Automatyczna budowa semantycznego modelu objawów chorobowych na bazie korpusu słownego

G. SZOSTEK, M. JASZUK, A. WALCZAK
grazyna.szostek@gmail.com

Wydział Cybernetyki Wojskowej Akademii Technicznej w Warszawie
Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

Opisane w artykule badania dotyczą danych z dziedziny medycyny. Wyniki badań diagnostycznych rejestrowane są na różne sposoby. Mogą mieć postać tabel, wykresów, obrazów. Niezależnie od oryginalnego formatu danych możliwe jest sporządzenie ich opisu słownego, który koncentruje się na opisie zaobserwowanych objawów chorobowych. Opisy takie tworzą korpusy słowne dotyczące poszczególnych technologii diagnostycznych. W podobny sposób zapisywana jest wiedza dotycząca jednostek chorobowych. Ma ona postać korpusów tekstowych, w których zawarte są opisy objawów specyficznych dla poszczególnych schorzeń. Za pomocą narzędzi przetwarzania języka naturalnego możliwe jest automatyczne wydobycie z tekstów modeli semantycznych, opisujących poszczególne technologie diagnostyczne oraz choroby. Pewne utrudnienie stanowi fakt, że wiedza medyczna może zostać zapisana w języku naturalnym na wiele sposobów. Zastosowanie formatu semantycznego pozwala wyeliminować te niejednoznaczności zapisu. W konsekwencji dostajemy ujednolicony model wiedzy medycznej, zarówno od strony wyników technologii diagnostycznych opisujących stan pacjenta, jak i wiedzy dotyczącej jednostek chorobowych. Daje to możliwość dokonania fuzji danych pochodzących z różnych źródeł (danych heterogenicznych) do postaci homogenicznej. Artykuł przedstawia metodę generowania modelu semantycznego wiedzy medycznej, wykorzystującą analizy leksykalne korpusów słownych.

Słowa kluczowe: sieć semantyczna, ontologia, przetwarzanie języka naturalnego.

1. Wprowadzenie

Szybki wzrost ilości informacji i mała skuteczność metod do ich przetwarzania dały początek pracom nad metodami opartymi zarówno na formalnej reprezentacji wiedzy – ontologii, jak i na sieciach semantycznych. W związku z dużą ilością pojęć i zależności między nimi co raz częściej wykorzystuje się automatyczną konstrukcję semantycznego modelu [2], [10]. Podstawowym zasobem wykorzystywanym przez takie metody jest korpus tekstów. Metody przetwarzania języka naturalnego [1], [3], [4], [5] opracowane dla korpusów tekstów w języku angielskim nie mają bezpośredniego przełożenia na języki fleksyjne o swobodnym szyku wyrazów w zdaniu, takie jak język polski. Prace nad konstrukcją sieci semantycznej dla języka polskiego [6], [9] przyczyniły się do powstania automatycznych metod wykrywania leksykalnych relacji semantycznych.

W wielu dziedzinach dane (np. wojskowe, ekonomiczne, medyczne) nie mają jednolitej postaci: tabela, obraz, wykres itd. Każda z nich wymaga stosowania dedykowanych metod przetwarzania i analizy. Od kilku lat sieci semantyczne są wykorzystywane jako narzędzie do jednorodnego zapisu heterogenicznych danych [8]. To podejście odkrywa nowe możliwości przetwarzania danych, zastosowanie tych samych metod analizy do obrazów, tekstów, tabel itd. Format zapisu danych powoduje zmniejszenie ich rozmiaru (np. obrazowych danych), co ma wpływ na szybkość przesyłania danych przez sieć.

