ISSN 2080-1807

TORUŃSKIE STUDIA BIBLIOLOGICZNE
2015, nr 2 (15)

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. Od eksperymentu Cranfield do laboratoriów TREC i CLEF. Geneza i metody

DOI: http://dx.doi.org/10.12775/TSB.2015.022

Streszczenie: W niniejszym artykule prezentujemy rozwój metod i miar służących do oceny efektywności systemów informacyjno-wyszukiwawczych. Zostały w nim opisane założenia eksperymentu Cranfield, jako długoletniego wyznacznika metodologii ewaluacyjnej, oraz zarzuty stawiane organizacji samego eksperymentu. Ważną częścią artykułu jest także opis ewolucji powszechnie dziś stosowanej metodologii ewaluacji systemów informacyjno-wyszukiwawczych, wypracowanej podczas dorocznych konferencji TREC (Text REtrieval Conference), a także omówienie najpowszechniej obecnie stosowanych miar ewaluacyjnych w tym zakresie. Artykuł przedstawia również organizację laboratoriów ewaluacyjnych CLEF (Conference and Labs of the Evaluation Forum) ze szczególnym uwzględnieniem panelu CHiC (Cultural Heritage in CLEF), a na gruncie języka polskiego – Polish Task in CHiC.

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

Wprowadzenie

Wraz z rozwojem oraz wzrostem powszechności komputerowych systemów informacyjno-wyszukiwawczych, w latach 60. XX w., oferujących m.in. możliwości automatycznego indeksowania zasobów piśmiennictwa[1], pojawiły się pytania o ocenę skuteczności tych systemów oraz użytych w nich algorytmów wyszukiwania. Postulaty, które im towarzyszyły, dotyczyły dostarczenia narzędzi do sprawdzenia efektywności stosowanych metod indeksowania i wyszukiwania informacji oraz wypracowania metodologii do porównań różnych rozwiązań w tym zakresie. Pierwszą systematyczną próbą odpowiedzi na to zapotrzebowanie były tzw. eksperymenty Cranfield[2], zaprojektowane i przeprowadzone przez Cyrila Wiliama Cleverdona na Uniwersytecie Cranfield w latach 60. XX w. (1957–1966). Polegały one na ocenie efektywności języków informacyjno-wyszukiwawczych i wykazały, że możliwe jest pełnotekstowe indeksowanie zasobów. W latach 60. była to teza śmiała, podobnie jak sam sposób przeprowadzenia porównań, który wywarł duży wpływ na dalsze badania nad wyszukiwaniem informacji.

Na potrzeby eksperymentu przygotowano *kolekcję testową* (ang. *test collection*), składającą się z 1400 dokumentów z zakresu aerodynamiki oraz 225 *zapytań testowych* (ang. *topics*) jako symulację potrzeb informacyjnych użytkowników. Zastosowanie listy zapytań wygenerowanych zamiast wprowadzanych przez realnych użytkowników stanowiło *novum* w badaniach nad skutecznością wyszukiwania informacji (obecnie prak-

[1] W latach 60. XX w. automatyczne indeksowanie polegało głównie na operacji generowania słów kluczowych z opisów rzeczowych dokumentów. Zob. *Indeksowanie automatyczne*, [w:] *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, pod red. B. Bojar, Warszawa 2002, s. 86.

[2] C. W. Cleverdon, *Evaluation of operational information retrieval systems*. P. 1: *Identification of criteria. Cranfield*, 1964; tenże, *The Cranfield tests on index language devices*, „ASLIB proceedings" 1967, vol. 19, no. 6, s. 173–193; tenże, *Progress in documentation: evaluation tests on information retrieval systems*, „Journal of Documentation" 1970, vol. 26, no. 1, s. 55–67. Cyt za J. Woźniak, *Kategoryzacja. Studium z teorii języków informacyjno-wyszukiwawczych*, Warszawa 2000, s. 22.

tyka taka przyjęta jest jako standardowa). Dla każdego zapytania podane były również listy dokumentów relewantnych znajdujących się w kolekcji. Listy zgodnych odpowiedzi służyły jako miara skuteczności testowanych języków wyszukiwawczych.

Od początku lat 90. Badacze związani m.in. z National Institute of Standards and Technology [dalej: NIST] w Stanach Zjednoczonych, uznając wprawdzie doniosłość eksperymentów Cranfield oraz ich kontynuacji, wskazali na ich słabe punkty. Dwa najpoważniejsze zarzuty były następujące:

1. Eksperymenty nie zawsze przeprowadzane były z wykorzystaniem tych samych danych czy takich samych procedur i miar ewaluacyjnych. Brakowało również porównania wyników uzyskanych przy różnych strategiach wyszukiwawczych. W związku z tym utrudnione lub wręcz niemożliwe było wskazanie systemu lub metody sprawdzającej się najlepiej w sytuacji odzwierciedlanej przez warunki eksperymentu. Rezultaty tak prowadzonych eksperymentów dawały wyniki doraźne, pozwalające ocenić skuteczność jednego systemu lub metody. Nie pozwalały jednak na porównanie całych systemów lub metod wyszukiwawczych.
2. Rozmiary stosowanych kolekcji testowych były niewielkie, a przez to niereprezentatywne. Wyniki uzyskiwane dla małych kolekcji nie były wiarygodne dla producentów komercyjnych systemów informacyjno-wyszukiwawczych, które pracowały i pracują z wielkimi zbiorami dokumentów[3].

