Wyszukiwanie artykułów medycznych w MEDLINE z wykorzystaniem UMLS

Julian Szymański

Streszczenie

Przedstawiono opis systemu PubMed dostarczającego narzędzi do wyszukiwania artykułów w bazie MEDLINE. Opisano powiązaną z nim ontologię medyczną UMLS z uwzględnieniem Metatezaurusa, Sieci semantycznej oraz Leksykonu. Opisany został sposób indeksowania artykułów medycznych z wykorzystaniem pojęć słownika MeSH, oraz przedstawiono główne usługi sieciowe realizujące dostęp do zasobów MEDLINE. Wykorzystując opisane zasoby UMLS przedstawiony został szkic koncepcji metody wyszukiwania, która może zostać użyta jako rozszerzenie silnika Entrez.

keywords:
bazy artykułów medycznych, słowniki maszynowe, sieci semantyczne, UMLS, PubMed, wyszukiwarki internetowe, MeSH.

1 Silnik wyszukiwania PubMed

PubMed jest silnikiem wyszukującym, działającym na bazie danych o nazwie MEDLINE, która zawiera artykuły dotyczące tematyki związanej z naukami biomedycznymi. Serwis jest dostępny przez strony internetowe pod adresem http://www.pubmed.gov. Użytkowanie systemu jest bezpłatne, jednak dostęp do większości indeksowanych artykułów wymaga subskrypcji.

1.1 Historia i stan obecny

Historia systemu PubMed, sięga roku 1986, kiedy do wyszukiwania artykułów medycznych powstała eksperymentalna baza danych korzystająca z silnika Entrez. Twórcą systemu jest *National Center for Biotechnology Information* (NCBI) przy *National Library of Medicine* (NLM) wchodzącej w skład *National Institutes of Health* (NIH), jednego z największych ośrodków naukowych na świecie. Baza danych i sam silnik rozwijają się bardzo prężnie. Na początku 2007 roku szacowano, że MEDLINE zawiera ponad 17.000.000 wpisów odnoszących się do artykułów pochodzących z około 5000 czasopism publikowanych w Stanach Zjednoczonych oraz w 80

innych krajach na świecie. W ostatnim czasie rozpoczęto również proces indeksowania rozdziałów książek o tematyce medycznej. Możliwy jest również dostęp do bazy OLDMEDLINE, zawierającej wpisy sprzed roku 1966 (z której ostatnio artykuły pochodzące z lat 1951-1966 zostały włączone do głównej bazy). Większość indeksowanych artykułów dotyczy medycyny, weterynarii, stomatologii oraz systemów opieki zdrowotnej. W bazie dostępne są również artykuły niemedyczne a dotyczące nauk pomocniczych (np. chemii). Nieliczne artykuły są dostępne w językach narodowych (tzn. nie po angielsku).

1.2 Baza danych MEDLINE

Baza danych MEDLINE stanowi główny składnik systemu PubMed. Historycznie jest ona kontynuacją pierwszego amerykańskiego elektronicznego systemu indeksowania danych o artykułach medycznych o nazwie MEDLARS (*Medical Literature Analysis and Retrieval System*) powstałego w 1966 roku.

Dla indeksowanych artykułów baza danych zawiera dodatkowe informacje mające na celu usprawnienie procesu wyszukiwania. Informacje te zorganizowane są w postaci pól określonego typu, odnoszących się do wpisów w słownikach mających cechy tezaurusów. Wyróżnić należy trzy główne leksykony:

- Chemical Name (pole /CN) - słownik nazw chemicznych,
- Cotrolled Term (pole /CT) - słownik pojęć medycznych,
- MeSH Tree Number (pole /MN) - numer węzła w ontologii MeSH, do którego przypisany jest dany artykuł. Przypisanie do węzła ma na celu powiązanie tematyki artykułu z dziedziną biomedycyny (w szczególności również z konkretną chorobą, organem ludzkim, typem komórki itp.). W dalszej części artykułu MeSH zostanie omówiony bardziej szczegółowo.

Wszystkie pola zapisane są w tzw. deskryptorze wpisu w bazie danych MEDLINE. Nie dotyczą one wpisów ze starszej bazy OLDMEDLINE oraz tych oznaczonych jako *in-process*, czyli wpisów, które zostały zapisane w bazie MEDLINE lecz nie zostały jeszcze ręcznie opracowane przez edytora, który przyporządkowuje artykułom pojęcia ze zbioru MeSH. W tabeli 1.2 przedstawiono pełen opis rekordu w bazie danych MEDLINE.

1.3 Wyszukiwanie w PubMed - formułowanie zapytań

Formułowanie złożonych, precyzyjnych zapytań jest podstawą efektywnego wyszukiwania. Ze względu na wielkość bazy danych MEDLINE, niemal każde ogólne zapytanie zwraca setki tysięcy wyników. PubMed oferuje liczne mechanizmy wyszukiwania charakterystyczne dla popularnych wyszukiwarek, takie jak np.:

Nazwa pola	Opis
AB	Abstrakt
AN	Numer dostępowy
AU	Autor
CM	Komentarz
CN	Nazwa chemiczna (wpis z tezaurusa nazw chemicznych)
CS	Źródło korporacyjne
CT	Wyrażenie kontrolne (wpisy z tezaurusa MeSH)
CY	Państwo, z którego pochodzi publikacja (zawiera również numer węzła w drzewie MeSH)
DN	Identyfikator PubMed
DT	Typ dokumentu
ED	Data wpisu
EM	Miesiąc wpisu
EML	Adres email
FS	Segment pliku
GEN	Nazwa genu (wpis z tezaurusa genów)
ISN	Międzynarodowy standardowy numer dokumentu
JT	Tytuł czasopisma
JTA	Skrócony tytuł czasopisma
JTF	Pełen tytuł czasopisma
LA	Język
NA	Nazwa osoby, której artykuł dotyczy (w przypadku biografii, bądź podobnych artykułów)
NC	Numer kontraktu/grantu
NCT	Numer testu klinicznego
NR	Numer raportu
OS	Inne źródła
PD	Data publikacji
PY	Rok publikacji
RN	Numer CAS (numer w słowniku nazw chemicznych)
SO	Źródło
ST	Wyrażenia spokrewnione
TC	Kod zabiegu
TI	Tytuł

Tabela 1: Pola rekordu opisujące artykuł bazie MedLine

- wyszukiwanie ciągu znaków oraz fraz dzięki użyciu znaków " ",
- wyszukiwanie przy pomocy tzw. *wildcards* (użycie znaku *),
- wyszukiwanie z użyciem operatorów logicznych AND, OR, NOT.

Ponadto w interfejsie www zakładka *limits* pozwala na dalsze uszczegółowienie zapytania, które jest możliwe dzięki opcjom :

- wyszukiwania po nazwisku autora,
- wyszukiwania po nazwie czasopisma,
- wyszukiwania po czasie publikacji lub czasie dodania do systemu PubMed,
- wyboru języka artykułu,
- wskazania czy szukany artykuł dotyczy ludzi, czy zwierząt,
- wskazania czy szukany artykuł dotyczy kobiet, czy mężczyzn,
- wyboru wieku osób, których dotyczy artykuł,
- wyboru typu artykułu,
- wyboru dziedziny artykułu.

Dalszym ułatwieniem w precyzyjnym wyszukiwaniu są znaczniki (*tags*) powiązane z polami rekordu opisującego artykuł w bazie MEDLINE. Dzięki nim możliwe jest nadanie znaczenia frazom zapytania wpisanym w wyszukiwarce. Najpopularniejsze z używanych znaczników to:

[au] – autor,

[dp] – data publikacji,

[la] – język,

[pg] – numer pierwszej strony artykułu,

[pmid] – PubMed ID – numer identyfikacyjny artykułu w bazie,

[pt] – typ publikacji, np.: review, article,

[ti] – słowa w tytule,

[vol] – numer tomu.

Typowy proces przeszukiwania MEDLINE jest trójstopniowy. W pierwszym etapie użytkownik wpisuje terminy (określające przedmiot wyszukiwania) z ewentualnym użyciem spójników logicznych, znaków cudzysłowu itp. W drugim etapie system próbuje przetłumaczyć podane słowa na terminy słownika MeSH lub wyrażenie logiczne złożone z kilku pojęć MeSH (sposób wykonywania tej transformacji nie jest ujawniony). Ostatni etap polega na stworzeniu listy publikacji najlepiej pasujących do wyrażenia powstałego w oparciu o zadane słowa kluczowe.

