ISSN 2080-1807

TORUŃSKIE STUDIA BIBLIOLOGICZNE
2016, nr 1 (16)

Piotr Malak
Instytut Informacji Naukowej i Bibliologii
Uniwersytet Mikołaja Kopernika w Toruniu
e-mail: piomk@umk.pl

Adam Pawłowski
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytet Wrocławski
e-mail: adam.pawlowski@ibi.uni.wroc.pl

Ewaluacja skuteczności systemów wyszukiwania informacji. Wyniki eksperymentu Polish Task realizowanego w ramach Conference and Labs of the Evaluation Forum (CLEF) 2012

DOI: http://dx.doi.org/10.12775/TSB.2016.010

Streszczenie: W niniejszym artykule prezentujemy realizację laboratorium ewaluacyjnego CLEF (Conference and Labs of the Evaluation Forum) ze specjalnym uwzględnieniem kampanii CHiC (Cultural Heritage in CLEF). Opisujemy realizację oraz wyniki zadania Polish Task in ChiC. W artykule zaprezentowano wnioski z realizacji zadania. Zostały omówione wyniki uzyskane przez uczestników zadania przy użyciu różnych strategii indeksowania oraz wyszukiwania zasobów. Porównaliśmy efektywność metod tf-idf, OKAPI, DFR oraz data fusion.

Słowa kluczowe: CLEF, ewaluacja systemów informacyjno-wyszukiwawczych, laboratorium ewaluacyjne CLEF, wyszukiwanie informacji w języku polskim.

Wprowadzenie

Niniejszy artykuł jest kontynuacją wcześniejszej publikacji zawartej w „Toruńskich Studiach Bibliologicznych”[1], w której omówiono genezę i stan obecny eksperymentów TREC i CLEF, a także metody badań skuteczności wyszukiwania w dużych zbiorach dokumentów. Poniżej przedstawiamy przebieg oraz wyniki eksperymentu Polish Task, realizowanego w latach 2012–2013 przez Uniwersytet w Neuchatel, Uniwersytet Mikołaja Kopernika w Toruniu, Uniwersytet Wrocławski i Uniwersytet w Padwie (obsługa oceny efektywności wyszukiwania). Część danych empirycznych, jakie uzyskano dzięki tym badaniom, była przedmiotem kilku publikacji anglojęzycznych[2], natomiast zabrakło syntezy przygotowanej dla polskiego czytelnika. Niniejsza publikacja uzupełnia tę lukę.

Przestrzeń cyfrowa oferuje niespotykane wcześniej możliwości przechowywania i udostępniania szerokiemu gronu odbiorców dziedzictwa kulturowego ludzkości. Chociaż odbiorca takich treści nie ma fizycznego kontaktu z oryginalnym dziełem, a jedynie z jego cyfrowym odpowiednikiem, możliwe jest wzbogacenie obiektu wirtualnego o metadane dotyczące kontekstu powstania, nośnika, wytwórcy i samej treści. Skrócona tekstowa reprezentacja treści utworu jest szczególnie cenna w przypadku obiektów wizualnych – zdigitalizowanych pocztówek czy obrazów. W odróżnieniu od zdeponowanych w repozytoriach i bibliotekach cyfrowych tekstów, na których można prowadzić zaawansowane wyszukiwania i analizy, obraz (a ściślej przedstawiona na nim treść) wciąż wymaga ręcznego opracowania. Dopiero opisy w języku naturalnym pozwalają

---

[1] P. Malak, A. Pawłowski, *Ewaluacja skuteczności systemów wyszukiwania informacji. Od eksperymentu Cranfield do laboratoriów TREC i CLEF. Geneza i metody*, „Toruńskie Studia Bibliologiczne” 2015, nr 2 (15), s. 137–156.

[2] M. Akasereh, P. Malak, A. Pawłowski, *Evaluation of IR Strategies for Polish*, [w:] *Advances in Natural Language Processing. 9th International Conference on NLP, PolTAL 2014,Warsaw, Poland, September 17–19, 2014. Proceedings*, ed. by A. Przepiórkowski, M. Ogrodniczuk, Heidelberg [et al.] 2014, s. 384–391 (Lecture Notes in Computer Science; vol. 8686); P. Malak, *Information searching over Cultural Heritage objects, and press news*, [w:] *Human language technologies as a challenge for computer science and linguistics: 6th Language & Technology Conference, December 7–9, 2013, Poznań, Poland: proceedings*, ed. by Z. Vetulani, H. Uszkoreit, Poznań 2013, s. 434–438.

w pełni wykorzystać możliwość technik indeksacji i wyszukiwania – nie tylko zresztą w tradycyjnym środowisku unilingwalnym, ale także multilingwalnym. Jako przykład tekstowej reprezentacji treści obrazu można zacytować opis dodany do prezentowanego w encyklopedii Europeana zdjęcia pomnika ułana z dziewczyną[3]. W opisie tym czytamy m.in.:

> Pomnik ułana z dziewczyną na koronie ceglanych murów przy ul. Spichrzowej na Starym Mieście w Grudziądzu. Obie postacie są lekko przytulone do siebie, jakby właśnie do siebie podbiegły. Trzymają też bukiecik kwiatów. Autorem rzeźby, odsłoniętej 23 sierpnia 2008 r., jest miejscowy rzeźbiarz Ryszard Kaczor. Odlew wykonał A. Maczewski[4].