Eksperymenty Cranfield, mimo swoich niedociągnięć, stały się wzorem dla kolejnych badaczy efektywności wyszukiwania[4]. Przyjęte w nich założenia utworzenia testowej kolekcji dokumentów, zestawu zapytań

---

[3] Por. E. M. Vorhees, D. K. Harman, *The Text Retrieval Conference*, [w:] *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, ed. by E. M. Vorhees, D. K. Harman, Cambridge 2005, s. 4–5; D. K. Harman, *The TREC test collections*, [w:] *TREC: Experiment and Evaluation…*, s. 21–23.

[4] Warto przytoczyć tu m.in. eksperymenty Ewy Głowackiej z Uniwersytetu Mikołaja Kopernika w Toruniu, polegające na ocenie dokładności i kompletności wyszukiwania informacji dla zapytań informacyjnych wyrażonych w dwóch wersjach języka haseł przedmiotowych oraz w języku słów kluczowych. Badaczka wykazała, że język haseł przedmiotowych oferuje wyższą kompletność, język słów kluczowych zaś większą dokładność. Por. E. Głowacka, *Badania efektywności języków informacyjno-wyszukiwawczych (komunikat*

symulujących potrzeby informacyjne użytkowników oraz listy dokumentów trafnych były i są powszechnie wykorzystywane w badaniach porównawczych nad efektywnością wyszukiwania informacji. Pierwszą usystematyzowaną próbę dopracowania założeń eksperymentów porównawczych dla efektywności wyszukiwania informacji podjęli w latach 90. ubiegłego wieku badacze związani z wzmiankowanym NIST. W owym czasie dysponowano już komputerami o mocach obliczeniowych, które pozwalały na efektywne budowanie i przetwarzanie rozległych kolekcji dokumentów. W 1991 r. powołano TREC (Text REtrieval Conference)[5] jako forum naukowe działające w formule warsztatowej, którego zadaniem jest dostarczenie infrastruktury dla wielkoskalowych działań ewaluacyjnych w zakresie systemów wyszukiwania informacji[6]. Celem TREC jest wyeliminowanie niedociągnięć eksperymentów Cranfield przez dostarczenie wielkich kolekcji testowych, ujednoliconych procedur oceny oraz środowiska wymiany metod i wyników. Konferencje TREC składają się z prezentacji różnego rodzaju zadań (ang. *tasks*), wśród których rolę centralną odgrywają tzw. zadania *ad-hoc* (ang. *ad-hoc tasks*). Służą one ocenie możliwości systemów informacyjno-wyszukiwawczych w zakresie dokładności i kompletności generowania rankingowych list odpowiedzi na potrzeby informacyjne wyrażone za pomocą 50 zapytań[7].

Eksperymenty typu *ad-hoc* polegają na indeksowaniu przez system wyszukiwawczy dokumentów z kolekcji testowej, a następnie na automatycznym wygenerowaniu i posortowaniu odpowiedzi na przekazane do systemu zapytania. System wyszukiwawczy powinien wskazać dokumenty trafne dla danego zapytania z listy bez odwoływania się do dodatkowej wiedzy o potrzebach informacyjnych użytkowników, wyłącznie na podstawie zapytania oraz analizy kolekcji dokumentów do przeszukania[8]. Zespoły biorące udział w eksperymencie, zwane uczestnikami, mają

z badań), [w:] *Komputeryzacja bibliotek: materiały konferencji 24–26 maja 1993 r.*, Toruń, pod red. B. Ryszewskiego, Toruń 1994, s. 209–210.

[5] *Text REtrieval Conference (TREC)* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://trec.nist.gov/.

[6] E. M. Vorhees, D. K. Harman, dz. cyt., s. 3.

[7] Tamże, s. 5.

[8] *CHIC 2012. Tasks* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/tasks.

pełną swobodę wyboru i implementacji strategii indeksowania oraz wyszukiwania i ustalania rankingu odpowiedzi. Wyniki pojedynczego eksperymentu, zwanego *przebiegiem* (ang. *run*), dostarczane są przez uczestników w postaci list rankingowych odpowiedzi uzyskanych na każde z zapytań testowych, przy zastosowaniu jednej procedury wyszukiwawczej[9]. Uczestnicy laboratoriów mogą przesyłać wyniki określonej liczby eksperymentów.