Ciekawą opcją są zapytania telegramowe (*telegram–style*). Dają one możliwość wpisywania zapytań w formie zbliżonej do języka naturalnego (np.: „State of vitreous body (of the eye) and time of death? A review, perhaps?" czy „cure for pain in back"). PubMed przeprowadza dla takich zapytań algorytm, w którym kolejne słowa są odwzorowywane na słowa znajdujące się w słowniku MeSH. Algorytm trwa do momentu, gdy liczba zwracanych wyników na zapytanie osiągnie satysfakcjonująco małą liczbę. Używanie zapytań telegramowych nie wymaga znajomości pojęć ze słownika MeSH oraz konstrukcji wykorzystujących operatory logiczne. Możliwość tworzenia zapytań w języku zbliżonym do naturalnego są ważne z punktu widzenia użytkowników systemu, którzy nie dysponują doświadczeniem i obyciem w jakiejkolwiek formie wyszukiwania internetowego, co dotyczyć może w szczególności starszych specjalistów.

Oferowane przez PubMed techniki przeszukiwania MEDLINE dają osobie wyszukującej spore możliwości precyzyjnego formułowania zapytania. Wymagają one jednak dodatkowej wiedzy o indeksowanych zasobach, której sposób organizacji i techniki wspierające zostaną przedstawione dalej bardziej szczegółowo.

2 Repozytorium pojęć UMLS

UMLS (*Unified Medical Language System*) [1] jest repozytorium pojęć dla biomedycznych słowników opracowanym przez NLM. Projekt systemu zainicjował w 1986 roku dyrektor NLM, doktor medycyny Donald Lindberg. System jest nieustannie rozwijany przez NLM, a dostęp o charakterze niekomercyjnym do zasobów UMLS jest bezpłatny po zaakceptowaniu warunków licencji UMLS[1].

UMLS można traktować również jako kompleksowy tezaurus bądź ontologię pojęć wokół biomedycyny, czy szeroko pojętego zdrowia. UMLS zawiera ponad 2 miliony nazw dla około 900.000 pojęć z ponad sześćdziesięciu słowników biomedycznych. Liczba powiązań pomiędzy pojęciami przekracza obecnie 12 milionów [2]. Poza danymi o pojęciach biomedycznych, UMLS zawiera narzędzia do efektywnego korzystania z Metatezaurusa, do tworzenia z pojęć słów – LVG (*Lexical Variants of Concept Names Generator*) i do wyodrębniania pojęć UMLS z tekstu (MetaMap).

U podstaw powstania UMLS leżą problemy związane z dużą liczbą źródeł o charakterze biomedycznym i potrzebą ich efektywnej organizacji. Ilość materiału

[1] https://kscas.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login

zwracana przy wyszukiwaniu w dużych bazach tekstowych często jest nieakceptowanie duża. UMLS stara się poprawić dostęp do biomedycznej literatury poprzez rozwój systemu, który rozumie język, jakim posługują się specjaliści z licznych dziedzin biomedycznych [3]. Za dwie główne bariery związane z przetwarzaniem języka naturalnego w postaci tekstów do pokonania uważa się:

- różnorodność sposobów, przy użyciu których te same pojęcia są wyrażane w różnych systemach i przez różnych ludzi,
- rozproszenie użytecznej informacji na wiele oddzielnych baz danych i systemów.

Jak dotychczas, w kontekście realizacji powyższych postulatów, UMLS jest największą i najbardziej zaawansowaną platformą integrującą różne słowniki źródłowe. Większość współczesnych słowników używa niekompatybilnych sposobów klasyfikowania wiedzy medycznej. Przykładowo, ta sama choroba może być klasyfikowana przez jeden słownik jako ostra lub przewlekła, inny słownik może ja organizować na podstawie miejsca jej ulokowania albo w kontekście innej choroby, z którą jest związana. Unifikacja wszystkich słowników w jedna hierarchię, jeden super-tezaurus, jest od lat jednym z priorytetowych celów informatyki medycznej, która pracuje nad powstaniem uniwersalnego rejestru medycznego (*Universal Medical Record*) dla wszystkich pacjentów.

W dalszej części artykułu omówione zostaną trzy główne składowe systemu UMLS: Metatezaurus, Sieć semantyczna oraz Leksykon SPECIALIST.

2.1 Słownikowa baza danych Metatezaurus

Metatezaurus to wielozadaniowa i wielojęzykowa słownikowa baza danych, która zawiera informacje o relacjach i pojęciach z szeroko pojętej dziedziny nauk biomedycznych. Budowana jest w oparciu o elektroniczne wersje licznych tezaurusów, klasyfikacji czy list kontrolowanych pojęć, używanych w dziedzinach opieki nad pacjentem, statystyk medycznych, indeksowania literatury biomedycznej, badań klinicznych i gromadzenia ich wyników itp. [4].

Obecnie bazę tworzą 123 słowniki źródłowe. Pełna lista słowników dostępna jest na stronie National Institute of Health[2]. Struktura Metatezaurusa pozwala na dołączanie tłumaczeń słowników na języki inne niż angielski. Największej liczby tłumaczeń doczekały się słowniki MeSH oraz International Classification of Primary Care. Wśród dotychczasowych 17 tłumaczeń językowych brakuje języka polskiego, a spora cześć zasobów jest ciągle dostępna tylko w języku angielskim.

Metatezaurus, integrując różnorodne terminologie, łączy nie tylko poszczególne słowniki, ale również reprezentowane przez nie subdomeny wiedzy biomedycznej. Do terminologii zintegrowanych z Metatezaurusem należą m.in.: NCBI Taxonomy,

[2]http://www.nlm.nih.gov/MeSH/subhierarchy2008.html

wykorzystywana przy identyfikowaniu organizmów, Gene Ontology używana do jednoznacznego opisu komponentów komórkowych, Digital Anatomist Symbolic Knowledge Base czyli cyfrowa symboliczna baza wiedzy dla anatomii, SNOMED jako repozytorium wiedzy klinicznej, OMIM – terminologia z zakresu genetyki klinicznej, oraz słownik MeSH wykorzystywany do indeksowania artykułów przez MEDLINE. Pozostałe kategorie terminologii obejmują inne specjalistyczne dyscypliny, takie jak psychiatria czy pielęgniarstwo oraz elementy systemu informacji klinicznych np.: choroby, środki farmaceutyczne, procedury. Struktura Metatezaurusa pozwala na zachowanie oryginalnych struktur wszystkich składowych terminologii, a zarazem ich wzajemne powiązanie na poziomie pojęć. Zasiąg bazy zdeterminowany jest przez zasięg słowników źródłowych: jeśli pojęcie nie pojawiło się w żadnym ze słowników źródłowych, nie znajdzie się również w Metatezaurusie. Metatezaurus zachowuje wszystkie terminy i relacje pojawiające się w integrowanych słownikach. Gdy dwa różne słowniki źródłowe używają tej samej nazwy dla różnych pojęć, Metatezaurus zapisuje obydwa znaczenia, umożliwiając wskazanie na słownik, z którego każde z nich pochodzi. Gdy to samo pojęcie pojawia się w różnych kontekstach hierarchicznych, w różnych słownikach, Metatezaurus zachowuje każdą z hierarchii. Innymi słowy: Metatezaurus nie reprezentuje wyłącznie pojedynczego kompleksowego widoku świata, nie jest też kompleksową ontologią – przechowuje różne widoki świata obecne w słownikach źródłowych, ponieważ różne widoki mogą być przydatne dla różnych zadań.

Sposób organizacji wiedzy w Metatezaurusie odróżnia go od klasycznego podejścia słownikowego opartego na pojedynczych słowach. Podstawową jednostką organizacji wiedzy jest tutaj znaczenie. Wszystkie słowa i frazy synonimiczne formułują dla danego znaczenia osobne pojęcie. Powiązania między znaczeniami reprezentowane są za pomocą relacji, które definiują semantyczne otoczenie pojęcia. Metatezaurus można zatem postrzegać jako sieć wiążącą alternatywne nazwy i określenia tego samego pojęcia i umożliwiającą identyfikacje użytecznych relacji miedzy nimi. Wszystkie pojęcia Metatezaurusa mają przypisany przynajmniej jeden typ semantyczny z Sieci semantycznej. Zapewnia to konsekwentny podział na kategorie przy relatywnie ogólnym poziomie reprezentacji w Sieci semantycznej. Wiele słów i terminów wielowyrazowych, które pojawiają się w nazwach pojęć, pojawia się również w leksykonie SPECIALIST. Narzędzia leksykonu używane są do generowania słów ze znaczeń w określonym kontekście, słów znormalizowanych oraz znormalizowanych indeksów znakowych na potrzeby Metatezaurusa.

Problem wieloznaczności słów (polisemia) rozwiązuje separacja znaczeń oraz ich umiejscowienie w różnych otoczeniach semantycznych. Słownik klasyczny zwróci dla słowa wieloznacznego pojedynczy wpis z wieloma definicjami, Metatezaurus zwróci tyle pojęć, ile znaczeń ma zadane słowo, a każde z nich osadzone zostanie w innym kontekście.