Przedstawiona tutaj forma opisu pozwala na automatyczną selekcję terminów kluczowych, a także ich przekład na inne języki. W wypadku obiektów starszych dostrzegalne są już jednak zmiany językowe, utrudniające procesy wyszukiwawcze. „Oryginały" mogą mianowicie zachowywać opisy zgodne ze starą ortografią, a także przenosić do współczesności inne, nie zawsze zrozumiałe dla odbiorców informacji znaczenia. Jako przykład możne podać wyraz *sejm*, który w dokumentach XVIII- i XIX-wiecznych ma pisownię *seym*, w dokumentach wcześniejszych *sjem*, *syem* czy *siyem*, natomiast w opisach obiektów zawsze występuje w formie współczesnej – *sejm*. Należy zauważyć, że dla większości cyfrowych zasobów dziedzictwa kulturowego właśnie metadane są podstawowym i przeważnie jedynym źródłem treści pozwalających na indeksowanie i wyszukiwanie.

Rozdrobnienie repozytoriów zawierających cyfrowe obiekty dziedzictwa kulturowego nie sprzyja ani pracy badawczej, ani codziennemu korzystaniu z tych danych. W celu ułatwienia dostępu do takich zasobów tworzone są agregatory treści cyfrowych. Jednym z nich jest Europeana – „metaagragator" i wyszukiwarka europejskich zasobów dziedzictwa kulturowego, gromadzący zasoby udostępniane przez europejskie

---

[3] M. Wieczorek, *Pomnik ułana z dziewczyną* [fot.] [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.europeana.eu/portal/record/2023815/LinkedHeritage_Update_ProvidedCHO_Pictures_bank_eu__ICIMSS__37250.html.

[4] Tamże.

biblioteki, muzea i archiwa[5]. Ze względu na rozległość zbioru, zróżnicowanie formatów danych oraz różnorodność językową metadanych zasoby Europeany zostały włączone w cykl laboratoriów ewaluacyjnych CLEF (Conference and Labs of the Evaluation Forum) w ramach zadania ChiC (Cultural Heritage in CLEF) w latach 2011–2013. Ich przedmiotem było wyszukiwanie informacji kulturowej w językach angielskim, niemieckim oraz francuskim. W 2013 r. do tej listy języków została dołączona sesja Polish Task[6]. Podczas badań, których wyniki przedstawiamy poniżej, zostały wykorzystane techniki przetwarzania języka naturalnego (ang. *Natural Language Processing*, NLP) oraz wyszukiwania informacji (ang. *Information Retrieval*, dalej: IR).

Polish Task – opis eksperymentu

Uczestnicy eksperymentu pracowali na tym samym zestawie danych (korpus oraz 50 zapytań), prowadząc ewaluacje wyszukiwania automatycznego oraz manualnie wzbogaconego. Wyszukania automatyczne mogły wykorzystywać dane zawarte w tezaurusach, dedykowanych ontologiach lub innych ustrukturyzowanych zasobach cyfrowych. Opcja wzbogacania manualnego dopuszczała wprowadzanie przez użytkowników własnych modyfikacji zapytań, dostosowujących je do różnych potrzeb informacyjnych oraz poziomu wiedzy potencjalnych użytkowników[7].

---

[5] *Europeana – O nas* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.europeana.eu/portal/aboutus.html.

[6] Autorzy artykułu są współorganizatorami oraz uczestnikami zadania Polish Task. Informacje dotyczące zadania, jego organizacji i przebiegu oraz wyników zob. V. Petras, T. Bogers, E. Toms, M. Hall, J. Savoy, P. Malak, A. Pawłowski, N. Ferro, I. Masiero, *Cultural Heritage in CLEF (CHiC) 2013*, [w:] *Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Information Access Evaluation. Multilinguality, Multimodality, and Visualization – 4th International Conference of the CLEF Initiative, CLEF 2013*, *Valencia, Spain, September 2013, Proceedings*, ed. by P. Forner [et. al.], Berlin–Heidelberg 2013, s. 192–211; P. Malak, *The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun Runs*, [w:] *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013*, ed. by P. Forner, R. Navigli, D. Tufis [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.clef-initiative.eu/documents/71612/b00f7561-fadb-47a8-ab67-74f116ce062a.

[7] Więcej o zadaniu zob. *CHiC 2013. Polish Task* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/tasks/polish-task;

Kolekcja

Przygotowana przez Europeanę kolekcja dla języka polskiego składała się z 1 093 705 dokumentów. W tym przypadku dokumentem jest opis obiektu zawarty w metadanych. Wewnętrzną strukturę zbioru prezentuje tabela 1.

Tabela 1. Struktura zbioru badawczego

Typ opisanego dokumentu	Liczba dokumentów	Udział w kolekcji
Tekst	975,818	89,222%
Obraz	117,075	10,704%
Video	582	0,053%
Dźwięk	230	0,021%

Źródło: opracowanie własne.