Kolejnym zadaniem wykonanym przez badaczy uczestniczących w konferencjach TREC było dostarczenie uniwersalnych miar pozwalających obiektywnie ocenić oraz porównać skuteczność poszczególnych systemów i metod wyszukiwania informacji. Spośród miar ewaluacyjnych zaproponowanych przez TREC dwie zostały uznane za wystarczająco wiarygodne i obiektywne do tego, by rekomendować je jako narzędzia oceny efektywności systemów informacyjno-wyszukiwawczych, stosowanych w praktyce. Są to:

1. Uśredniona dokładność

AP (ang. *Average Precision*) jest miarą dokładności dla każdego dokumentu zgodnego z zapytaniem, uśrednioną przez liczbę wszystkich dokumentów wskazanych przez system dla danego zapytania. Przy uwzględnieniu miar uśrednionej dokładności dla wszystkich zapytań testowych uzyskujemy **średnią dokładność uśrednioną** (**MAP**, ang. *Mean Average Precision*). MAP jest miarą najczęściej wykorzystywaną w badaniach nad efektywnością wyszukiwania informacji.

Aby zdefiniować powyższe zmienne, należy wprowadzić parametr **dokładności** (**P**). W teorii informacji **P** (ang. *precision*) jest miarą efektywności metody wyszukiwawczej i wyraża stosunek liczby dokumentów trafnych podanych w odpowiedzi do liczby wszystkich zwróconych dokumentów[10]:

$$P = \frac{\text{liczba zwróconych dokumentów trafnych}}{\text{liczba zwróconych dokumentów}}. \quad [1]$$

[9] Ch. Buckley, E. M. Voorhees, *Retrieval System Evaluation*, [w:] *TREC: Experiment and Evaluation…*, s. 53.

[10] Por. A. Mykowiecka, *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*, Warszawa 2007, s. 270.

Z kolei uśredniona dokładność (**AP**) opisuje sumaryczną dokładność zbioru wyników dla każdej kolejnej trafnej odpowiedzi, obliczaną dla każdego pojedynczego wyszukania w zbiorze według wzoru [1]. Można opisać ją wzorem:

$$AP = \left(\sum_{n=1}^{N} P_n \right) - R. \qquad [2]$$

gdzie:

P_n oznacza precyzję dla n-tego trafnego dokumentu,

R – liczbę dokumentów relewantnych.

Na przykład, jeśli na 10 podanych odpowiedzi wszystkie 4 trafne dokumenty znajdują się na pozycjach 1, 3, 5 i 6, AP dla tego zapytania będzie wynosić:

AP = (1/1 + 2/3 + 3/5 + 4/6)/4 = 0,73

Dla opisywanych warunków dokładność wyniesie: P = 4/10 = 0,4

MAP jest miarą uśrednionych dokładności dla wszystkich zapytań w zestawie i oblicza się ją według wzoru:

$$P = \left(\sum_{n=1}^{N} AP_n \right) - Q. \qquad [3]$$

gdzie:

AP_n oznacza średnią dokładność dla n-tego zapytania (*topic*),

Q – liczbę ocenianych zapytań.

Współczynnik MAP jest wprost proporcjonalny do efektywności systemu wyszukiwawczego – jego wysoka wartość wskazuje na lepszą skuteczność.

AP i związana z nią MAP są miarami bardzo wrażliwymi na modyfikacje mechanizmów wyszukiwawczych i rankingowych. Wszelkie zmiany liczby oraz pozycji rankingowej dokumentów trafnych wpływają na zmianę wartości AP i MAP dla danego eksperymentu. Jeśli w podanych przykładowych wynikach zmienimy pozycje rankingowe trafnych dokumentów na 1, 2, 5, 6 (przesunięcie drugiego wyniku o jedną pozycję w górę na liście rankingowej – w wyniku wyższej oceny stopnia zgodności z zapytaniem), to średnia dokładność wyniesie:

AP = (1/1 + 2/2 + 3/5 + 4/6)/4 = 0,82

podczas gdy dokładność nadal będzie wynosiła P = 4/10 = 0,4

Wartość MAP, odzwierciedlająca średnią dokładność dla wszystkich zapytań, również ulegnie wtedy zmianie.

2. Dokładność dla K dokumentów

P(K) lub P@K (ang. *Precision at document cutoff K*) wyraża dokładność wyszukania przy ograniczeniu zbioru zwróconych dokumentów do zadanej wielkości. W praktyce miara ta jest wykorzystywana głównie do ewaluacji skuteczności algorytmów rankingowych zastosowanych w systemie wyszukiwania informacji. Przy ograniczeniu do pierwszych dziesięciu pozycji listy odzwierciedla ona dokładność odpowiedzi podanych na pierwszej stronie wyników wyszukiwania – stąd jej popularność w postaci P(10). Modyfikacje parametru K nie wskazują efektywności systemu wyszukiwawczego, lecz jedynie skuteczność zastosowanych w systemie algorytmów rankingowych (oceniających poziom zgodności wskazanego dokumentu z zapytaniem) dla odpowiedzi wskazanych przez system. Miara ta pełni funkcje pomocnicze w stosunku do MAP czy AP. Dla rankingu, w którym na dziesięć pierwszych pozycji znajdziemy 6 trafnych dokumentów, P(10) = 6/10 = 0,6, w przypadku zaś czterech trafnych dokumentów P(10) = 4/10 = 0,4, co nie ma związku z dokładnością dla całego zbioru wyników.