Każde pojęcie w bazie posiada atrybuty definiujące jego znaczenie: typ semantyczny, kategorię, do której przynależy, pozycje w hierarchii słownika źródłowego,

definicję.

Relacje miedzy pojęciami wynikają ze struktury słowników źródłowych lub generowane są przez edytorów Metatezaurusa. Semantykę relacji definiuje jej typ oraz opcjonalnie dołączane atrybuty. Relacje symboliczne mogą być hierarchiczne (np.: parent–child, broader–narrower_than z atrybutami is_a, kind_of oraz part_of), wynikające z hierarchii, asocjacyjne (z atrybutami location_of, caused_by, treats ...) lub definiujące stopień bliskoznaczności pojęć (similar, source asserted synonym, possible synonym). Leksykalne relacje typu synonim służą do klasteryzacji biomedycznych terminów w grupy znaczeniowe. Relacje asocjacyjne zapewniają powiązania miedzy różnymi subdomenami wiedzy biomedycznej.

2.1.1 Organizacja pojęć w Metatezaurusie

Każda nowa nazwa słownikowa, która nie zostanie znaleziona w żadnym z dotychczas przetwarzanych słowników, ani nie zostanie powiązana przez człowieka z istniejącym już w bazie pojęciem, otrzymuje swój unikatowy niezmienny identyfikator pojęcia CUI. Każda unikatowa nazwa pojęcia (łańcuch znaków) dla każdego języka posiada swój unikatowy identyfikator SUI (np. dla terminów headache, łańcuchy headache i Headache otrzymają różne identyfikatory). Każde wystąpienie danej nazwy (łańcucha znaków) w różnych słownikach źródłowych jest opatrzone unikatowym identyfikatorem AUI. Jeśli dokładnie ten sam łańcuch pojawi się w tym samym słowniku jako nazwa dla różnych pojęć (np. *cold*: w odniesieniu do temperatury, w odniesieniu do przeziębienia, czy jako akronim dla *chronic obstructive lung disease*), każde wystąpienie otrzyma unikatowe AUI. Pojedynczy identyfikator AUI jest zawsze dowiązany do pojedynczego identyfikatora pojęcia. Dla słownika angielskiego każdy łańcuch jest dowiązany do grupy wszystkich łańcuchów, które są dla siebie nawzajem leksykalnymi wariantami. Grupę taką nazywa się terminem i opatruje identyfikatorem LUI. Podobnie jak identyfikator łańcucha, LUI może być dowiązany do więcej niż jednego pojęcia. Ma to miejsce w sytuacji gdy łańcuchy, które są swoimi leksykalnymi wariantami, mają różne znaczenia. Każdy identyfikator łańcucha i każdy identyfikator atomowego wystąpienia może być dowiązany do pojedynczego LUI [5].

2.1.2 Struktura

Struktura Metatezaurusa jest wynikiem funkcji, które ma on spełniać. Głównymi funkcjonalnościami, które udostępnia są:

- odnajdywanie definicji danego wyrazu,
- odnajdywanie bliskoznaczności i synonimiczności dwóch wyrazów,
- odnajdywanie słowników źródłowych, z których dany termin pochodzi,
- umiejscawianie słowa w Sieci semantycznej systemu UMLS,

Pojęcie CUI	Terminy LUI	Nazwy SUI	Atomy AUI
C0004238 Atrial Fibrillation (preffered) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)	A0027665 Atrial Fibrillation (from MSH)
			A0027667 Atrial Fibrillation (from PSY)
		S0016669 Atrial Fibrillations	A0027668 Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	S0016899 Auricular Fibrillation (preferred)	A0027930 Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	A0027932 Auricular Fibrillations (from MSH)

Tabela 2: Przykład obrazujący organizację pojęć medycznych w Metatezaurusie

- odnajdywanie innych form gramatycznych tego samego wyrazu i sprawdzanie identyczności wyrazów z dokładnością co do formy gramatycznej,
- stwierdzanie relacji hiperonimicznych (nadrzędności) i hiponimicznych (podrzędności),
- badanie częstości współwystępowania pojęć.

Kompletna baza słownikowa dostępna jest w postaci plików tekstowych zapisanych w relacyjnym formacie: RRF (*Rich Release Format*). Każdy z plików bazy powiązany jest z jedną z czterech logicznych grup danych przedstawionych na rysunku 1 w postaci ERD. Istnieje możliwość odwzorowania struktury bazy RRF na relacyjna bazę danych oraz załadowanie zasobów UMLS do lokalnej bazy danych, a w konsekwencji uzyskanie dostępu do danych poprzez język zapytań SQL.

Rysunek 1: Główne tabele Metatezaurusa

2.1.3 Tabela MRCONSO

MRCONSO jest tablicą klasyfikującą dany ciąg znaków do określonego pojęcia. Do najważniejszych pól tej tablicy należy zaliczyć:

CUI, LUI, SUI, AUI – identyfikatory przynależności odpowiednio do pojęcia, formy fleksyjnej (terminu), ciągu znaków (nazwy), atomu,

LAT – identyfikator języka w jakim występuje termin,

ISPREF – informacja czy dany ciąg znaków jest preferowany w obrębie formy fleksyjnej,

TS – określa czy dana forma fleksyjna jest preferowana w obrębie pojęcia,

SCUI, SDUI, SAUI – identyfikatory pojęcia, deskryptora i atomu w słownikach źródłowych (opcjonalne),

STR – łańcuch znaków określający nazwę,

SAB – skrót nazwy słownika źródłowego. Np umożliwia wyodrębnienie terminów słownika MeSH używanych do indeksowania artykułów przez PubMed Central,

TTY – skrót typu terminu w słowniku źródłowym (opcjonalny),

CODE – identyfikator słownika źródłowego,

SUPPRESS – flaga zawieszenia określająca na ile dany wiersz jest aktualny (np. w jak dawnych publikacjach atom był użyty ostatni raz),

STT – typ nazwy – określa czy dana nazwa jest preferowana nazwa terminu, a jeśli nie jest, to jakim jest jego wariantem (kolejność słów, wielkość liter itp.),

SRL – poziom ograniczenia słownika źródłowego – określa, czy producent słownika nałożył dodatkowe restrykcje na jego używanie, a jeśli tak, to jakiego są typu,

CVF – flaga widoku zawartości - flaga bitowa oznaczająca wiersze włączone do danego widoku zawartości. Widok zawartości może specyfikować predefiniowany podzbiór Metatezaurusa, użyteczny dla specyficznych celów.

Ważną informacją są liczby encji w MRCONSO: 1.06 mln unikalnych pojęć (CUI), 2.07 mln unikalnych ciągów znaków (STR), 2.4 mln atomów, 14 języków, w których występują pojęcia, 99,6% haseł jest w języku angielskim, 70% pojęć wyrażanych jest jednym atomem.

Tablica AMBIGLUI zawiera 67 tys. wpisów w postaci par forma fleksyjna – pojęcie. Wpisy w tablicy AMBIGSUI występują wtedy, gdy pewne warianty w obrębie form fleksyjnych są niejednoznaczne, a inne nie. Przykładem może być tu wyraz „Alcohol", który używany jest w dwóch znaczeniach, raz jako środek sporzywczy – alkohol , lub jako związek chemiczny alkohol etylowy C2H5(OH), innym razem jako grupa związków węglowodoropochodnych z aktywną grupą OH. Jak się okazuje, inne formy fleksyjne np. alcohols używane są już jednoznacznie (w trzecim znaczeniu). Tablica ta zawiera 40 tys. wierszy.

Tablice AMBIGSUI i AMBIGLUI uzupełniają funkcjonalności Leksykonu systemu UMLS (opisanego dalej), który pozwala sprowadzić dany termin lub ciąg terminów do ich form podstawowych. Tablice AMBIGSUI i AMBIGLUI ułatwiają powiązanie uzyskanych form podstawowych z właściwymi pojęciami. Tablice te pełnią jedynie funkcje pomocnicze, gdyż tablica MRCONSO zawiera w pełni kompletne odwzorowania (w razie potrzeby redundantne) pojęcia na formy i grupy fleksyjne. Rozmiar MRCONSO znacząco wpływa na czas przetwarzania zapytań i ustalania istnienia wieloznaczności pojęć.

Opisane wcześniej tablice pozwalają na grupowanie słów w pojęcia, które jest tu jedynie abstrakcyjnym tworem reprezentowanym przez ośmio znakowy identyfikator. UMLS dostarcza również definicji pojęć. Funkcję tę pełni tablica MRDEF, której najważniejsze pola to:

AUI – identyfikator atomu, którego dotyczy definicja,

CUI – identyfikator pojęcia, którego dotyczy definicja,

DEF – treść definicji,

SAB – słownik źródłowy z którego pochodzi definicja,

SUPPRESS – flaga zawieszenia, określa stopień aktualności definicji,

ATUI – unikatowy identyfikator atrybutu (definicji),

SATUI – identyfikator atrybutu w danym słowniku źródłowym (opcjonalnie),

SAB – skrót nazwy słownika źródłowego.