Wszystkie dokumenty składające się na kolekcję dostarczone były w postaci plików XML. Przykładową zawartość jednego dokumenty prezentuje tabela 2.

Tabela 2. Zawartość przykładowego dokumentu kolekcji

```
<ims:metadata ims:identifier="http://www.europeana.eu/resolve/record/92033/5F13
392D14630ECE62BFE40B0166F9C454C5C872" ims:namespace="http://www.europeana.
eu/" ims:language="pol">
<ims:fields>
<dc:date> [ca 1920]</dc:date><dc:format>text/html</dc:format>
<dc:identifier>http://193.59.172.16/szzz/ShowStart.do?id=23809</dc:identifier>
<dc:language>pol</dc:language>
<dc:publisher>M. Gladbach : B. Kühlen</dc:publisher>
```

Polish Track at CLEF 2013 [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/; Informacje dla uczestników: *Guidelines for Participation and Submission* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/Participation.html.

Tabela 2. Zawartość przykładowego dokumentu kolekcji (cd.)

```
<dc:publisher>Zakład Reprografii i Digitalizacji Zbiorów Bibliotecznych Biblioteki Narodowej, 2008</dc:publisher>
<dc:rights>Biblioteka Narodowa</dc:rights>
<dc:source>Biblioteka Narodowa, Poczt.13924</dc:source>
<dc:subject>Grodno (Białoruś) – ikonografia</dc:subject>
<dc:title>Grodno, klasztor oo. franciszkanów [Dokument ikonograficzny]</dc:title>
<dc:type>pocztówka</dc:type>
<europeana:country>poland</europeana:country>
<europeana:dataProvider>The National Library of Poland - Biblioteka Narodowa</europeana:dataProvider>
<europeana:isShownAt>http://193.59.172.16/szzz/ShowStart.do?id=23809</europeana:isShownAt>
<europeana:isShownBy>http://193.59.172.16/szzz/IsShownBy.do?id=23809</europeana:isShownBy>
<europeana:language>pl</europeana:language>
<europeana:object>http://193.59.172.16/szzz/IsShownBy.do?id=23809</europeana:object>
<europeana:provider>The European Library</europeana:provider>
<europeana:rights>http://creativecommons.org/publicdomain/mark/1.0/</europeana:rights>
<europeana:type>IMAGE</europeana:type>
<europeana:uri>http://www.europeana.eu/resolve/record/92033/5F13392D14630ECE62BFE40B0166F9C454C5C872</europeana:uri>
<europeana:year>1920</europeana:year>
</ims:fields>
</ims:metadata>
```

Źródło: opracowanie własne.

W strukturze metadanych wykorzystywano następujące schematy:

- Dublin Core (znaczniki zaczynające się prefiksem dc:),
- Qualified Dublin Core (znaczniki zaczynające się prefiksem dcterms:),
- Europeana Semantic Elements (znaczniki zaczynające się prefiksem europeana:).

W celu przyśpieszenia przetwarzania obszernych zasobów widocznych za pośrednictwem Europeany indeksowano tylko wybrane pola: <dc:contributor>, <dc:creator>, <dc:date>, <dc:language>, <dc:subject>, <dc:title>, <dc:type>, <dcterms:alternative>, <dcterms:created>, <europeana:language>, <europeana:type>, <europeana:uri>, <europeana:year>.

Zapytania

Uczestnikom zadania udostępniono zbiór 50 zapytań kontrolnych (ang. *topics*) w postaci krótkich haseł odzwierciedlających rzeczywiste potrzeby informacyjne użytkowników. Opracowano je na podstawie logów wyszukiwań w zasobach Europeany. Osobom testującym przedstawiono je w postaci plików XML w dwóch językach – po polsku na potrzeby wyszukiwania oraz po angielsku jako tłumaczenie informujące (zob. tab. 3).

Tabela 3. Zawartość przykładowego zapytania

```
<topic lang=”pl”>
<identifier>CHIC-2013-PL-001</identifier>
<title>meblarstwo polskie</title>
<description>prace poświęcone polskim meblom, polskiemu meblarstwu</description>
</topic>
```

Źródło: opracowanie własne.

Zawartość pola <description> służyła późniejszej ocenie zgodności odpowiedzi z zapytaniem i nie mogła być użyta w trakcie wyszukiwania informacji[8].

Zapytania składały się łącznie ze 141 wyrazów, co daje średnio 2,82 wyrazu na zapytanie. Spośród nich 10 było jednowyrazowych, 11 dwuwyrazowych i 29 dłuższych niż dwuwyrazowe. Cztery najdłuższe zapytania składały się z sześciu wyrazów (wraz ze spójnikami i przyimkami).

Realizacja eksperymentu

Pierwszym etapem badań z zakresu NLP jest wstępne przetworzenie tekstu (ang. *Pre-processing*). Służy ono ujednoliceniu form reprezentacji poszczególnych elementów (tokenów lub „egzemplarzy”) analizowanego tekstu. Elementami tymi są najczęściej wyrazy – rozumiane jako tekst

[8] Zasady udziału w eksperymencie opisano w dokumencie: *Guidelines for participation...*

ciągły ograniczony znakami przestankowymi lub spacjami. Kolejnymi etapami są: indeksowanie treści oraz dopasowywanie zapytań wyszukiwawczych i zindeksowanych dokumentów.