3. Dokładność dla R relewantnych dokumentów

R-Precision jest miarą dokładności uzyskanego zbioru odpowiedzi przy ograniczeniu do wskazanej liczby (R) dokumentów trafnych. Od dokładności dla K dokumentów odróżnia ją to, że wyraża ocenę przeprowadzonych obliczeń po otrzymaniu wskazanej liczby dokumentów **relewantnych**, podczas gdy P(K) dotyczy ogólnej liczby dokumentów w odpowiedzi. Miarę tę wykorzystuje się jako wyznacznik ogólnej skuteczności badanego systemu, podczas gdy P(K) wyznacza skuteczność konkretnego zastosowania algorytmów rankingowych[11]. R-dokładność wykazuje silne powiązanie z wartościami MAP, lecz można ją również zastosować do oceny

[11] Ch. Buckley, E. M. Voorhees, dz. cyt., s. 59.

skuteczności algorytmów porządkowania zbioru odpowiedzi[12] lub do oceny kompletności zbioru wyników. Wtedy nazywana jest **R-Recall**.

Oprócz wymienionych i powszechnie używanych miar, zaproponowanych przez TREC, do oceny efektywności systemów wyszukiwawczych można zastosować średnią harmoniczną dokładności i kompletności, określaną jako *F-measure* (lub *F-score*). Jej zaletą jest uwzględnianie dokładności (ang. *precision*) i kompletności (ang. *recall*) wyników dla systemu wyszukiwawczego.

Średnia harmoniczna dokładności i kompletności

Miara ta oznaczana jest wzorem:

$$F_1 = \frac{2 \times \text{dokładność} \times \text{kompletność}}{\text{dokładność} + \text{kompletność}}. \quad [4]$$

F-measure może także odzwierciedlać wagi poszczególnych wartości w procesie ewaluacji. Wybraną miarę mnożymy w tym celu przez współczynnik wagi i otrzymujemy średnią harmoniczną ważoną:

$$F_\alpha = \frac{(\alpha + 1) \times \text{dokładność} \times \text{kompletność}}{\alpha \times \text{dokładność} + \text{kompletność}}. \quad [5]$$

W ramach konferencji TREC zaproponowane zostały również inne miary skuteczności, jednakże mają one ograniczone zastosowanie, np. tylko dla specyficznych warunków lub typów danych. Miary sprawdzone i zalecane przez TREC są powszechnie wykorzystywane w badaniach nad skutecznością systemów wyszukiwania informacji, w tym m.in. przez laboratoria ewaluacyjne **CLEF** (*Conference and Labs of the Evaluation Forum*).

Konferencje i laboratoria ewaluacyjne CLEF

Konferencje CLEF (Conference and Labs of the Evaluation Forum) organizowane są cyklicznie od 2000 r. (do 2010 r. pod nazwą *Cross-Language*

[12] A. Mykowiecka, dz. cyt., s. 271.

Evaluation Forum)[13]. Metodologia prac ewaluacyjnych CLEF jest taka sama, jak ta, którą zaproponowano w TREC. W ramach CLEF organizowane są laboratoria badawcze poświęcone różnym aspektom wyszukiwania informacji, włączając w to również pracę z zasobami nietekstowymi (grafika, video). Organizatorzy udostępniają testowe korpusy dokumentów oraz listy zapytań z informacją, jakie dokumenty uznawane będą za trafne dla każdego z zapytań. Uczestnicy zaś opracowują lub dostosowują metody wyszukiwania, przeprowadzają zadane procesy indeksowania, wyszukiwania i ustalania rankingu odpowiedzi, a następnie przesyłają do organizatorów uzyskane wyniki. Laboratoria CLEF zapewniają również narzędzia do oceny i porównania uzyskanych wyników. Rezultaty nadsyłane przez uczestników są łączone w jeden *zbiór odpowiedzi na poszczególne zapytania* (ang. *pooling*), a następnie oceniana jest ich zgodność z zapytaniem. Ocena dokonywana jest zazwyczaj manualnie przez specjalistów, dysponujących wskazówkami, jakie dokumenty należy uznać za trafne. Rezultaty oceny przekazywane są uczestnikom, którzy mogą sprawdzić skuteczność przyjętej metody wyszukiwania.

Podobnie jak TREC, konferencja CLEF oferuje każdego roku kilka ścieżek tematycznych, które tworzą tzw. zadania *ad-hoc* (ang. *ad-hoc tasks*), służące ewaluacji systemów wyszukiwawczych, oraz zadania opracowane dla konkretnych typów danych (np. ImageCLEF – analiza i anotacja plików graficznych, WebCLEF – wielojęzyczne wyszukiwanie informacji w Internecie, GeoCLEF – wyszukiwanie geograficznych jednostek nazewniczych w tekstach). Większość współczesnych laboratoriów CLEF dotyczy zasobów wielojęzykowych, jednakże zdarzają się również eksperymenty dotyczące tylko jednego języka.