Zauważyć można, że definicje przypisane są na poziomie atomów a nie na poziomie pojęć. Wynika to z faktu, że UMLS powstał z połączenia ponad 100 słowników, zdarzać się więc może, że to samo słowo występuje w różnych znaczeniach. Jeśli różnica znaczeń słowa jest wystarczająca powstaje nowe pojęcie, jednakże może zdarzyć się, że różnica ta jest zbyt mała aby ergonomiczne było wyodrębnienie osobnego pojęcia. Jednakże, aby nie gubić niepotrzebnie informacji, warto zapamiętać obie definicje. Analiza ilościowa danych dostarcza kolejnych ważnych informacji. Definicje posiada 97.5 tys atomów należących do 82.5 tys pojęć, co stanowi mniej niż 10% ogółu pojęć. Można to interpretować dwojako, po pierwsze, że stan bazy UMLS w dziedzinie definicji pozostawia wiele do życzenia, po drugie, że 90% haseł posiada tak specjalistyczne znaczenie, że podawanie ich definicji jest zbędne z punktu automatycznego przetwarzania tekstu (stopień szczegółowości jest tu o rząd wielkości za duży).

Tablica MRSAT jest jedną z największych tablic w systemie UMLS. Tablica ta zawiera skojarzenie argumentów z pojęciami bądź słowami, tak aby precyzyjniej można było zarówno określić znaczenie danego obiektu, jak i kontekst jego wystąpienia w bazie źródłowej. Pola tablicy MRSAT zawierają następujące informacje:

ATN, ATUI – nazwa atrybutu oraz jego unikatowy identyfikator,

ATV – wartość atrybutu,

CUI,LUI,SUI – identyfikatory słowa do którego odnoszą się atrybuty,

SAB – skrót od nazwy bazy z której pochodzi atrybut,

SATUI – identyfikator atrybutu w bazie źródłowej,

STYPE – nazwa kolumny tablicy MRCONSO i MRREL, do której odnosi się dany atrybut.

Rysunek 2: relacje

W opisie pola STYPE wspomniano, że atrybuty odnoszą się do dwóch tablic MRCONSO i MRREL (w praktyce do tabeli MRREL parametry odnoszą się jedynie pośrednio). Wszystkie powyższe atrybuty dotyczą sposobu transformacji z słowników źródłowych. Nie zawierają informacji stricte medycznych, a jedynie wyjaśniają aspekty techniczne budowy bazy danych systemu UMLS czy też precyzują pochodzenie określonych pojęć (np. w jakich latach były indeksowane, jaki posiadają status w rodzimych bazach itp.). Można tę tabelę obrazowo porównać do strychu systemu UMLS, są tam wszystkie te atrybuty, które nie pasowały nigdzie indziej, ale szkoda było je usuwać.

2.1.4 Relacje

Przedstawiony powyżej fragment Metatezaurusa pozwalał na wiązanie słów w odpowiadające im pojęcia i znajdowanie ich znaczeń. Istotnym elementem tezaurusa są dane wiążące pojęcia w struktury wyższego rzędu. Są to najbardziej wartościowe informacje z punktu widzenia automatycznego przetwarzania informacji.

MRREL jest tablicą zawierającą relacje między pojęciami. Do najważniejszych pól tablicy MRREL należą:

CUI1,AUI1 – identyfikator pojęcia i atomu pierwszego argumentu relacji,

CUI2,AUI2 – identyfikator pojęcia i atomu drugiego argumentu relacji,

REL – skrót określający typ relacji,

RELA – atrybut relacji,

STYPE1 – nazwa kolumny z MRCONSO, która zawiera identyfikator pierwszego pojęcia/pierwszego atomu pojawiającej/ pojawiającego się w słowniku źródłowym w kontekście danej relacji,

STYPE2 – analogicznie jak wyżej dla drugiego atrybutu relacji.

W tablicy MRREL występują nastęujące typy relacji:

RB *Broader relationship* – łączy pojęcie bardziej ogólne z mniej ogólnym,

RN *Narrower relationship* – łączy pojęcie mniej ogólne z bardziej ogólnym (przeciwne do RB),

PAR *Parent relationship* – relacja analogiczna do RB,

CHD *child* – relacja analogiczna do RN,

AQ *Allowed qualifier* – stanowi legalny kwalifikator dla,

QB *Can be qualified by* – może być kwalifikowany przez,

RQ *Related possibly synonymus* – relacja prawdopodobnie synonimiczna,

SY *Synonym* – pojęcia synonimiczne tymczasowo rozdzielone z powodu różnych źródeł,

SIB *Sibilings* – relacja łącząca dwa wierzchołki połączone relacją CHD z tym samym wierzchołkiem (wspólny rodzic),

RO *Relation Other* – relacja nieokreślona.

Szczególną uwagę należy zwrócić na pary relacji RB–RN oraz PAR–CHD. Ciężko zauważyć różnicę między ”pojęcie A jest pojęciem potomnym B” oraz ”pojęcie A jest specjalizacją pojęcia B” zgodnie z [6] relacje PAR i CHD są relacjami hierarchicznymi odziedziczonymi z innych słowników i jako takie nie zawsze muszą opisywać tego samego typu relacji we wszystkich relacjach nawet jeśli posiadają ten sam atrybut. Relacje RB–RN są natomiast zawsze jednoznaczne gdyż powstały dopiero w wyniku analizy samej bazy UMLS.

Relacje RB–RN, PAR–CHD, AQ–QB są relacjami przeciwstawnymi. Relacje RQ, SIB, SY, RO są symetryczne więc nie mają swoich relacji przeciwnych. Istotny jest również kierunek relacji, tak więc wiersz, gdzie REL przyjmuje wartość PAR oznacza, że (CUI2, AUI2) jest w relacji PAR z (CUI1, AUI1). Jeśli relacja posiada

charakter asymetryczny, to występują w tablicy MRREL wpisy określające obie relacje.

W wypadku relacji symetrycznych RQ, SIB, SY, RO występują one parami (jeśli istnieje relacja A RQ B to istnieje również B RQ A) Kolejnym atrybutem wymienionym w tabeli MRREL jest RELA. Oznacza on atrybut relacji, uściślając jej interpretacje. Najczęstsze atrybuty wraz z ich liczbami wystąpień przedstawiono poniżej:

CHD is_a (200 tys) part_of(20 tys) branch_of(5400) tributary_of (1600)

RN mapped_to(240 tys) isa (195 tys.) tradename_of (78140) part_of (24000)

QB Brak atrybutów (550 tys wystąpień)

SIB sib_in_isa (553 tys) sib_in_part_of (136 tys)

sib_in_branch_of (96 tys), sib_in_tributary_of (9000)

RQ Brak atrybutów (40 tys) alias_of(38 tys) classifies (33 tys) use (10 tys)

RO ingredient_of (191 tys) may_threat (115 tys) component_of (98000) consist_of(90 tys) dose_form_of(85 tys)

SY Brak relacji (473 tys) Expanded form of(140 tys) permuted term of(81 tys) alias_of (28 tys)

W tablicy MRREL, liczba pojęć biorących udział w jakiejkolwiek relacji wynosi 1.04 mln. Oznacza to, że pojęć niepowiązanych żadnymi relacjami jest zaledwie 43.000, liczba samych relacji to 10 mln mamy więc do czynienia z grafem rzadkim. Uwzględniając kierunkowość relacji, średnia liczba pojęć powiązanych z daną wynosi około 5. Dodatkowo istotny jest fakt, że ponad 500 tys. pojęć z zawartych w tablicy MRREL pochodzi ze słownika MeSH. Tablica ta zawiera bardzo dokładne informacje na temat relacji między pojęciami.

Tablica MRREL zawiera informacje o relacjach wiążących pojęcia. Relacje te budowane były ręcznie przez specjalistów w oparciu o ich wiedzę z różnych dziedzin nauki. Dane zawarte w tablicy MRCOC tworzone są w oparciu o analizę współwystępowania słów w w ramach tych samych publikacji. Współwystępowanie oznacza tu, że oba pojęcia zostały uznane za hasło kluczowe tego samego artykułu. Eliminuje to ryzyko przypadkowego zindeksowania pary słów jako określonych pojęć, zupełnie wbrew kontekstowi ich użycia. Głównymi polami w tabeli MRCOC są:

AUI1,CUI1 – identyfikatory atomu i pojęcia pierwszego argumentu relacji,

AUI2,CUI2 – identyfikatory atomu i pojęcia drugiego argumentu relacji,

COA – atrybut relacji,

COF – częstość relacji,

COT – typ współwystępowania.

