie.

Pontes, E. L., Huet, S., Torres-Moreno, J. M. i Linhares, A. C. (2020). Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, *125*, 101763. https://doi.org/10.1016/j.datak.2019.101763

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A. i Zhang, Z. (2004). MEAD – a platform for multidocument multilingual text summarization. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004).* Lizbona, Portugalia. https://doi.org/10.7916/D8MG7XZT

Rajasekaran, A. i Varalakshmi, R. (2018). Review on automatic text summarization. *International Journal of Engineering & Technology*, *7*(2.33), 456–460. https://doi.org/10.14419/IJET.V7I2.33.14210

Slamet, C., Atmadja, A. R., Maylawati, D. S., Lestari, R. S., Darmalaksana, W. i Ramdhani, M. A. (2018). Automated text summarization for Indonesian article using vector space model. *IOP Conference Series: Materials Science and Engineering*, *288*, 012037. IOP Publishing. 24 sierpnia, Bandung, Indonezja. https://doi.org/10.1088/1757-899x/288/1/012037

Suszczańska, N. i Kulików, S. (2003). A Polish Document Summarizer. W M. H. Hanza (red.), *Applied Informatics* (s. 369–374). Proceedings of the 21st IASTED International Multi-Conference on Applied Informatics. 10–13 lutego 2003, Innsbruck, Austria. IASTED/ACTA Press.

Swamy, A. i Srinath, S. (2019). Automated Kannada text summarization using sentence features. *International Journal of Recent Technology and Engineering*, *8*(2), 470–474. https://doi.org/10.35940/ijrte.b1531.078219

Świetlicka, J. (2010). *Metody maszynowego uczenia w automatycznym streszczaniu tekstów* (praca magisterska). Uniwersytet Warszawski.

Xiang, X., Xu, G., Fu, X., Wei, Y., Jin, L. i Wang, L. (2018). Skeleton to abstraction: An attentive information extraction schema for enhancing the saliency of text summarization. *Information*, *9*(9), 217. https://doi.org/10.3390/info9090217

Zhang, Y., Li, D., Wang, Y., Fang, Y. i Xiao, W. (2019). Abstract text summarization with a convolutional Seq2seq Model. *Applied Sciences*, *9*(8), 1665. https://doi.org/10.3390/app9081665

Zhu, T. i Li, K. (2012). The similarity measure based on LDA for automatic summarization. *Procedia Engineering*, *29*, 2944–2949. https://doi.org/10.1016/j.proeng.2012.01.419

Zhuang, H., Wang, C., Li, C., Li, Y., Wang, Q. i Zhou, X. (2018). Chinese language processing based on stroke representation and multidimensional representation. W *IEEE Access*, *6*, 41928–41941. https://doi.org/10.1109/access.2018.2860058

Piotr Glenc jest informatykiem, uczestnikiem studiów doktoranckich z zakresu nauk o zarządzaniu, asystentem w Katedrze Projektowania i Analizy Komunikacji na Uniwersytecie Ekonomicznym w Katowicach. Zajmuje się problematyką związaną z automatyzacją analizy komunikacji w organizacjach, zwłaszcza w obszarze analizy komunikatów i dokumentów tekstowych oraz projektowaniem narzędzi informatycznych pozwalających na organizację i analizę komunikacji zachodzącej wewnątrz organizacji oraz między organizacjami a ich otoczeniem.