Przykładem ścieżki badawczej poświęconej wyszukiwaniu informacji w jednym języku była ścieżka CHiC (Cultural Heritage in CLEF), organizowana w latach 2011–2013, a współorganizowana m.in. przez encyklopedię Europeana[14]. Europeana jest inicjatywą mającą na celu udo-

[13] Informacje dotyczące minionych oraz bieżących tematów badawczych realizowanych w ramach laboratoriów CLEF można znaleźć w witrynie: http://www.clef-initiative.eu/track/series [dostęp 15 grudnia 2015]. Archiwalne informacje dotyczące edycji 2000–2009 dostępne są pod adresem: http://www.clef-campaign.org [dostęp 15 grudnia 2015].

[14] Więcej na temat ChiC 2012 zob. *ChiC 2012* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2012/home.

stępnienie szerokiemu gronu odbiorców europejskich zasobów cyfrowych dziedzictwa kultury, nauki i sztuki. Pełni funkcje meta-agregatora oraz wyszukiwarki obiektów dziedzictwa kulturowego (ang. *Cultural Heritage Objects*). Portal gromadzi dane udostępniane przez muzea, biblioteki i archiwa europejskie[15].

W zadaniu *ad-hoc retrieval task*[16] eksperymenty dotyczyły wyszukiwania informacji w zasobach dziedzictwa kulturowego dla trzech języków: angielskiego, niemieckiego oraz francuskiego. Wyszukiwanie odbywało się dla każdego z języków pojedynczo, dla par języków oraz dla wszystkich trzech języków równocześnie. Każdej z opcji przyświecał cel badawczy związany z innym typem wyszukiwania informacji.

Do 2013 r. w ramach CLEF nie odbyła się żadna sesja badawcza poświęcona w całości językowi polskiemu. Organizatorzy uznali jednak, że ze względu na objętość zasobów i swoją pozycję w Europie polszczyzna zasługuje na odrębną sesję badawczą w zakresie wyszukiwania informacji. Odzwierciedleniem tego przekonania stało się zorganizowanie w roku 2013 odrębnego zadania w ramach CHiC – *Polish task*, poświęconego wyszukiwaniu informacji w dokumentach polskojęzycznego i/lub polskiego dziedzictwa kulturowego. Poniżej zamieszczono pierwszą część opisu eksperymentów, przeprowadzonych w polskiej ścieżce CHiC 2013. Opis uzyskanych wyników ukaże się w kolejnym numerze „Toruńskich Studiów Bibliologicznych".

Język polski z punktu widzenia wyszukiwania informacji

Pod względem genetycznym język polski należy do rodziny indoeuropejskiej, grupy słowiańskiej, a w jej ramach do podgrupy zachodniosłowiańskiej. Dzięki indoeuropejskiemu pochodzeniu wykazuje wiele podobieństw do języków z grupy germańskiej i romańskiej, reprezentujących rdzeń kultury europejskiej. Zaletą polszczyzny jest także posiadanie alfabetu łacińskiego, wzbogaconego o dziewięć dodatkowych znaków, utworzonych przez dodanie cech diakrytycznych do liter łacińskich (tzw.

---

[15] *Europeana: pomyśl o kulturze* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.europeana.eu/portal/aboutus.html.

[16] Por. *ChiC 2012: Tasks*…

„polskie" znaki: ą, ę, ó, ł, ń, ć, ś, ż, ź). Chociaż nie jest to uznaną normą, znaki te można bez znaczącej deformacji sensu zastępować ich łacińskimi odpowiednikami, co ułatwia przetwarzanie informacji przez niektóre systemy automatyczne. Oprócz „przyjaznego" alfabetu procesy wyszukiwawcze ułatwia też, z lingwistycznego punktu widzenia, leksyka z tych obszarów tematycznych, które są reprezentowane w zbiorach dziedzictwa kulturowego (terminologia specjalistyczna i naukowa). Leksyka ta, jeśli chodzi o etymologię i rdzenie słów, jest w znacznej części wspólna wszystkim językom europejskich, co ogranicza negatywne efekty zjawiska polisemii na jakość wyszukiwania (jednak ich nie eliminuje).

Więcej uwagi należy poświecić cechom typologicznym polszczyzny, do których należy jej umiarkowanie syntetyczny charakter, objawiający się fleksyjnością i bogatą morfologią. W języku angielskim, stanowiącym nieformalny prototyp wielu języków opisu, fleksja nie jest tak rozwinięta, co tworzy fałszywe wrażenie trudności języka polskiego. Trudność może polegać jedynie na tym, że systemy komputerowe nie przewidują daleko idącej wariantywności form i koreferencji, preferując leksemy hasłowe, czego przykładem są słowa kluczowe – zawsze w mianowniku liczby pojedynczej, oczywiście z wyjątkiem form *plurale tantum*. Współczesna inżynieria języka rozwiązuje jednak te problemy coraz skuteczniej. Należy tutaj podkreślić, że tradycyjne opisy systemu gramatycznego polszczyzny nie są dla systemów automatycznego przetwarzania języka wystarczające. Na ich podstawie powstały opisy nowe, lepiej dostosowane do oznaczania morfoskładniowego korpusów tekstów. Przytaczanie ich w całości nie jest tutaj konieczne, warto natomiast podać główne cechy gramatyczne, wpływające na wyniki wyszukiwania informacji.