Polem szczególnie godnym uwagi jest typ współwystępowania. Wartości przyjmowane przez pole COT mogą być następujące:

L – współwystępowanie w słowach kluczowych artykułu,

KP – współwystępowanie zgodnie z korelacją w Metatezaurus,

KN – współwystępowanie pomimo braku korelacji w Metatezaurusie,

LQB – drugie pojęcie jest kwalifikowane pierwszym i obie występują w ramach jednego artykułu,

LQ – drugie pojęcie stanowi kwalifikator pierwszego, oba występują w obrębie jednego artykułu, lub (gdy brak drugiego pojęcia) sumaryczna liczba wystąpień pojęcia,

MP – współwystępowanie pary modyfikator - problem jako przypadku klinicznego,

PP – współwystępowanie dwóch problemów (schorzeń) w obrębie jednego przypadku klinicznego.

W relacjach zawartych w tablicy MRCOC bierze udział zaledwie około 24 tys. pojęć i tworzą one aż 16.7 mln relacji, tak więc struktura relacji tego pliku okazuje się być przeciwieństwem relacji zawartych w MRREL. Graf zawiera stosunkowo niewielką liczbę pojęć, a powiązane są one dużą liczbą relacji (ponad 1.5 tys. na pojęcie), to czyni te dane trudnymi w przetwarzaniu. Jednocześnie informacje zawarte w MRCOC cenne są ze względu na możliwość oceny stopnia pokrewieństwa między danymi terminami, a tym samym lepszą ocenę relewantności dokumentu zawierającego daną terminologię.

Tablica MRHIER jest kolejną z tablic istniejących z racji genezy UMLS jako unifikacji wielu dawnych słowników. Zawiera ona informacje o tym, w jakich hierarchiach występował atom w słowniku źródłowym. Jeśli dany słownik nie był hierarchiczny, to atomy z niego pochodzące nie tworzą wierszy w MRHIER, jeśli słownik jest wieloherarchiczny to atomy z tego słownika mogą generować wiele wierszy w MRHIER Główne kolumny tablicy MRHIER przedstawiono poniżej:

CUI, AUI – unikalne identyfikatory opisywanego atomu i pojęcia,

CTX – numer kontekstu, jeśli atom występuje w wielu klasyfikacjach,

HCD – kod klasyfikujący z słownika źródłowego (jeśli takowy istnieje),

PAUI – unikalny identyfikator dla atomu nadrzędnego w klasyfikacji,

PTR – pełna ścieżka do wierzchołka klasyfikacji,

RELA – atrybut typu relacji,

SAB – słownik źródłowy.

Tablica MRHIER jest, podobnie jak wiele tablic w UMLS, nadmiarowa. Trzy kolumny CUI, AUI i PAUI dają pełną informację na temat struktury tworzonej hierarchii. Pomimo to istnieje kolumna PTR zawierająca powielone dane, które mogą być odbudowane z użyciem pozostałych kolumn tabeli.

Wartości parametru RELA, odpowiedzialnego za opis typu relacji w tablicy MRHIER może przyjmować wartości:

is_a określa, że AUI jest podpojęciem dla PAUI (509 tys.),

Part_of określa, że AUI jest elementem składowym PAUI (288 tys.),

branch_of określa, że AUI jest poddziedziną dla PAUI (5.8 tys.),

tributary_of określa, że AUI jest podproblemem dla PAUI (1.2 tys.),

NULL określa, brak wiedzy o typie relacji (427 tys.).

W tablicy MRHIER 560 tys. atomów przynależnych do 500 tys. pojęć opisanych jest przez 1.2 mln relacji hierarchicznych. Pewnym problemem jest fakt pochodzenia tych danych z 22 różnych baz, co gorsza, z bazy MeSH (indeksującej PubMed Central) pochodzi zaledwie 50 tys. relacji. Hierarchie określone tutaj podobne są do tych, które określone zostały w tablicy MRREL.

Podsumowując paragraf opisujący Metatezaurus UMLS wskazać można na jego kilka generalnych cech:

1. bardzo duży rozmiar, pokrywający znaczną część pojęć z obszaru tematyki medycznej,
2. rozdzielenie warstwy pojęć od ich reprezentacji tekstowej pozwalające na przetwarzanie dokumentów na wyższym poziomie abstrakcji,
3. istnienie relacji hierarchicznych wiążących pojęcia w bardziej złożone struktury,
4. problemy z spójnością hierarchii wynikającą z niekompletnej unifikacji baz różnej tematyki, jest to największa część systemu UMLS.

2.2 Sieć semantyczna

Sieć semantyczna zapewnia trwały i konsekwentny podział na kategorie wszystkich pojęć reprezentowanych przez Metatezaurus oraz dostarcza użytecznych relacji pomiędzy nimi [7]. Wszelkie informacje na temat pojęć zawarte są w Metatezaurusie, natomiast Sieć semantyczna dostarcza informacji na temat zbioru bazowych typów semantycznych lub też kategorii, które mogą być przypisywane tymże pojęciom. Sieć semantyczna definiuje zbiór relacji mogących zajść pomiędzy typami semantycznymi. Obecna wersja Sieci semantycznej zawiera 135 typów semantycznych[3] oraz 54 typy relacji[4]. Wychodząc od typów semantycznych najwyższego rzędu: *Entity* oraz *Event* Sieć semantyczna tworzy strukturę hierarchiczną opartą na podstawowej relacji is_a. Poprzez wprowadzenie dodatkowych relacji między węzłami reprezentującymi typy semantyczne powstaje Sieć semantyczna. Relacje niehierarchiczne podzielono na 5 głównych kategorii: *physically related to*, *spatially related to*, *temporally related to*, *functionally related to*, *conceptually related to*. Relacje asocjacyjne w sieci są słabe, tzn. relacja wiążąca typy semantyczne nie musi stosować się do wszystkich instancji tych typów, np.: relacja *evaluation_of* zachodzi miedzy typami *Sign* oraz *Organism Attribute*. Sensowna jest relacja miedzy pojęciami *overweight – evaluation of – body weight* czy *fever – evaluation of – temperature*, ale relacja *overweight – evaluation of – temperature* nie ma już sensu.

Relacje miedzy typami semantycznymi dziedziczone są za pośrednictwem relacji *is_a* przez wszystkich potomków rozpatrywanych typów, np.: relacja *process_of* umiejscowiona miedzy typami Biologic Function oraz Organism, zachodzi również miedzy typem Organ or Tissue Function (potomkiem typu Biologic Function) a typem Animal (potomkiem typu Organism). Możliwe jest jednak zaistnienie konfliktu miedzy umiejscowieniem typów w sieci a relacją, którą powinny w konsekwencji swojego umiejscowienia odziedziczyć. W takich przypadkach dziedziczenie zostaje jawnie zablokowane, np.: w wyniku dziedziczenia typy *Mental Process* oraz *Plant* powinna łączyć relacja *process_of*. W związku z faktem, iż rośliny nie są istotami świadomymi, dziedziczenie takie zostało zablokowane. Zdarza się również, że natura relacji nie pozwala na jej dziedziczenie przez potomków typów semantycznych, które łączy. Wówczas relacja zostaje oznaczona jako zdefiniowana tylko dla typów, których jawnie dotyczy a zablokowana dla ich potomków, np.: relacja *conceptual_part_of* łączy typy *Body System* oraz *Fully Formed Anatomical Structure*, ale nie powinna wiązać typu *Body System* z wszystkimi potomkami typu *Fully Formed Anatomical Structure*, takimi jak *Cell* czy *Tissue* [8].

Każde pojęcie w Metatezaurusie posiada przynajmniej jeden typ semantyczny i jest to zawsze najbardziej specyficzny typ dostępny w hierarchii, np.: pojęcie *Macaca* otrzyma typ semantyczny *Mammal*, ponieważ sieć nie przewiduje bardziej specyficznego typu (takim mógłby być np. *Primate*). Poziom ziarnistości Sieci semantycznej

[3] http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html
[4] http://www.nlm.nih.gov/research/umls/META3_current_relations.html

jest różny dla różnych jej obszarów, co wpływa istotnie na sposób interpretacji typu semantycznego dowiązanego do pojęcia. Np. typ *Manufactured Object* (poddrzewo węzła *Physical Object*) posiada dwóch potomków: *Medical Device* oraz *Research Device*. Istnieją jednak pojęcia, których nie można zakwalifikować do żadnej z tych kategorii, chociaż w ogólności zaliczają się do *Manufactured Object* – w takim przypadku zostaną dowiązane do bardziej ogólnego typu *Manufactured Object*.

2.2.1 Struktura sieci semantycznej

Siec semantyczna dostępna jest w postaci relacyjnego formatu RTF (*Relational Table Format*). Cała struktura sieci zorganizowana jest w czterech głównych plikach: SRDEF, SRSTR, SRSTRE1, STSTRE2, oraz w dwóch pomocniczych SRFIL i SRFLD, które zawierających informacje o zawartości poszczególnych plików i zawartości poszczególnych kolumn. Wszystkie typy semantyczne i relacje identyfikowane są za pomocą 4-znakowego identyfikatora postaci „T001".