Dane medyczne są dobrym przykładem różnorodności zapisu danych. Opisana w pracy budowa modelu semantycznego jest elementem szerszych badań mających na celu dokonanie fuzji danych pochodzących z różnych źródeł (danych heterogenicznych) do postaci homogenicznej. Ten sam objaw może mieć różną postać (fraza w tekście, element tabeli, fragment obrazu, punkt na wykresie), co stanowi problem dla dalszego przetwarzania takiej informacji. Istnieje więc potrzeba zapisu objawów w postaci jednolitej, co daje możliwość zastosowania takich samych metod i narzędzi matematycznych w procesie wspomagania stawiania diagnozy. Podstawą do budowy modelu semantycznego objawów są opisy wyników technologii diagnostycznych i opisy jednostek chorobowych, które tworzą korpus słowny. W artykule zostanie zaprezentowana metoda

Praca została zrealizowana w ramach projektu POIG.02.03.03-00-013/08, dofinansowanego ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach Programu Operacyjnego „Innowacyjna Gospodarka”.

generowania modelu semantycznego na bazie korpusu słownego. Technika ta zostanie wykorzystana do budowy ontologii poszczególnych technologii diagnostycznych i jednostek chorobowych w celu ujęcia w jednolitą strukturę zapisu wiedzy diagnostycznej.

Układ artykułu jest następujący. W sekcji 2 jest pokazana różnorodność form zapisu danych medycznych i propozycja homogenicznej postaci ich zapisu. Sekcja 3 stanowi krótkie wprowadzenie do modelu formalnego ontologii i sieci semantycznej. Następnie, w sekcji 4 jest szczegółowo omówiony proces wykrywania fraz oznaczających objawy i cech je specyfikujących. Sekcja 5 podsumowuje wyniki uzyskane w sekcji 4. Ogólny opis procesu budowy ontologii jest przedstawiony w sekcji 6.

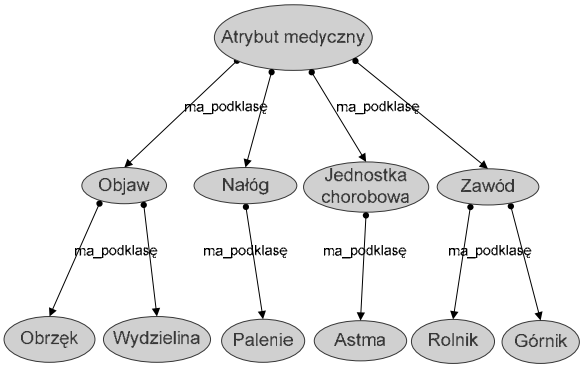
2. Reprezentacja heterogenicznych danych

Opis stanu zdrowia pacjenta i opis jednostki chorobowej składają się z opisów objawów zdiagnozowanych za pomocą różnych technologii. W diagnostyce chorób są m.in. stosowane następujące technologie diagnostyczne (TD):

- badanie podmiotowe
- przedmiotowe
- testy skórne
- gazometria
- RTG
- scyntygrafia
- ultrasonografia
- spirometria
- bronchoskopia
- badanie laboratoryjne itd.

W zależności od TD wyniki mają różną postać. Wyniki mogą być zapisane w postaci tabelarycznej (spirometria, badanie morfologiczne krwi), w postaci wykresu (spirometria), obrazu (RTG), opisu słownego (badanie podmiotowe) itd. Badania obrazowe najczęściej mają dodatkową postać wyniku – opis słowny wykonany przez lekarza specjalistę. Wraz z wykresami mogą być dostarczone najistotniejsze parametry wykresu (w postaci tabeli). Podsumowując, wyniki TD, ze względu na formę zapisu, można podzielić na trzy postacie: tabelaryczną, opisu słownego, pliku grafiki cyfrowej lub analogowej.

W ramach danej TD dokonuje się pomiaru/ obserwacji wielu parametrów związanych ze stanem pacjenta. Można stwierdzić istnienie objawów chorobowych, ale można także zarejestrować dużo dodatkowych faktów dotyczących pacjenta. Przykładami mogą być: przynależność do grupy zawodowej (np. rolnik, górnik), nałogi (np. palenie), istnienie określonych chorób w rodzinie badanego (choroby dziedziczne), wiek itp. Informacje te będą opisywane i przetwarzane przez nasz system w sposób identyczny jak objawy. Zaistniała więc potrzeba zgrupowania wszystkich rodzajów informacji pomocnych przy diagnozowaniu JCH. Zrobimy to przez wprowadzenie osobnej klasy semantycznej, której podklasy będą reprezentowały interesujące nas informacje. Nazwaliśmy tę klasę atrybut medyczny (AM). Rysunek 1 przedstawia przykładową hierarchię klas wywodzących się z klasy Atrybut medyczny.