W literaturze przedmiotu, również polskiej, często występuje pojęcie *tagset.* Termin ten oznacza, w największym skrócie, listę etykiet (znaczników) wybranych do oznaczenia formy wyrazów danego języka[17]. W praktyce tagset jest to zestaw zasad znakowania morfosyntaktycznego tekstów danego języka wraz z zestawem używanych w tym celu oznaczeń[18]. W tagsecie polszczyzny wyróżnia się dwie podstawowe kategorie opisu: tzw. fleksemy („zbiór form jednolicie lub niemal jednolicie zróż-

[17] Cyt. za. A. Mykowiecka, dz. cyt., s. 74.

[18] M. Woliński, *System znaczników morfosyntaktycznych w korpusie IPI PAN*, „Polonica" t. XXII–XXIII: 2003, s. 39–55.

nicowanych ze względu na właściwe im kategorie gramatyczne"[19]) oraz „zwykłe" kategorie gramatyczne. Janusz Bień wyróżnił 35 fleksemów, które wprowadzono m.in. w oznaczeniach Narodowego Korpusu Języka Polskiego. Za fleksemy uznano m.in. różne formy czasownika (np. bezosobnik *jadano*, bezokolicznik, odsłownik *jadanie*[20]), dwie formy rzeczownika (nienacechowaną i deprecjatywną), a także wiele form trudnych do zaklasyfikowania (tzw. kubliki, burkinostki, ciała obce, interpunkcję). Fleksemom przyporządkowano kategorie gramatyczne, które tworzą bogactwo i elastyczność polszczyzny, wymagają jednak narzędzi innych niż te, które stosuje się do języków o tendencji analitycznej (np. angielski). I tak, gdy chodzi o kategorie współczesne i statystycznie dominujące, polszczyzna ma dwie liczby (sg i pl), siedem przypadków gramatycznych, pięć rodzajów (męski osobowy, męski zwierzęcy, męski rzeczowy, żeński, nijaki), trzy osoby gramatyczne, trzy stopnie przymiotnika i przysłówka, dwa aspekty (dokonany i niedokonany, w tagsecie pominięto pozostałe), dwie postaci negacji, dwie postaci akomodacji, dwie formy akcentowe (*go* vs *jego*)[21].

Wyliczenie i opis tych kategorii powinien dać wyobrażenie o tym, jakie trudności pokonuje się w automatycznych systemach przetwarzania języka polskiego i języków podobnych pod względem gramatycznym. Należy jednak mieć na uwadze fakt, że wyszukiwanie informacji w opisach dokumentów uwzględnia przede wszystkim nazwy proste lub wielowyrazowe, czyli rzeczowniki lub złożone wyrażenia nominalne, natomiast analiza składni całych zdań nie jest konieczna. Problemem jest więc hasłowanie (lematyzacja) opisów, czyli sprowadzenie wyrazów odmienionych do postaci podstawowej (np. rzeczowników do liczby pojedynczej i mianownika), oraz rozpoznawanie homografii. W opisanych dalej eksperymentach, realizowanych w ramach ścieżki Polish Task, zastosowano w tym celu tzw. *light stemming* czyli sprowadzanie do wspólnego rdzenia (skróconej postaci wyrazu, niekoniecznie poprawnej gramatycznie) głównie rzeczowników oraz niektórych przymiotników – dzielących wspólne

[19] Ł. Szałkiewicz, A. Przepiórkowski, *Anotacja morfoskładniowa*, [w:] *Narodowy Korpus Języka Polskiego*, pod red. A. Przepiórkowskiego, M. Bańko, R. L. Górskiego, B. Lewandowskiej-Tomaszczyk, Warszawa 2012, s. 62.

[20] Tamże, s. 63.

[21] Pozostałe patrz: tamże, s. 65.

wzory odmian z rzeczownikami. Kolejnym elementem ujednolicania zapisów był etap manualnego wzbogacania zapytań – zaangażowani w eksperyment specjaliści podawali terminy rozszerzające zapytania w formach nominalnych.

Polish Task[22] – organizacja eksperymentu

Program Polish task został zorganizowany przy współpracy Uniwersytetu Neuchatel w Szwajcarii, Uniwersytetu Mikołaja Kopernika w Toruniu oraz Uniwersytetu Wrocławskiego. Było to typowe zadanie *ad-hoc* dla jednego języka, stanowiące kontynuację laboratoriów monolingwistycznych ChiC 2012. Celem zadania była ocena skuteczności różnych technik wyszukiwania informacji w zasobach opisanych językiem o rozbudowanej morfologii i fleksji. Przyjęto założenie, że cechy morfosyntaktyczne polszczyzny będą mieć wpływ na wyniki indeksowania tekstów napisanych w tym języku oraz na efektywność i trafność wyszukiwania informacji.