Tabela SRDEF zawiera poszerzone informacje na temat znaczenia danych typów semantycznych oraz relacji. Poszczególne pola oznaczają:

RT – typ rekordu: STY(Semantic Type) lub RL(Relation),

UI – unikalny identyfikator typu semantycznego lub relacji,

STY/RL – Unikalny identyfikator hierarchiczny; Numer w drzewie typów semantycznych lub relacji,

STN/RN – nazwa typu lub relacji,

DEF – opis znaczenia typu lub relacji,

EX – przykładowe obiekty z tezaurusa o tym typie,

UN – uwagi dotyczące dowiązania do danego typu semantycznego do pojęcia (dotyczy tylko typów semantycznych),

ABR – skrót typu semantycznego lub relacji,

RIN – relacja odwrotna (dotyczy tylko wybranych relacji), np. dla relacji *affects* pole RIN = *affected_by*.

Informacje o strukturze sieci przechowuje tablica SRSTR . Zawiera ona informacje o relacjach łączących typy semantyczne. Wiersze tej tablicy opisują relację między parą typów lub relacji.

STY/RL – typ elementu STY(Semantic Type) lub RL(Relation)

RL – nazwa relacji. Np. is_a jeśli jest to relacja hierarchiczna

LS – status relacji (D - dziedziczna,DNI -niedziedziczna)

Tablice SRSTRE1 oraz STSTRE2 są rozszerzeniem informacji z tablicy SRSTR. Dostarczają kompletnego zbióru powiązań między pojęciami z uwzględnieniem dziedziczenia. Dane przechowywane są tu w postaci: typ semantyczny – relacja – typ semantyczny. Na pierwszą tablice składają się trójki identyfikatorów, a na druga zawiera trójki nazw:

UI/STY – argument 1 powiązania: identyfikator lub nazwa typu semantycznego (identyfikator UI z tablicy SRDEF dla wskazanego typu semantycznego, STY tekstowa nazwa z tablicy SRDEF dla pierwszego typu semantycznego)

UI/RL – relacja: identyfikator lub nazwa niehierarchicznej relacji,

UI/STY – argument 2 powiązania: identyfikator lub nazwa typu semantycznego.

Dane kojarzące pojęcia Metatezaurusa z typami semantycznymi składowane są w tabeli MRSTY.

CUI – unikatowy identyfikator pojęcia

TUI – unikatowy identyfikator typu semantycznego

STN – numer strukturalny typu semantycznego

STY – opis tekstowy typu semantycznego

Każde pojęcie z UMLS posiada w tablicy MRSTY co najmniej jeden wiersz, oznacza to 1.06 mln unikatowych identyfikatorów pojęć. Tablica MRSTY może wiązać pojęcie do więcej niż jednego typu semantycznego, jak się okazuje, nie jest to zjawisko częste – łączna liczba dowiązań to 1.25 mln, zaś pojęć należących do więcej niż jednej kategorii jest 186 tys.

Ważnym parametrem jest oczekiwana wielkość grupy dla losowego parametru. Wynosi ona znacznie więcej niż minimalna możliwa wartość 1.25 mln/135 = 9259, aż 81060, jest to wartość bardzo duża, wskazująca, że rozmiar Sieci semantycznej jest dalece niewystarczający do przeprowadzania jakiegokolwiek wnioskowania, może ona stanowić co najwyżej wskazówkę dla laików w dziedzinie medycyny odnośnie klasyfikacji pojawiających się pojęć. Jakiekolwiek wykorzystanie Sieci semantycznej do innych celów wymagałoby jej uszczegółowienia, o czym wspomina [9]. W tym miejscu należy nadmienić, że 135 typom semantycznym towarzyszy ponad 6200 relacji typu is_a, *cause*, *co-ocurs_with* i innych, które mogłoby być cennym źródłem informacji, niestety wielkość pojedynczych domen (typów semantycznych) powoduje, że powyższe relacje nie zostały użyte w systemie.

2.3 Leksykon SPECIALIST i narzędzia językowe

Trzecią częścią systemu UMLS jest Leksykon. Leksykon został stworzony w celu dostarczenia informacji i narzędzi niezbędnych dla przetwarzania języka naturalnego (NLP). Nad budowa i utrzymaniem Leksykonu pracuje zespół lingwistów przeszkolonych przez NLM. Jego tworzenie wspomaga narzędzie LexBuild, które zapewnia kompletność oraz poprawność ręcznie tworzonych wpisów oraz ich zgodność z przyjętą strukturą wpisu leksykalnego.

Wpis leksykalny posiada strukturę opartą na ramach [10]. Tworzą ją tzw. sloty oraz ich wypełnienia. Każdy rekord posiada formę bazową (*slot base=*), kategorię (*slot cat=*) oraz unikatowy identyfikator wpisu – EUI (*slot entry=*). Ponieważ w języku angielskim niektóre pojęcia składają się z kilku słów, np. *ice cream*, dlatego aby unikać niepotrzebnych konfliktów miedzy różnymi znaczeniami słowa, leksykon operuje na elementach leksykalnych, które pokrywają zarówno pojedyncze jak i wielowyrazowe jednostki leksykalne. Wpis leksykalny posiada informacje o ortografii, morfologii oraz składni dla elementu leksykalnego, którego dotyczy.

Możliwe warianty pisowni elementu leksykalnego wskazuje *slot base=* (gdy element ma tylko jeden wariant pisowni) oraz *spelling variant=* (dla alternatywnych wariantów pisowni). Leksykon rejestruje tylko dozwolone warianty pisowni.

Leksykon, podobnie jak Metatezaurus, dostępny jest w postaci plików w relacyjnym formacie RRF pozwalającym na zmapowanie na relacyjną bazę danych.

Leksykon zawiera przede wszystkim dane o słownictwie angielskojęzycznym. Zawiera on zarówno słowa powszechnego użytku jak i specjalistyczną terminologię biomedyczną. Leksykon dostarcza zarówno informacji składniowych, ortograficznych jak i morfologicznych o terminach w nim zawartych. Cztery zasadnicze części to forma bazowa, część mowy, unikatowy identyfikator i wszelkie możliwe warianty pisowni. Leksykon SPECIALIST ma za zadanie być generalnym anglojęzycznym leksykonem zawierającym liczne terminy biomedyczne, typowe słownictwo używane w języku angielskim, pojęcia znalezione w bazie danych MEDLINE oraz w UMLS.

Przez Leksykon rozumie się również zbiór dołączonych narzędzi służących do przetwarzania języka naturalnego. W skład narzędzi wchodzą:

- *Normalizer* służący do sprowadzania ciągu słów do formy podstawowej. Normalizacja polega m.in. na usunięciu końcówek fleksyjnych, znaków interpunkcyjnych, zmianie dużych liter na małe, usunięciu słów pospolitych (znajdujących się na tzw. stop liście).
- *Word Index generator* służący do rozbijania ciągów słów na zbiór pojedynczych wyrazów (terminów) zaindeksowanych w leksykonie. Umożliwia on przetworzenie ciągu znaków celem otrzymania formy odpowiedniej do wyszukiwania w indeksie słów.
- *Lexical wariant generator* służący do generowania wszystkich możliwych form gramatycznych z podstawowej formy gramatycznej słowa.

2.4 UMLS a PubMed

PubMed umożliwia dostęp do największego repozytorium biomedycznych artykułów naukowych w bazie MEDLINE. W tym celu wykorzystuje specjalizowany indeks pozwalający określić poszczególne zasoby biblioteczne. Do indeksowania artykułów w MEDLINE wykorzystuje się specjalnie stworzony na tę potrzebę i regularnie aktualizowany hierarchiczny słownik *Medical Subject Headings* (MeSH), wchodzący w skład źródłowych słowników integrowanych przez Metatezaurus. Pojęcia słownika MeSh używane są jako słowa kluczowe artykułów mające oddawać najlepiej charakter i treść indeksowanej publikacji. Artykuły przesłane za pomocą PubMedu drogą elektroniczną, otrzymują status *PubMed – as supplied by publisher*. Podczas indeksowania status zmienia się na *PubMed – in process*, a po ostatecznym dołączeniu do bazy MEDLINE - *indexed for MEDLINE*.