Wstępne przetwarzanie tekstu

Etap pre-processingu ma na celu dostarczenie materiału badawczego do dalszych analiz kwantytatywnych. Materiał ten powinien zapewniać ekonomiczne wykorzystanie zasobów komputerowych oraz jak najwyższy poziom ujednolicenia znaczeń przekazywanych w tekście.

Pierwszy warunek realizowany jest zazwyczaj za pomocą usunięcia z tekstu wyrazów nieznaczących (ang. *stop words*), czyli takich, które nie przenoszą istotnej treści. Należą do nich tzw. wyrazy gramatyczne, czyli spójniki, przysłówki, zaimki itp. oraz słowa o najwyższych częstościach w tekstach danego języka. Ustalając ich listę, należy uwzględnić rozkład frekwencji w konkretnym korpusie, a nie w języku jako całości. Analiza frekwencyjna korpusu badawczego Europeany pozwoliła wyodrębnić 304 wyrazy o zdecydowanie wyższych frekwencjach w korpusie. Spośród nich ostatecznie 138 zostało wpisanych na listę słów nieznaczących (tzw. stoplistę). Została ona następnie użyta do skrócenia tekstów z korpusu badawczego oraz zapytań – skoro dane słowo zostanie usunięte z korpusu, nie ma sensu rozpatrywanie go w zapytaniu, ponieważ nie zostanie dla niego znalezione dopasowanie. Należy zauważyć, że podejście czysto frekwencyjne przy tworzeniu list słów nieznaczących może obniżyć skuteczność systemu wyszukiwawczego. W przypadku danych Europeany jednymi z najczęściej występujących w polskim korpusie słów były *Polska* (oraz formy gramatyczne) i polskie nazwy własne. Pozostawienie ich na liście wyrazów do pominięcia spowodowałoby wykluczenie wielu dokumentów z procesu wyszukiwania. Najlepszą strategią tworzenia rozbudowanej stoplisty wydaje się uwzględnienie słownictwa strefy gramatycznej oraz najczęściej występujących – w danym korpusie tekstów – wyrazów ze strefy słownictwa podstawowego.

W przypadku języków fleksyjnych, a takim jest język polski, poszczególne wyrazy mogą występować w różnych formach gramatycznych, a tym samym pojawiać się w postaci różnych zapisów graficznych, np. *zdjęcie*, *zdjęcia*, *zdjęciu*. Ze względu na to poszczególne formy gramatyczne tego

samego wyrazu są traktowane przez systemy wyszukiwawcze jako odrębne jednostki. W celu ujednolicenia znaczeń stosuje się jedną z dwóch najpopularniejszych metod, czyli wyznacza się automatycznie rdzeń wyrazu (*stemming*) lub wskazuje jego podstawową formę gramatyczną (czynność też określa się jako hasłowanie lub lematyzację). Ze względu na wysoki poziom złożoności oraz koszty operacyjne (głównie czas przetwarzania) lematyzacji nie przeprowadzono. Oferuje ona wprawdzie wyższą skuteczność niż stemming, jednakże wydłuża czas oczekiwania na odpowiedź, a więc parametr istotny w zadaniach typu ad-hoc (a takim było Polish Task in CHiC).

Wyznaczanie rdzenia wyrazu, czyli *stemming*, jest standardowym podejściem przy konstrukcji wyszukiwarek internetowych. Również na potrzeby badań w ramach omawianego zadania wykorzystano to właśnie podejście. Należy tutaj zaznaczyć, że rdzenie uzyskiwane w wyniku tej procedury nie odpowiadają gramatycznym rdzeniom wyrazów[9]. Są to raczej ich skrócone graficzne reprezentacje, które dla systemu wyszukiwawczego przechowują określoną grupę znaczeń, np. **kole**: kole**ga**, kole**gi**, kole**dze** itd. Opierając się na wcześniejszych doświadczeniach z przetwarzaniem języka czeskiego[10], na potrzeby realizacji badań w ramach Polish Task in CHiC zastosowano tzw. *light stemming.* Dla języków fleksyjnych stemming powinien obejmować wszystkie główne klasy gramatyczne. Natomiast przytaczane badania dla języka czeskiego wykazały porównywalną skuteczność wyszukiwania informacji przy zastosowaniu stemmingu tradycyjnego oraz przy ograniczeniu operacji ujednolicania wyłącznie do rzeczowników – stąd nazwa angielska *light.*

Dla języka polskiego istnieją publicznie dostępne narzędzia ujednolicające, np. Stempel[11] czy Morfologik[12], jednak ze względu na problemy

---

[9] W przeciwieństwie do lematów uzyskiwanych podczas procesu lematyzacji – lematy zawsze są podstawowymi formami gramatycznymi danego wyrazu.

[10] Por. C. Fautsch, J. Savoy, *Algorithmic Stemmers or Morphological Analysis: An Evaluation*, „Journal of American Society for Information Science and Technology" 2009, vol. 60, iss. 8, s. 1616–1624; J. Savoy, *Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages*, [w:] *Proceedings. SAC '06 Proceedings of the 2006 ACM symposium on Applied computing*, New York 2006, s. 1031–1035.