Kolekcję testową stanowiły opisy polskich zasobów encyklopedii Europeana, dla których organizatorzy przygotowali zestaw pięćdziesięciu zapytań. Kolekcja testowa jest częścią zbioru wielojęzycznego, wyko-

---

[22] Autorzy artykułu są współorganizatorami oraz uczestnikami zadania Polish Task. Informacje dotyczące zadania, jego organizacji i przebiegu oraz wyników zob. 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; 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 15 grudnia 2015]. Dostępny w World Wide Web: http://www.clef-initiative.eu/documents/71612/b00f7561-fadb-47a8-ab67-74f116ce062a.

rzystywanego w kampaniach badawczych ChiC (ang. *multilingual task*) w latach 2012 i 2013. Część zapytań przygotowanych dla opisywanego projektu została włączona w zestaw zapytań dla zadania „multilingual task ChiC 2013"[23]. W tym drugim przypadku zapytania zostały przetłumaczone na pozostałych 13 języków.

W laboratorium Polish task uczestnicy mieli możliwość przesłania wyników dwóch rodzajów wyszukiwania (osobno dla każdego z nich):

1. automatycznego,
2. manualnie wzbogaconego (ang. *manually enriched*).

W pierwszym przypadku można było zastosować dowolnie wybrane metody indeksowania, wyszukiwania i ustalania rankingu zgodności wyników, pracując na oryginalnych zbiorach tekstów i zapytań. Dozwolone było również automatyczne rozbudowywanie zapytań. W tym celu można było skorzystać z tezaurusów, dedykowanych ontologii lub z materiałów dostępnych w sieci Internet. Druga opcja dopuszczała manualne modyfikacje tekstowe zarówno opisów dokumentów, jak i samych zapytań. Uczestnicy mogli więc samodzielnie dodawać dowolnie wybrane wyrażenia, które w ich opinii wzbogacały lingwistycznie opis lub zapytanie. Zmiany te miały wyrażać potrzeby informacyjne oraz poziom wiedzy różnych potencjalnych użytkowników[24].

Wszystkie nadesłane przez uczestników odpowiedzi zostały następnie poddane procesowi oceny zgodności (ang. *relevance assessment*), przeprowadzonemu przez specjalistów, dla których język polski był językiem ojczystym. Składał się on z następujących etapów:

1. Ustalenie zakładanej potrzeby informacyjnej na podstawie zawartości pola *<description>* każdego zapytania. Opisy dostarczane w polach *<description>* zawierały informację, jakie dokumenty mogą zostać uznane za zgodne z zapytaniem. Podczas tworzenia opisów przyjęto, że odzwierciadlają one „uśrednioną potrzebę informacyjną".

---

[23] Por. *Multilingual task ChiC 2013* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/tasks/multilingual-task.

[24] Więcej o zadaniu zob. *CHiC 2013. Polish Task* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/tasks/polish-task; *Polish Track at CLEF 2013* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/; Informacje dla uczestników: *Guidelines*

2. Grupowanie zgłoszonych rezultatów. Na tym etapie z każdego nadesłanego zbioru odpowiedzi automatycznie wybierano ustaloną liczbę najlepszych rezultatów, dalej następowało usuwanie duplikatów i budowa plików odpowiedzi, które były oceniane przez specjalistów.
3. Ocena zgodności. Uwzględniając informacje dodatkowe, specjaliści oceniali każdą odpowiedź, przypisując jej jedną z trzech wartości:
   - zgodny (ang. *fully relevant*),
   - częściowo zgodny (ang. *partially relevant*),
   - niezgodny (ang. *not relevant*).

Grupowanie rezultatów i ocena zgodności odpowiedzi z zapytaniem były realizowane za pomocą systemu DIRECT (Distributed Information Retrieval Evaluation Campaign Tool)[25]. Jako główny wyznacznik skuteczności wyszukiwania przyjęto miarę MAP oraz dodatkowo miarę dokładności dla dziesięciu pierwszych odpowiedzi: P@10.

Kolekcja

Pełna kolekcja udostępniona przez Europeanę na potrzeby laboratorium ChiC obejmuje 23.300.932 dokumenty. Są to opisy obiektów prezentowanych w encyklopedii Europeana, reprezentujące ok. 80% wszystkich dostępnych zasobów europejskiej cyfrowej encyklopedii kultury na rok 2012. Kolekcja została podzielona na trzynaście zbiorów, w zależności od języka opisu, czternasty zbiór tworzą języki, dla których zarejestrowano mniej niż sto tysięcy dokumentów. Języki reprezentowane w omawianej kolekcji to: angielski, duński, francuski, grecki, hiszpański, holenderski, niemiecki, norweski, polski, słoweński, szwedzki, węgierski, włoski[26]. Pliki zbiorów dla poszczególnych języków wraz z opisem wa-

for Participation and Submission [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/Participation.html.

[25] DIRECT obsługuje również kampanie ewaluacyjne TREC; *Distributed Information Retrieval Evaluation Campaign Tool* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://direct.dei.unipd.it/.

[26] *CHiC 2013.CHiC: Cultural Heritage in CLEF*. [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/home.

runków korzystania z zasobów dostępne są na stronie http://ims.dei.unipd.it/data/chic/.