Dzięki MeSh, dysponując nazwą pojęcia, mechanizm kros-referencji umożliwia użytkownikowi systemu dostęp do literatury dotyczącej wskazanego zagadnienia. Przed wykonaniem zapytania, PubMed wykorzystując mechanizm automatycznego mapowania terminu, próbuje przetłumaczyć treść zapytania i w miarę możliwości przypisać słowa zapytania do terminów słownika MeSH. Dysponując pojęciami silnik wyszukiwania tworzy zapytanie, dla którego po przeanalizowaniu tabeli odpowiedników MeSH (*MeSH Translation Table*), proponowane jest bardziej precyzyjne zapytanie. W rezultacie użytkownik końcowy otrzymuje zbiór odnośników do artykułów opatrzonych identyfikatorem PMID oraz skojarzonymi z innymi pojęciami MeSH. Dysponując nazwami skojarzonych z zapytaniem pojęć oraz posługując się operatorami logicznymi, można skonstruować bardziej szczegółowe zapytanie, które zwróci nie tak liczne, ale bardziej trafne rezultaty.

2.5 słownik pojęć MeSH

Uważa się, że liczbę osób piszących artykuły naukowe należy szacować w milionach. Ten fakt w sposób istotny utrudnia proces indeksowania artykułów i wymaga ujednolicenia sposobu używania słów kluczowych w swoich artykułach [11].

Próbą rozwiązania tego problemu jest MeSH (*Medical Subject Heading*)– kontrolowany słownik pojęć i powiązań między nimi. Jego zadaniem jest umożliwienie efektywnego i spójnego indeksowania, katalogowania oraz wyszukiwania artykułów i książek. Historia MeSH sięga połowy XX wieku. Pierwsza edycja MeSH pochodzi z roku 1960 i zawierała 4400 deskryptorów. Trzy lata później liczba deskryptorów wynosiła 5700. Obecnie[5] zawiera ich 24,767.

MeSH jako kontrolowany słownik jest zbiorem deskryptorów (*descriptors*) opisujący pojęcia. Zbiór ten jest zorganizowany zarówno w sposób hierarchiczny, jak i alfabetyczny. Wyróżniamy trzy zasadnicze typy rekordów MeSH: deskryptory, kwalifikatory, pomocnicze rekordy konceptualne (*Supplementary Concept Records* –

[5] 2008 rok

SCRs).

Hierarchiczne uporządkowanie ma formę drzewa, w którym położenie identyfikowane przez specjalne oznaczenie, zwane *Tree Number*. Ustrukturyzowanie słownika MeSH w postaci drzewa ma bardzo istotny wpływ na proces wyszukiwania artykułów o tematyce medycznej. Dzięki ścisłej hierarchizacji pojęć możliwe jest nie tylko wyszukanie artykułów związanych ze słowem wpisanym przez użytkownika, lecz również z pojęciami powiązanymi z zadanym określeniem, np: bardziej szczegółowymi frazami z danej dziedziny medycyny. Np. wpisanie pojęcia „układ pokarmowy" spowoduje nie tylko wyszukanie artykułów indeksowanych pojęciem „układ pokarmowy", ale również takich, które opisane są pojęciami „żołądek", „przełyk", „dwunastnica" itp. Fakt, że frazy wyszukiwawcze są wzbogacane o pojęcia zależne, pociąga za sobą potrzebę przeszukiwania drzewa pojęć w głąb, co jest kosztowne obliczeniowo. W celu zapewnienia efektwności przetwarzania danych ograniczane jest przeglądanie drzewa hierarchii do 11 poziomów. Ten rodzaj wyszukiwania nazywany jest wyszukiwaniem typu *concept exploding*.

MeSH jest zapisywany w jednym z dwóch formatów : XML lub ASCII[6].

Deskryptory (*Main Headings*) używane są do indeksowania cytowań w bazie danych MEDLINE, do katalogowania publikacji i innych baz danych. Deskryptor zawiera takie informacje jak: Nazwa opisywanego pojęcia Identyfikator niepodlegający modyfikacjom Opis pojęcia Pojęcia spokrewnione. Lista synonimów i pojęć bliskoznacznych Kwalifikatory - pojęcia używane do katalogowanie (grupowania) pojęć zgrupowanych w drzewie MeSH.

Deskryptory są poddawane zmianom raz do roku, chociaż jeśli sytuacja tego wymaga, w niektórych sytuacjach mogą być aktualizowane częściej. Deskryptory posiadają oznaczenie *Tree Number* (TR) określające pozycję w drzewie pojęć MeSH. Deskryptory mogą się znajdować w wielu miejscach w drzewie hierarchii MeSH jednocześnie, stąd mogą posiadać wiele "TR". Każdy deskryptor ma dodatkowy numer identyfikacyjny, co do którego mamy pewność, że nigdy nie ulegnie zmianie. Deskryptory zawierają krótki opis, listę deskryptorów powiązanych, linki do nich oraz listę synonimów lub pojęć bardzo podobnych (w PubMed oznaczonych jako „*entry terms*"). Synonimy, formy alternatywne i bardzo blisko spokrewnione pojęcia dla zadanego rekordu MeSH używane są zamiennie z pojęciami zadanymi w wyszukiwaniu. Wprowadzenie ich miało na celu poprawę znajdywania poprzez zwiększenie liczby odwołań, po przez które artykuły są indeksowane.

Charakterystyki publikacji są wyróżnionymi deskryptorami. W przeciwieństwie do deskryptorów MeSH, nie wskazują na to, o czym jest obiekt, lecz przechowują informację o tym, czym obiekt jest (np. wskazując na rodzaj artykułu). Są to w zasadzie metadane (dane o danych), nie wskazujące na rzeczywistą zawartość. Dlatego nie wyszukujemy ich przy użyciu znacznika [MH], lecz specjalnego znacznika [PT] (*Publication Type*). Są wyspecyfikowane w kategorii "V" struktury drzewiastej

[6] http://www.nlm.nih.gov/MeSH/filelist.html

MeSH.

Deskryptory geograficzne (*Geographics*) zawierają dane na temat kontynentów, krajów, regionów, stanów itp. Są używane do charakterystyki lokalizacji obiektu. Znajdują się w kategorii "Z" struktury MeSH.

Obok powiązań hierarchicznych między deskryptorami mogą występować jeszcze relacje typu: *see related* określająca niesynonimiczne powiązane artykułu,

consider also stosowana najczęściej do wskazywania relacji pomiędzy deskryptorami o powiązanych korzeniach lingwistycznych.

entry Combination określająca ona kombinacje deskryptorów/kwalifikatorów, które nie są dozwolone. Zakazana kombinacja jest zapisana w elemencie <EntryCombination> XML MeSH lub w adekwatny sposób w formacie ASCII.

Obok deskryptorów głównych w Mesch występują tzw. kwalifikatory (zwane również podnagłówkami (*subheadings*)), które jest używane do indeksowania i katalogowania w połączeniu z deskryptorami. Kwalifikatory pozwalają na wygodne grupowanie tych cytowań, które są związane z pewnym konkretnym aspektem tematu. Dla przykładu: „*Liver/drug effects*" wskazuje, że szukamy artykułu na temat wątroby, ale wyszukiwanie jest zawężone do temat wpływu leków na wątrobę poprzez wyszukiwanie odpowiednich kwalifikatorów, oznaczanych znacznikiem [SH]. Istnieją obecnie 83 kwalifikatory, a ich szczegółową hierarchię kwalifikatorów można znaleźć na stronach National Institute of Health[7]. .

Do indeksowania związków chemicznych, lekarstw i innych pojęć dla bazy danych MEDLINE używane są tzw. pomocnicze rekordy pojęciowe *Supplementary Chemical Records* wyszukiwane przy użyciu znacznika [NM]. Ich uporządkowanie jest nieco nietypowe, nie są bezpośrednio związane z drzewiastą strukturą MeSH, nie posiadają *Tree Numbers*, lecz są połączone z jednym lub wieloma deskryptorami przy użyciu znacznika *HeadingMappedTo*. Są aktualizowane co tydzień, a więc dużo częściej niż inne rekordy. Obecnie istnieje ponad 170.000 pomocniczych rekordów pojęciowych.

2.6 Usługi Entrez

Obok interfejsu użytkownika realizującego dostęp do bazy MEDLINE przez WWW realizowany jest przez zestaw usług Web Service wchodzących w skład Entrez-Utilities.

Entrez Programming Utilities stanowią zbiór narzędzi umożliwiających dostęp do globalnego systemu wyszukiwawczego Entrez. Jedną z baz dostępnych z poziomu tych usług jest baza MEDLINE, dostępny jest również jej podzbiór, z którego korzysta PubMed Central. Wszystkie narzędzia zorganizowane są jako usługi sieciowe: zapytanie następuje poprzez przekazanie w url odpowiednich parametrów, zaś odpowiedzią jest plik XML o określonej strukturze charakterystycznej dla danego narzędzia. Entrez realizuje następujące usługi:

[7] http://www.nlm.nih.gov/MeSH/subhierarchy2008.html

- EInfo – usługa zwraca podstawowe informacje statystyczne na temat danej bazy danych takie jak: liczba publikacji liczba zaindeksowanych terminów, liczba autorów, czasopism źródłowych.