[11] *Stempel – Algorithmic Stemmer for Polish Language* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.getopt.org/stempel/.

[12] *Morfologik/morfologik-stemming* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: https://github.com/morfologik/morfologik-stemming.

techniczne z ich wykorzystaniem przygotowano autorski system wyznaczania rdzenia wyrazu dla rzeczowników. Zastosowano w nim podejście całkowicie algorytmiczne: rdzeń wyrazu był wyznaczany na podstawie długości wyrazu oraz rozpoznanej końcówki funkcyjnej, np. obraz**-u**, pol**-ach**, kole-**gami**. Modyfikowane były wyrazy o długości równej lub większej niż cztery znaki, a usuwane końcówki miały długość od 1 do 3 znaków. Dla różnych długości wyrazów stosowane były różne zestawy reguł. Podczas oceny zgodności odpowiedzi z zapytaniem okazało się, że dla języka polskiego rdzenie o długości 4 lub nawet 5 znaków tworzą słowoformy zbyt polisemiczne. Można z nich wyprowadzić wiele klas znaczeniowych, co z kolei prowadzi do zwiększenia liczby trafień negatywnych (ang. *false positives*), czyli dopasowań prawidłowych pod względem formalnym, ale nieprawidłowych pod względem znaczeniowym.

Ocena stopnia zgodności

Wszystkie nadesłane przez uczestników odpowiedzi zostały poddane procesowi oceny zgodności przez specjalistów, dla których język polski był językiem ojczystym. Na tym etapie odpowiedzi systemu oceniano według trzystopniowej skali o następujących wartościach:

- zgodny (ang. *fully relevant*),
- częściowo zgodny (ang. *partially relevant*),
- niezgodny (ang. *not relevant*).

Grupowanie rezultatów (ang. *pooling*) przeprowadzono dla 100 najwyżej notowanych na liście rankingowej trafień z każdego nadesłanego zestawu odpowiedzi. Wartości trafień powtarzających się w kilku zbiorach odpowiedzi były uśredniane i umieszczane w zbiorze zgrupowanym jako niepowtarzalne. Zbiór dokumentów do oceny dla wszystkich 50 zapytań liczył 32 144 dokumenty. Przy wymagającym klasyfikowaniu, przyjmującym za trafne jedynie te odpowiedzi, które oznaczono jako *w pełni zgodne*, na zapytanie przypadało średnio 170 trafnych wyszukań. Przy łagodnym podejściu, przyjmującym jako akceptowalne odpowiedzi *zgodne* i *częściowo zgodne*, średni poziom trafności wynosił 256 zgodnych dokumentów na zapytanie (minimalnie 22, maksymalnie 562).

Proces grupowania oraz ocena zgodności były realizowane za pomocą systemu DIRECT (Distributed Information Retrieval Evaluation

Campaign Tool)[13]. Głównym wyznacznikiem skuteczności wyszukiwania była miara MAP (ang. *Mean Average Precision*), a jako dodatkową miarę zastosowano wskaźnik P@10 – odzwierciedlający skuteczność danego podejścia na pierwszej stronie z wynikami wyszukiwania danego systemu (obie te miary omawiamy w poprzednim artykule)[14].

Rozwiązania – strategie wyszukiwawcze

W zadaniach typu *ad hoc* poszczególne podejścia nazywane są po angielsku *run*. Ich wyniki są przesyłane do organizatorów poszczególnych zadań w postaci rankingowych list odpowiedzi. Zgodnie z przyjętym algorytmem wyszukiwania dokumenty występujące na początku takiej listy są najbardziej trafne w stosunku do zapytania. Za punkt odniesienia do porównań skuteczności poszczególnych rozwiązań przyjmuje się wyniki rozwiązania bazowego (ang. *baseline run*). W przypadku Polish Task było to zastosowanie algorytmu OKAPI bez wstępnego przetwarzania tekstów. Uczestnicy zadania przesłali rozwiązania zarówno z grupy podejść automatycznych, jak i manualnie wzbogacanych.

Rozwiązania automatyczne

Łącznie zgłoszonych zostało sześć rozwiązań automatycznych (w tym jedno już po terminie nadsyłania rozwiązań, jednakże zostało ono uwzględnione w porównaniach). W tabeli 4 zaprezentowano porównanie oceny skuteczności dla rozwiązań automatycznych.

Dla współczynnika MAP każdego z rozwiązań wykonano t-test (dla $\alpha = 5\%$) w porównaniu do wartości bazowej. Miał on wykazać występowanie znaczących statystycznie różnic pomiędzy rozwiązaniem bazowym a rozwiązaniami nadesłanymi (znaczne różnice oznaczono w tabeli 2 symbolem ±). Spośród sześciu nadesłanych prób automatycznych aż trzy

[13] *Distributed Information Retrieval Evaluation Campaign Tool* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://direct.dei.unipd.it/. System DIRECT obsługuje również kampanie ewaluacyjne TREC.