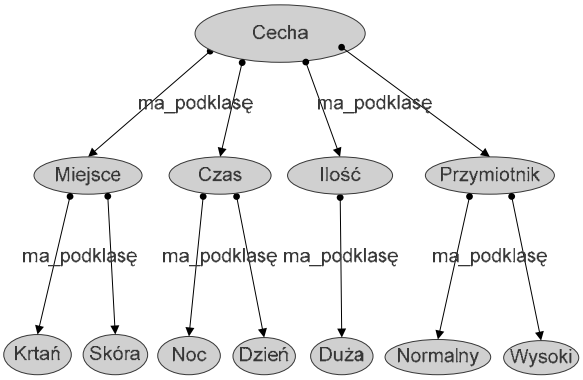


Rys. 1. Przykładowa hierarchia klas wywodzących się z nadrzędnej klasy Atrybut medyczny

Z większością AM można powiązać wiele dodatkowych parametrów, które są cechami charakteryzującymi dany atrybut – cech. Do cech można zaliczyć:

- przymiotnik charakteryzujący atrybut (np. wydzielina: gęsta, zalegająca, obfita itp.)
- miejsce występowania (np. obrzęk: błony śluzowej nosa, powiek, błony śluzowej gardła)
- czas występowania lub nasilania (np. kaszel występujący okresowo lub codziennie, w nocy, między 4–5 rano)
- substancję wywołującą objaw (np. sierść zwierząt, kurz, pyłki roślin, antygeny)
- sytuację, w której występuje objaw (np. stres, zmęczenie, wysiłek fizyczny).

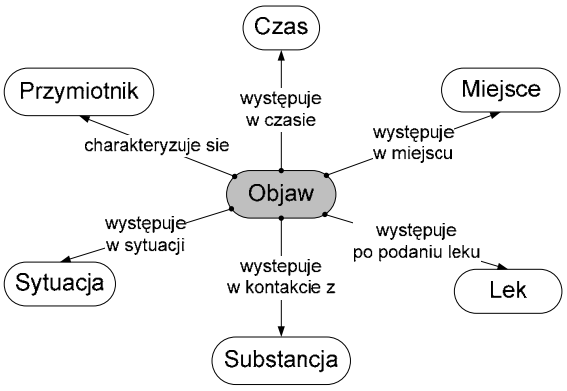
Dla każdej z wymienionych cech można stworzyć osobną klasę semantyczną, która będzie podklasą ogólnej klasy Cecha, rysunek 2.



Rys. 2. Gałąź hierarchii reprezentująca cechy charakteryzujące atrybut medyczny

Struktura hierarchii tworzonej w systemie ma więc dwie główne gałęzie: Atrybuty medyczne i Cechy. Wszystkie pozostałe klasy są podklasami jednej z tych dwóch.

Cechy i AM będą powiązane przez relacje. Ogólną charakterystykę AM reprezentuje fragment ontologii, w którym centralnym węzłem jest konkretny atrybut. Węzeł ten będzie się wiązał z węzłami cech charakteryzujących atrybut. Relacje pomiędzy węzłami będą zależne od typu AM i cechy (rysunek 3). Zarówno AM, jak i ich cechy zostaną zidentyfikowane w tekstach z wykorzystaniem metod przetwarzania języka naturalnego.



Rys. 3. Przykładowe powiązania pomiędzy objawem a cechami go charakteryzującymi

3. Model formalny ontologii i sieci semantycznej

Ontologia

Ontologia jest hierarchicznie i strukturalnie uporządkowanym zbiorem pojęć służących do opisu danej dziedziny:

$$O = \langle C, R, L \rangle . \tag{1}$$

Zbiór C jest zbiorem wszystkich pojęć wykorzystywanych w budowanym modelu.

Element R ontologii O jest zbiorem relacji między pojęciami:

$$R = \{\Re : \Re \subset C \times C\}. \tag{2}$$

Zbiór relacji dzieli się dodatkowo na zbiór relacji strukturalnych oraz relacji hierarchicznych.