Zbiór opisów obiektów dziedzictwa kulturowego Europeany dostępnych w języku polskim składa się z 1.093.705 dokumentów i jest to dziewiąty co do wielkości podzbiór całej kolekcji. Polskie zasoby reprezentowane przez opisy dostępne w tym zbiorze obejmują 975.818 dokumentów tekstowych, 117.075 plików graficznych, 582 plików wideo oraz 230 dokumentów dźwiękowych. W opisach tych wykorzystywane są 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 procesu indeksowania analizowanych zasobów proces ograniczono do następujących pól: <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

Zapytania przygotowane na potrzeby ścieżki *Polish Task* zawierają pytania ogólne oraz szczegółowe. Wyrażono je w języku polskim wraz z dodatkowym tłumaczeniem na język angielski. Większość spośród 50 krótkich zapytań miało na celu odzwierciedlenie rzeczywistych potrzeb informacyjnych użytkowników encyklopedii Europeana. Opracowano je na podstawie logów wyszukiwań systemu europeana.eu. Ponadto, z okazji 150. rocznicy powstania styczniowego w puli znalazło się kilka zapytań dotyczących polskiej historii i terenów polskich w 18. i 19. wieku, a także pytania poświęconych konkretnym okresom historycznym i współczesnej historii Polski. Przygotowane zapytania zawierają:

1. **Zapytania chronologiczne:**

a) 8 zapytań z podanymi ramami czasowymi (18 lub 19 wiek),
b) 8 zapytań dotyczących konkretnych okresów historycznych, jak barok, dwudziestolecie międzywojenne.

2. **Nazwy własne:**
a) 12 zapytań z nazwami osobowymi (generał Józef Bem, Matka Boska),
b) 6 zapytań z nazwami geograficznymi (Kraków, pałace Lubelszczyzny),
c) 5 zapytań zawierających nazwy historyczne (powstanie styczniowe, barok).
3. **Zapytania ogólne:**
a) 5 zapytań dotyczących religii oraz wiary (diabeł),
b) 7 zapytań dotyczących funkcji lub grup społecznych (robotnicy).

Ewaluacja

Ocena trafności nadesłanych odpowiedzi została dokonana przez ekspertów, posługujących się językiem polskim jako rodzimym, na podstawie informacji przekazanych w polach <description> każdego z zapytań. Ewaluacja według skali trzystopniowej odbywała się na połączonym zbiorze wyników, nadesłanych przez uczestników ścieżki *Polish Task*. Jak już wcześniej wspomniano, z uwagi na ograniczoną objętość artykułu wyniki przeprowadzonych badań, dotyczące skuteczności metod indeksowania i wyszukiwania informacji dla dokumentów w języku polskim, zostaną przedstawione w kolejnym numerze „Toruńskich Studiów Bibliologicznych".

Podziękowania

Opisywane badania są częścią badań prowadzonych 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 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/tasks.

CHiC 2013.CHiC: Cultural Heritage in CLEF [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://www.promise-noe.eu/chic-2013/home.

CHiC 2013. Polish Task [online] [dostęp 15 grudnia 2015]. 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 15 grudnia 2015]. 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 15 grudnia 2015]. 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 15 grudnia 2015]. 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. Bohdana Ryszewskiego, Toruń 1994, s. 209–213.

Guidelines for Participation and Submission [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://members.unine.ch/jacques.savoy/Polish/Participation.html.

Jagodzinski Grzegorz, *A Grammar of the Polish Language* [online] [dostęp 15 grudnia 2015]. 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 Pamela Forner, Roberto Navigli, Dun Tufis [online] [dostęp 15 grudnia 2015]. 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 Pamela 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 15 grudnia 2015]. Dostępny w World Wide Web: http://www.clef-initiative.eu/documents/71612/0cadb163-3e32-4f16-a659-b457480c2a29.

Polish Track at CLEF 2013 [online] [dostęp 15 grudnia 2015]. 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, pod red. Witolda Doroszewskiego; oprac. i red. Czesław Pankowski, Warszawa 1995.

Swan Oskar E., *Polish Grammar in a Nutshell* [online] [dostęp 15 grudnia 2015]. Dostępny w World Wide Web: http://polish.slavic.pitt.edu/firstyear/nutshell.pdf.

Szałkiewicz Łukasz, Przepiórkowski Adam, *Anotacja morfo składniowa*, [w:] *Narodowy Korpus Języka Polskiego*, pod red. Adama Przepiórkowskiego, Mirosława Bańki, Rafała L. Górskiego, Barbary Lewandowskiej-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.

Woliński Marcin, *System znaczników morfosyntaktycznych w korpusie IPI PAN*, „Polonica" 2003, t. XXII–XXIII, s. 39–55.

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

Evaluation of IR systems efficiency. From Cranfield to TREC and CLEF labs. Genesis and methods

Abstract: We present the genesis and evolution of methods and measures of IR systems evaluation. The design of the Cranfield experiment, a long-term model for evaluation methodology, is described. Evolution of current methodology of IR systems evaluation, developed at the annual TREC (Text REtrieval Conference) is provided, and the most popular and current measures described. The article presents also design of the CLEF (Conference and Labs of the Evaluation Forum) evaluation labs with special attention paid to CHiC (Cultural Heritage in CLEF). We describe the design of Polish Task in CHiClab and discuss conclusions from lab realisation.

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