[14] P. Malak, A. Pawłowski, *Ewaluacja skuteczności…*

osiągnęły wartość MAP znacząco niższą niż próba bazowa. W żadnej z tych prób nie zastosowano wyznaczania rdzeni wyrazów (*stemming*). Natomiast w próbach o skuteczności wyższej niż próba bazowa wyznaczono rdzenie dla rzeczowników i przymiotników (*light stemming*), co wydaje się potwierdzać hipotezę postawioną na podstawie wyników badań nad językiem czeskim. Jednakże ostateczne określenie, czy *stemming* „lekki" wystarcza do efektywnego przetwarzania języka polskiego, wymaga odrębnych badań porównawczych na większym korpusie tekstów.

Tabela 4. Porównanie oceny skuteczności dla rozwiązań automatycznych Polish Task

Lp.	Rozwiązanie	Parametry rozwiązania (pre-prpocessing; indeksowanie; wyszukiwanie)	MAP	P@10
1	Torun_Auto	Stop list, light stemming; tf.idf; koniunkcja logiczna wyrażeń	**0.348**	–
2	UniNE_Fusion	data fusion[1] (light stemming; trunc-5[2]; Okapi)	0.343	0.614
3	UniNE_DFR	Stop list, light stemming; DFR-I(n_e)B2[3]	0.331	0.568
4	UniNE_PRF	data fusion, PRF (Rocchio, 5 docs, 10 terms)	0.258 ±	0.494
5	UniNE_Baseline	Stop list; tf.idf (miara podobieństwa - cosinus),	0.257 ±	0.492
6	UniNEGramPRF	data fusion, 5-gram, PRF	0.220 ±	0.472
7	Baseline run	Stop list; OKAPI (tf-idf, cosinus)	0.314	0.520

[1] Z wykorzystaniem standaryzacji wartości w próbce – z-score.
[2] Trunc-n metoda wyznaczania rdzenia wyrazów w sytuacji braku skutecznego stemera dla danego języka – polega na skracaniu wszystkich wyrazów w danym dokumencie do *n* pierwszych znaków.
[3] DFR – *Divergence From Randomness* – jeden z modeli probabilistycznych stosowanych w IR.

Źródło: opracowanie własne.

Najwyższy współczynnik MAP dla prób automatycznych osiągnęła próba z dodatkowym łączeniem (koniunkcja logiczna) wyrażeń z zapy-

tania. W tym przypadku zastosowano tradycyjną, statystyczną metodę wyznaczania wagi wyrazu tf-idf dla pojedynczych słów z dokumentu. Następnie dodatkowo premiowane były te dokumenty, w których występowały wszystkie wyrażenia z zapytań lub większość nich. Takie podejście pozwoliło zmniejszyć liczbę nieprawidłowych dopasowań typu *false positives* i zwiększyć zgodność zbioru odpowiedzi z zapytaniem. Dla zapytania #24 *Fryderyk Szopen* w próbie z łączeniem wyrażeń uzyskano średnią dokładność (*Average Precision, AP*) równą 0,9959, co stanowiło najwyższy wskaźnik AP dla wszystkich zapytań. Również ta metoda dała najlepsze rezultaty dla zapytania #31 *Lech lub Jarosław Kaczyński.* W przypadku Polish Task metoda premiująca współwystępowanie wyrażeń wyszukiwawczych wykazała się niższą efektywnością niż OKAPI tylko w przypadku 12 zapytań. Zastosowanie dodatkowo koniunkcji wyrażeń z zapytania dawało wyższą zgodność odpowiedzi dla zapytań dłuższych, dwu- lub więcej wyrazowych, podczas gdy metoda probabilistyczna, OKAPI (BM25), dawała lepsze rezultaty w przypadku zapytań jednowyrazowych. Można ten fakt potraktować jako kolejne potwierdzenie wyższej skuteczności probabilistycznego ważenia wyrazów w stosunku do ważenia wyłącznie frekwencyjnego (jak ma to miejsce w przypadku metody tf-idf).

Rozwiązania manualnie wzbogacane

Przy manualnym wzbogacaniu zapytań przyjęto założenie odzwierciedlenia różnego poziomu doświadczenia użytkowników, jednak z uwzględnieniem specyficznej zawartości Europeany. Pięć prób miało na celu odzwierciedlenie strategii wyszukiwawczych następujących grup użytkowników:

- studentów informacji naukowej,
- użytkowników wykształconych (magister),
- osób z wykształceniem wyższym humanistycznym,
- specjalistów wyszukiwania informacji,
- specjalistów z dziedziny.

Dla trzech pierwszych grup wzbogacanie zapytań polegało głównie na dodaniu synonimów wyrażeń z zapytania. Natomiast dla grup symulujących zapytania specjalistyczne dodawano adekwatne terminy

z encyklopedii oraz innych profesjonalnych źródeł. W rezultacie operacji wzbogacania dla trzech pierwszych grup użytkowników zapytania łącznie składały się z 303 wyrazów (średnio 6,1 na zapytanie), a dla grup specjalistycznych było to 489 wyrazów (średnio 9,78 na zapytanie). Dodawane były zazwyczaj terminy o szerokim polu znaczeniowym.