Zbiór L jest nazywany leksykonem i jest określony następująco:

$$L = L_C \cup L_R , \tag{3}$$

gdzie:

L_C – zbiór słów stanowiący nazwy dla pojęć, nazywany dalej leksykonem pojęć;

L_R – zbiór słów stanowiący nazwy dla relacji, nazywany dalej leksykonem relacji.

Sieć semantyczna

Z przyjętej definicji ontologii wynika następujące formalne określenie sieci semantycznej:

$$SN^O = \langle I_C^O, I_R^O \rangle , \tag{4}$$

gdzie:

I_C^O – zbiór instancji wszystkich pojęć zdefiniowanych w ontologii O;

I_R^O – zbiór instancji wszystkich relacji pomiędzy pojęciami zdefiniowanymi w ontologii O.

Jeśli $Inst_c$ jest to zbiór instancji pojęcia c, to:

$$I_C^O = \bigcup_{c \in C} Inst_c . \tag{5}$$

Zbiór I_R^O można zapisać jako:

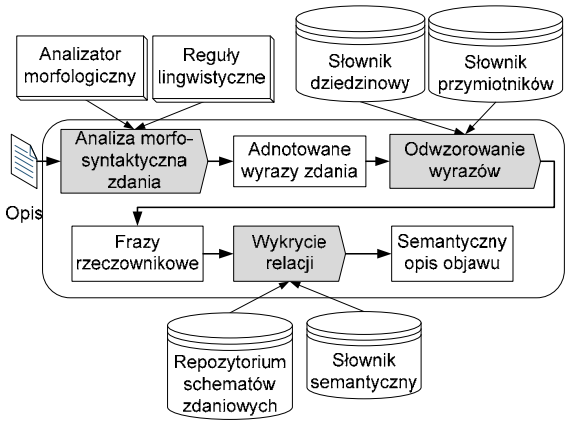
$$I_R^O = \bigcup_{\Re \in R} Inst_{\Re} . \tag{6}$$

Zbiór $Inst_{\Re}$ nazywamy zbiorem instancji relacji $\Re$.

4. Proces tworzenia opisu semantycznego objawów

Opisy wyników diagnostycznych są zapisane w języku naturalnym. Stąd sposób określenia fraz oznaczających objawy i ich cechy jest oparty na metodach przetwarzania języka naturalnego. Na rysunku 4 jest szczegółowo przedstawiony proces budowy opisu semantycznego objawu z pojedynczego zdania. Rezultatem tego procesu jest pojedyncza gałąź sieci

semantycznej, która jest modelem opisującym objaw lub objawy chorobowe. Na podstawie sieci semantycznej będzie budowana ontologia, oznaczymy ją O_{OS}.



Rys. 4. Proces tworzenia opisu semantycznego objawu

Zanim zostanie uruchomiony proces tworzenia opisu semantycznego, konieczne jest stworzenie odpowiednich słowników (słownik dziedzinowy, przymiotników, semantyczny) i zasobów słownikowych (schematy zdaniowe).

Słownik dziedzinowy i słownik przymiotników

Słownik dziedzinowy zawiera rzeczowniki i frazy rzeczownikowe specyficzne dla danego obszaru wiedzy, w omawianym przypadku – medycyny. Słownik jest budowany z wykorzystaniem korpusu tekstów medycznych (opisy wyników TD, opisy JCH, literatura medyczna itd.) i korpusu tekstów o tematyce ogólnej. Słownik przymiotników zawiera przymiotniki charakterystyczne dla rozważanej dziedziny. Przyjmując wcześniej wprowadzone oznaczenia, słownik dziedzinowy i słownik przymiotników stanowią leksykon pojęć L_{C_M} ontologii medycyny O_{M}

Słownik synonimów

Wiele wyrazów zgromadzonych w słowniku dziedzinowym i słowniku przymiotników ma podobne znaczenie, a więc występuje między nimi relacja synonimii. Po zidentyfikowaniu zbiorów takich fraz będą utworzone tzw. synsety, czyli zbiory fraz o tym samym lub zbliżonym znaczeniu, które jednocześnie definiują klasy będące węzłami ontologii. Każdy z synsetów reprezentuje pewne pojęcie. Pojęciem (nazwą synsetu) powinna zostać fraza o największej częstości występowania.