- ESearch – usługa zwraca identyfikatory artykułów, które spełniają zadane kryteria. Najważniejszymi parametrami usługi są:

  db – nazwa bazy danych, do której kierowane jest zapytanie,

  term – termin indeksowany lub wyrażenie logiczne złożone z terminów logicznych, dla których poszukujemy listy artykułów,

  field – pole w którym należy szukać terminów z parametru term np. auth (Autor), MeSH (hasła przedmiotowe),

  retstart – pole mówiące, od którego rekordu należy rozpocząć wypisywanie, przydatne gdy artykułów jest bardzo dużo i pożądane jest stronicowanie,

  retmax – pole mówiące ile rekordów ma być zwrócone na jednej stronie.

- ESummary – usługa zwraca podstawowe informacje na temat artykułu, którego identyfikator podany został jako parametr. Informacjami tymi są: autor, tytuł, wydawca, data wydania, dostępność abstraktu oraz pełnej wersji artykułu. Parametrami do wykonania usługi ESummary są m.in.:

  id – lista identyfikatorów artykułów umieszczana po przecinkach,

  retmax, retstart, db – rola parametrów identyczna jak w ESearch.

- EFetch – działanie tego narzędzia zależy od bazy, na której wykonane jest zapytanie (parametr db). Dla bazy PubMed Central zwracana jest cała dostępna zawartość artykułu (lub artykułów) o podanym id. EFetch zwraca informacje bibliograficzne w formacie DublinCore[12] oraz tekst artykułu (dostępny dla bazy PubMed Central), jak również link, pod którym znajduje się oryginalny zasób. EFetch używa następujących parametrów:

  id, retmax, retstart, db - rola parametrów identyczna jak w ESearch.

- EGQuery - usługa zwraca liczbę obiektów w różnych bazach danych o treści opisywanej hasłem podanej jako parametr zapytania.

- ESpell - usługa służy poprawianiu literówek w hasłach, parametrami zapytania są:

  db - baza danych z której ma pochodzić hasło,

  term - termin podany do poprawienia.

3 Propozycja udoskonalenia silnika wyszukującego

Sieć semantyczna organizująca pojęcia w obrębie biomedycznej dziedziny oraz powiązania artykułów poprzez ich słowa kluczowe z pojęciami Sieci semantycznej może zostać wykorzystana do poprawy jakości wyszukiwania dokumentów w bazie MEDLINE.

Wspomniane wyżej wyszukiwanie typu telegramowego, mające zwracać najlepiej dopasowane artykuły do zapytania użytkownika jest postulowanym idealnym rozwiązanie w systemach wyszukiwawczych. Realizacja jednak tego zadania jest trudna, między innymi z uwagi na fakt, że silnik wyszukiwania zakłada, że zapytanie składa się ze słów kluczowych, których używa się do wskazania artykułów.

Jedną z modyfikacji, którą można zrealizować w istniejącym obecnie wyszukiwaniu jest dodanie do mechanizmów wyszukiwania możliwości doprecyzowywania poszukiwanych treści oraz wprowadzeniu rankingu dokumentów spełniających zadane kryteria wyszukiwawcze.

Koncepcja zaprezentowanego tu rozszerzenia polega na dostarczeniu końcowemu użytkownikowi narzędzia, dzięki któremu będzie mógł on precyzyjniej określić systemowi informatycznemu jakiego rodzaju treści są dla niego interesujące. Realizacja tego zadania może się odbyć przy użyciu interaktywnej wizualizacji fragmentu Sieci semantycznej UMLS, w której w pierwszym etapie aktywowane są pojęcia najbardziej pasujące do zapytania użytkownika. Możliwość wskazania skojarzonych z nimi pojęć pasujących do wyszukiwanego tematu w precyzyjny sposób umożliwia zawężanie zbioru wyszukiwanych dokumentów.

W związku z możliwością wystąpienia dość znacznej liczby pojęć skojarzonych z wyszukiwanym zagadnieniem poprzez obliczenie zysku informacyjnego związanego z każdym z nich możliwe jest zawężenie tego zbioru do pojęć najbardziej istotnych w sensie klasyfikacji określonych dokumentów. Drugą modyfikacją mogąca w znaczny sposób poprawić trafność wyszukiwania artykułów w bazie MEDLINE, jest wprowadzanie w podzbiorze dokumentów będących rezultatem zwracanym końcowemu użytkownikowi porządku określającym ich istotność. W wyszukiwarce Google ranking dokumentów realizowany jest w oparciu o algorytm PageRank [13] działający na grafie powiązań między stronami. Analogiczny algorytm można przeprowadzić na sieci powiązań między pojęciami do których dołączone są artykuły. Wprowadzona na tej podstawie miara popularności określać może istotność zadanego artykułu dla osoby wyszukującej. Rozważyć można również szereg możliwości modyfikacji takiego podejścia wprowadzając zamiast globalnego określenia popularności dokumentu (jak ma to miejsce w PageRank [14]) miarę wiążącą popularność dokumentu dla określonej pytaniem użytkownika poddziedziny, jak ma to miejsce np. w algorytmie HITS [15].

4 Podsumowanie

Dynamiczny rozwój portalu PubMed spowodował, że ma on bardzo duży wpływ na świat dzisiejszej medycyny, poprzez zwiększenie dostępności artykułów dla specjalistów na całym świecie i łatwość wyszukiwania wiedzy na określony temat. Pewnym dowodem popularności, a więc pośrednio również użyteczności serwisu jest fakt, że już w 2004 system odnotowywał do 1300 odwołań na sekundę. Częstotliwość ta stale rośnie w tempie gwałtowniejszym niż rozmiar samego archiwum. PubMed staje się w sposób zauważalny „wyrocznią" świata medycznego, głównym źródłem informacji i wiedzy, podobnie jak ma to miejsce w przypadku Google czy Wikipedii. Obecnie zarówno specjaliści amerykańscy, jak i światowi zgodnie twierdzą, że publikacja artykułu w czasopiśmie nie indeksowanym przez bazę danych MEDLINE i nie obsługiwanym przez wyszukiwarkę dostępną na portalu PubMed mija się z celem. Tylko umieszczenie artykułu w pismach, które znajdują się na liście pism indeksowanych przez PubMed daje szansę na to, że praca będzie dostrzeżona.

W obliczu lawinowo rosnących zasobów w sieci rozwój metod efektywnego dostępu do informacji tekstowej staje się coraz bardziej palącym problemem. Pewną próbą organizacji tych zasobów jest wyszukiwarka Google, indeksująca znaczną część zasobów sieciowych na świecie. Jedną z niedogodności jest jednak brak możliwości doprecyzowywania tematu wyszukiwania. Przedstawiona w artykule koncepcja uzupełnienia tej niedogodności zastosowana może zostać dla bazy MEDLINE z wykorzystaniem UMLS. Pozytywne zweryfikowanie proponowanej idei wyszukiwania w ograniczonej dziedzinie medycznej, w kolejnym kroku pozwoli na wykorzystanie jej w wyszukiwarkach ogólnego zastosowania.

Literatura

[1] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research (**32**) 267–270

[2] Kleinsorge, R., Tilley, C., Willis, J.: Unified Medical Language System (UMLS) Basics. (online, http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf)

[3] Humphreys, B., Lindberg, D., Schoolman, H., Barnett, G.: The Unified Medical Language System An Informatics Research Collaboration (1998)

[4] Humphreys, B., MLS, D., Schoolman, H., Barnett, G.: Focus on The UMLS. Journal of the American Medical Informatics Association **5** (1998) 1

[5] Lindberg, D., Humphreys, B., McCray, A.: The Unified Medical Language System. Methods Inf Med **32** (1993) 281–91

[6] Bodenreider, O.: Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. In: Proceedings of AMIA Annual Symposium. Volume 57. (2001) 61

[7] McCray, A.: The UMLS semantic network. In: 13. Annual Symposium on Computer Applications in Medical Care. (1989) 503–507

[8] Kumar, A., Schulze-Kremer, S., Smith, B.: Revising the UMLS Semantic Network. Medinfo (2004) 7–11

[9] Chen, Z., Perl, Y., Halper, M., Geller, J., Gu, H.: Partitioning the UMLS semantic network. Information Technology in Biomedicine, IEEE Transactions on **6** (2002) 102–108

[10] Minsky, M.: Frame-system theory. Thinking: Readings in cognitive science (1977) 355–376

[11] Nelson, S., Johnston, D., Humphreys, B.: Relationships in Medical Subject Headings. Relationships in the organization of knowledge (2001) 171–184

[12] Core, D.: Dublin Core Metadata Initiative. online, http://www. dublincore. org (2004)

[13] Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems **30** (1998) 107–117

[14] Langville, A., Meyer, C.: Deeper inside pagerank. Internet Mathematics **1** (2004) 335–380

[15] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46** (1999) 604–632