Przy wzbogacaniu manualnym usuwano wyrazy nieznaczące, dwukrotnie zastosowano dalsze przetwarzanie z operacją *light stemming* oraz, alternatywnie, bez tej operacji. Dla pozostałych nadesłanych rozwiązań nie stosowano stemmingu. Wszystkie wzbogacone rozwiązania zostały następnie przetworzone za pomocą algorytmu OKAPI (BM25), a wyniki porównano do tego samego rozwiązania bazowego, do którego porównywano wyniki rozwiązań automatycznych.

Tabela 5. Porównanie oceny skuteczności dla rozwiązań wzbogaconych manualnie

Rozwiązanie	**Symulowana grupa użytkowników, pre-processing**	**MAP**	**% zmiana wartości MAP**	**P@10**
PLTO1EduLS	Wykształcenie wyższe, light stemmer	**0.2774**	−11.66%	0.454
PLTO1EduNO	Wykształcenie wyższe	0.2724	−13.25%	**0.460**
PLTO2HighLS	Specjaliści dziedzinowi, light stemmer	0.2709	−14.33%	0.528
PLTO2HighNO	Specjaliści dziedzinowi	0.2690	−13.73%	0.528
PLWR2Exp	Specjaliści wyszukiwania informacji	0.1795 †	−42.83%	0.378
PLWR1Edu	Wykształcenie wyższe humanistyczne	0.1529 †	−51.31%	0.350
PLWR3Stu	Studenci	0.1279 †	−59.27%	0.268
Base Line	–	**0.3140**	**nie dotyczy**	**0.552**

Źródło: opracowanie własne.

Wyniki oceny zgodności dla każdego z pięciu rozwiązań z manualnym wzbogacaniem zapytań były gorsze niż próby bazowej. Można wskazać dwie główne tego przyczyny. Po pierwsze, do rozszerzonych zapytań pasowało zbyt wiele dokumentów z kolekcji – zastosowane wyrażenia

wzbogacające były na tyle ogólne, że mogły znaleźć się w większości dokumentów z zakresu dziedzictwa kulturowego. Tym samym w wynikach dopasowań znalazło się zbyt wiele dokumentów niezwiązanych tematycznie z zapytaniem. Drugą przyczyną niższej wydajności tej metody było zastosowanie zbyt specjalistycznego słownictwa w symulacjach zapytań specjalistów. Część terminów specjalistycznych nie występowała w dokumentach kolekcji, część zaś występowała w dokumentach niezgodnych z zapytaniami. Wartości współczynnika MAP zostały zaprezentowane w tabeli 5.

Oczywiście nie wszystkie wzbogacone zapytania cechowały się zgodnością niższą niż bazowa. Zapytanie #29 *Warszawa w 19 wieku w sztuce* po wzbogaceniu dla użytkownika z wykształceniem wyższym uzyskał współczynnik MAP = 0,3463, podczas gdy dal próby bazowej MAP wynosił 0,001. Dla tego zapytania zostały dodane terminy: *architektura* oraz *dzielnica*. Przy czym termin *architektura* przyczynił się do zwiększenia skuteczności wyszukiwania.

Wnioski

Ocena zastosowanych rozwiązań pozwala stwierdzić, że przy ustalaniu list rankingowych zgodności dokumentów z zapytaniem warto, oprócz ważenia słów, dodać również koniunkcję logiczną terminów użytych w zapytaniu i wyżej premiować dokumenty zawierające większość tych wyrażeń. Bardzo popularna metoda ustalania wagi słowa w dokumencie (OKAPI (BM25)) stosuje dopasowanie unigramowe – pojedynczo dla każdego terminu indeksowanego w zapytaniu. Przewagę łączenia terminów wyszukiwawczych nad pojedynczymi dopasowaniami uwidacznia szczególnie przykład zapytania #31 *Jarosław lub Lech Kaczyński*. Dla tego zapytania metoda OKAPI wygenerowała bardzo dużo nieprawidłowych dopasowań: jedynie 16 na 731 (2%) zwróconych w odpowiedzi dokumentów było zgodnych z zapytaniem. Sytuacja ta wynikła z faktu bardzo bogatej reprezentacji aktów miejskich miasta *Jarosław* dostępnych w Europeanie – większość odpowiedzi dotyczyła właśnie miasta. W całym korpusie było 3318 dokumentów trafnych dla hasła *Lech*, 1049 trafnych dla hasła *Kaczyński* i 9253 trafne dla hasła *Jarosław*, zarówno dla imienia, jak i nazwy miasta.

Kolejny wniosek dotyczy traktowania nazw własnych osobowych. Dla każdego hasła tego typu dobrą strategią wydaje się nadawanie wyższej wagi nazwisku, ponieważ często imię jest w tekstach skracane do inicjału przy kolejnym wystąpieniu, natomiast nazwisko zazwyczaj pozostaje w formie niezmienionej.

Podziękowania

Opisywane badania są częścią projektu realizowanego w ramach grantu Sciex-NMS *POL 11.219 – IRP Information Retrieval and Texts Categorisation for Polish*. Prace badawcze zrelacjonowane w niniejszym artykule były możliwe dzięki wsparciu finansowemu PROMISE (Participative Research Laboratory for Multimedia and Multilingual Information Systems Evaluation, Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191).