Słownik synonimów jest zbiorem pojęć C_{M} ontologii O_{M}, a słownik dziedzinowy i przymiotników jest leksykonem pojęć L_{C_M}. W hierarchii klas semantycznych przedstawionej na rysunkach 1 i 2 węzły najniższego poziomu zawierają pojęcia ze słownika synonimów.

Słownik semantyczny

Słownik semantyczny grupuje elementy słownika synonimów ($c \in C_M$) poprzez połączenie ich relacją hierarchiczną z odpowiednimi nadklasami. Jak zostało przedstawione w sekcji 2 mamy następujące klasy reprezentujące grupy pojęć: [objaw], [jednostka chorobowa], [czas], [zawód] itd. Te klasy tworzą wyższe poziomy hierarchii klas (rysunek 1 i 2).

Analiza morfosyntaktyczna zdania

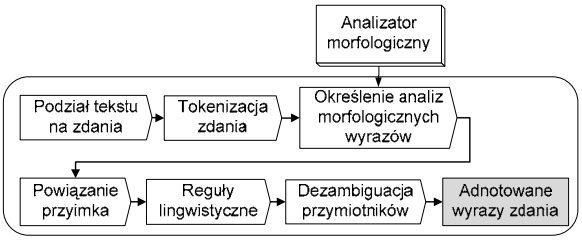
W procesie tworzenia opisu semantycznego brane są pod uwagę opisy wyników tylko z jednej TD lub JCH. Każdy opis składa się ze zbioru zdań. Wykrycie objawu i cech z nim związanych odbywa się w kontekście pojedynczego zdania.

Proces analizy morfosyntaktycznej zdania odbywa się z wykorzystaniem słownika morfologicznego i reguł lingwistycznych. Wynikiem procesu są adnotowane wyrazy zdania – do każdego wyrazu jest przypisany znacznik morfosyntaktyczny.

Schemat procesu analizy składa się z następujących etapów (rysunek 5):

- Tokenizacja

Zdanie jest poddawane tokenizacji, czyli podziałowi na tokeny: słowa, liczby, znaki interpunkcyjne.



Rys. 5. Schemat procesu analizy morfo-syntaktycznej zdania

- Określenie analiz morfologicznych

Dla każdego słowa jest generowany ciąg analiz fleksyjnych przez analizator morfologiczny. Generowane są wszystkie możliwe analizy.

- Powiązanie przyimka

Związek przyimka z rzeczownikiem jest wyrażany za pomocą końcówki rzeczownika charakterystycznej dla przypadku dopuszczalnego w tym połączeniu. Informacja ta umożliwia przeprowadzenie częściowej dezambiguacji niektórych rzeczowników, jak również zaimków, przymiotników i liczebników występujących w parze z przyimkiem.

- Reguła lingwistyczna rzeczownik + rzeczownik (dopełniacz)

W tej części procesu są analizowane wyrazy o klasie gramatycznej rzeczownik. Celem jest ujednoznacznienie kategorii przypadku. Z przeprowadzonych badań wynika, że gdy występują obok siebie w zdaniu „obok siebie" dwa rzeczowniki, np. błona śluzowa oskrzeli, światło oskrzela, świąd skóry itd., to ostatni z nich najczęściej jest w dopełniaczu. Z tej własności skorzystamy przy eliminacji dla tego rzeczownika analiz fleksyjnych zawierających przypadek inny niż dopełniacz.

- Dezambiguacja przymiotników

Zależność między rzeczownikiem i określającym go przymiotnikiem ma wykładnik formalny w postaci końcówek fleksyjnych charakterystycznych dla wspólnego obu wyrazom przypadku i dla wspólnej liczby. Ta własność umożliwia poszukiwanie par – rzeczownik i odpowiadający mu pod względem przypadku i liczby przymiotnik.

W trakcie całego procesu analizy morfosyntaktycznej jest budowana wiedza o podmiocie i orzeczeniu lub podmiotach i orzeczeniach w przypadku zdań złożonych. Opis tego procesu nie mieści się w temacie artykułu.

W trakcie analizy morfosyntaktycznej można wykryć instancje następujących relacji:

charakteryzuje_sie, miejsce_wystąpienia $\in R_{\mathrm{OS}}$,

gdzie R_{OS} – zbiór relacji między pojęciami ontologii O_{OS}.

Instancja relacji występuje między dwoma instancjami pojęć. Nie wszystkie analizowane wyrazy zdania są instancjami pojęć C_{OS} definiowanej ontologii O_{OS}. Część z nich nie występuje w słowniku dziedzinowym, część jest składową fraz rzeczownikowych. Zanim zostaną odrzucone wyrazy nieistotne z punktu widzenia opisu objawów i zanim zostaną wykryte frazy rzeczownikowe, przyjmijmy, że wyrazy zdania są instancjami pojęć ze zbioru C_{x} pewnej ontologii O_{x}.

O wystąpieniu między wyrazami określonej relacji semantycznej możemy wnioskować tylko na podstawie znaczników morfosyntaktycznych przypisanych do wyrazów, ponieważ na tym etapie analizy nie posiadamy informacji semantycznej, jedynie informację morfosyntaktyczną.

Dla relacji *charakteryzuje_się* argumentami są wyrazy z wykrytych fraz rzeczownikowo-przymiotnikowych:

$$\begin{aligned} &Inst_{charakteryzuje\_się} = \{(i,j) \in I_{C_x}^{O_x} \times I_{C_x}^{O_x} : \\ &a(i) \supset \{rzeczownik\}, \\ &a(j) \supset \{przymiotnik\}\}, \end{aligned} \tag{7}$$

gdzie: $a : I_{C_{OS}}^{O_{OS}} \to A$, A – zbiór znaczników morfosyntaktycznych. Relacja między wyrazami pary rzeczownik − rzeczownik w dopełniaczu istnieje, ale na tym etapie nie można określić, jaka to relacja i jaki jest jej kierunek. Dalej instancję relacji $\Re$ będziemy zapisywać:

$$instancja\_relacji(i, j), \tag{8}$$

gdzie: $instancja\_relacji \in Inst_{\Re}$, $i,\ j\ \in I_{C_{OS}}^{O_{OS}}$. Jeśli nie jest znany argument relacji, to zapisujemy „?". Kiedy relacja występuje, ale nie jest znany jej typ, zapisujemy „?relacja". Brak informacji o kierunku relacji, tzn. co jest jej lewym i prawym argumentem, zapisuje się (?lewy_arg, ?prawy_arg). Przynależność do klasy semantycznej zapisujemy jako relację typu hierarchicznego *ma_instancję*:

$$ma\_instancję(klasa\_semantyczna, i), \tag{9}$$

gdzie:

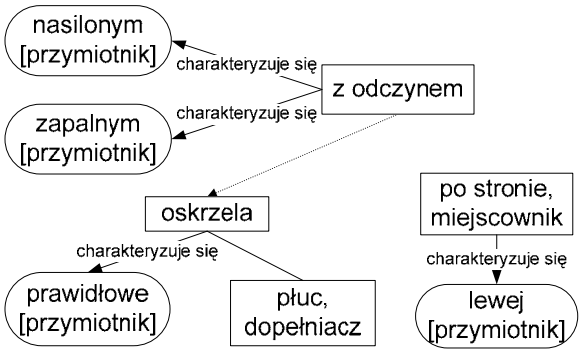
$klasa\_semantyczn \in C_{\mathrm{OS}}, i \in I_{C_{OS}}^{O_{OS}}$.

Przymiotnikom przypisujemy klasę semantyczną [przymiotnik].

Powiązania morfosyntaktyczne między wyrazami wykryte w wyniku analizy można przedstawić za pomocą grafu. W węzłach są umieszczone instancje pojęć, strzałki reprezentują instancje relacji. W węzłach prostokątnych są umieszone rzeczowniki, przymiotniki – w owalnych. Wykryte relacje semantyczne są reprezentowane pełnymi strzałkami. Relacje, które są do wykrycia – strzałkami przerywanymi.