Bibliografia

Akasereh Mitra, Malak Piotr, Pawłowski Adam, *Evaluation of IR Strategies for Polish*, [w:] *Advances in Natural Language Processing. 9th International Conference on NLP, PolTAL 2014,Warsaw, Poland, September 17–19, 2014. Proceedings*, ed. by Adam Przepiórkowski, Maciej Ogrodniczuk, Heidelberg [et al.] 2014, s. 384–391 (Lecture Notes in Computer Science; vol. 8686).

CHIC 2012. Tasks [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.promise-noe.eu/tasks.

CHiC 2013.CHiC: Cultural Heritage in CLEF [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/home.

CHiC 2013. Polish Task [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/tasks/polish-task.

Elektroniczny słownik języka polskiego XVII i XVIII wieku [online]. Polska Akademia Nauk, Instytut Języka Polskiego, 2008 [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://sxvii.pl/index.php?strona=haslo&id_hasla=9516&forma=RZE%C5%B9BA#9516.

Europeana: think culture [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.europeana.eu/portal/.

Fautsch Claire, Savoy Jacques, *Algorithmic Stemmers or Morphological Analysis: An Evaluation*, „Journal of American Society for Information Science and Technology" 2009, vol. 60, iss. 8, s. 1616–1624.

Feldstein Ron F., *A Concise Polish Grammar* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=4.

Głowacka Ewa, *Badania efektywności języków informacyjno-wyszukiwawczych (komunikat z badań)*, [w:] *Komputeryzacja bibliotek: materiały konferencji 24–26 maja 1993 r., Toruń*, pod red. B. Ryszewskiego, Toruń 1994, s. 209––213.

Guidelines for Participation and Submission [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/Participation.html.

Jagodzinski Grzegorz, *A Grammar of the Polish Language* [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://grzegorj.w.interia.pl/gram/en/gram00.html.

Malak Piotr, *Information searching over Cultural Heritage objects, and press news*, [w:] *Human language technologies as a challenge for computer science and linguistics: 6th Language & Technology Conference, December 7–9, 2013, Poznań, Poland: proceedings*, ed. by Zygmunt Vetulani, Hans Uszkoreit, Poznań 2013, s. 434–438.

Malak Piotr, *The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun Runs*, [w:] *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013*, ed. by P. Forner, R. Navigli, D. Tufis [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.clef-initiative.eu/documents/71612/b00f7561-fadb-47a8-ab67-74f116ce062a.

Mykowiecka Agnieszka, *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*, Warszawa 2007.

Petras Vivien, Bogers Toine, Toms Elaine, Hall Mark, Savoy Jacques, Malak Piotr, Pawłowski Adam, Ferro Nicola, Masiero Ivano, *Cultural Heritage in CLEF (CHiC) 2013*, [w:] *Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Information Access Evaluation. Multilinguality, Multimodality, and Visualization – 4th International Conference of the CLEF Initiative, CLEF 2013*, *Valencia, Spain, September 2013, Proceedings*, ed. by P. Forner [et. al.], Berlin–Heidelberg 2013, s. 192–211.

Petras Vivien, Ferro Nicola, Gäde Maria, Isaac Antoine, Kleineberg Michael, Masiero Ivano, Nicchio Mattia, Stiller Juliane, *Cultural Heritage in CLEF (CHiC)*

Overview 2012 [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://www.clef-initiative.eu/documents/71612/0cadb163-3e32-4f16-a659-b457480c2a29.

Polish Track at CLEF 2013 [online] [dostęp 31 maja 2016]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/.

Savoy Jacques, *Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages,* [w:] *Proceedings. SAC '06 Proceedings of the 2006 ACM symposium on Applied computing*, New York 2006, s. 1031–1035.

Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych, pod red. Bożenny Bojar, Warszawa 2002.

Słownik poprawnej polszczyzny PWN, red. Witold Doroszewski; oprac. i red. Czesław Pankowski Warszawa 1995.

Szałkiewicz Łukasz, Przepiórkowski Adam, *Anotacja morfo składniowa*, [w:] *Narodowy Korpus Języka Polskiego*, red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Warszawa 2012, s. 59–96.

TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing), ed. by Ellen M. Vorhees, Donna K. Harman, Cambridge 2005.

Woźniak Jadwiga, *Kategoryzacja. Studium z teorii języków informacyjno-wyszukiwawczych*, Warszawa 2000.

Evaluation of IR Systems Efficiency. Results of Experiment Polish Task within Conference and Labs of the Evaluation Forum (CLEF) 2012

Abstract: The article presents the design of CLEF (Conference and Labs of the Evaluation Forum) evaluation labs with special attention paid to CHiC (Cultural Heritage in CLEF). We describe design of Polish Task in CHiClab and discuss conclusions from lab realisation. We discuss results achieved by different participants using different indexing and matching approaches. Efficiency of tf-idf, OKAPI, DFR and data fusion was compared and analysed.

Keywords: CLEF evaluation lab, IR in Polish, IR systems evaluation.