Przykład 1

Dla zdania „Oskrzela płuc są prawidłowe z nasilonym odczynem zapalnym szczególnie po stronie lewej." powstaną poniżej przedstawione grafy.



Rys. 6. Powiązania morfo-syntaktyczne między wyrazami zdania z Przykładu 1

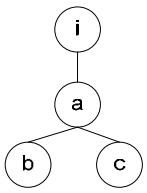
Odwzorowanie wyrazów

Opisany w poprzednim punkcie proces adnotacji opisuje pojedyncze wyrazy. Niektóre wyrazy, gdy występują bezpośrednio obok siebie, tworzą związki mające znaczenie jako całość. Stąd, przechodząc na poziom semantyczny, wymagane jest wykrycie fraz rzeczownikowych. Jak sama nazwa wskazuje, centralnym elementem frazy jest rzeczownik. Proces odwzorowania wyrazów we frazy odbywa się w kilku krokach z wykorzystaniem słownika dziedzinowego. Dla każdego rzeczownika w zdaniu:

$$\forall i \in I_{C_M}^{O_M} : a(i) \supset \{rzeczownik\} \qquad (10)$$

są wykonane kolejne kroki:

1. Utworzenie fraz dla rzeczownika, gdzie fraza zawiera jedno albo więcej słów, na podstawie adnotacji morfologicznej:
   - najpierw dla rzeczownika *i* są wyszukiwane wyrazy, które są z nim w relacji nie tylko bezpośredniej, ale też pośredniej. Niech zbiór Z_i zawiera powiązane z rzeczownikiem *i* wyrazy:
   $$Z_i = i \cup \Big\{ j : j \in I_{C_M}^{O_M} \wedge i \neq j \wedge \exists_{r_k \in B \subset I_{R_{OS}}^{O_{OS}}, k=1,\ldots,n, n \leq |B|} \big(r(i,j) \vee r_1(i,\cdot),\ldots,r_k(\cdot,j)\big)\Big\}, \qquad (11)$$
   gdzie: *B* – zbiór reguł analizowanego zdania.
   - na podstawie zbioru wyrazów Z_i jest generowany zbiór fraz-kandydatów *K*. Wyraz *i* występuje w każdej frazie, stąd są tworzone wszystkie podzbiory zbioru $Z_i/\{i\}$ i do tak powstałych fraz jest dodawany wyraz *i*. Przykładowo, jeśli rzeczownik *i* występuje w relacji z innymi wyrazami, tak jak pokazano na rysunku 6, to dla niego będą utworzone następujące frazy-kandydaci: *i, ia, ib, ic, iab, iac, ibc, iabc*.



Rys. 7. Przykładowe relacje rzeczownika *i* z innymi wyrazami

   - nie wszystkie elementy zbioru *K* mogą być frazami. Trzeba odrzucić takie frazy-kandydatów, w których elementy są nieosiągalne z węzła zawierającego rzeczownik *i*. Dla przykładu z poprzedniego punktu odrzucamy frazy *ib, ic*.
2. Eliminacja ze zbioru *K* fraz rzeczownikowych, które nie występują w słowniku dziedzinowym. Po wykonaniu tego kroku zbiór *K* zawiera tylko instancje pojęć ontologii – O_{OS}:
   $$K \subset I_{C_{OS}}^{O_{OS}} \subseteq I_{C_M}^{O_M}. \qquad (12)$$
3. Ze zbioru *K* jest wybierana najdłuższa dopasowana fraza *k*.
4. Na podstawie frazy *k* są modyfikowane odpowiednie reguły w bazie reguł *B*, tak aby połączyć wyrazy frazy *k* w jeden węzeł.

Wykrycie relacji

Opis objawu składa się z wielu elementów semantycznie powiązanych ze sobą. Jak zostało wcześniej zaznaczone, takimi elementami są nie tylko charakteryzujące go przymiotniki, ale również elementy opisujące miejsce wystąpienia objawu, czas wystąpienia, czynniki wywołujące objaw itd. W procesie dotychczasowej analizy zostały wykryte relacje morfosyntaktyczne istniejące między wyrazami zdania, tworząc grafy powiązań. Na tym etapie analizy wymagane jest wykrycie relacji semantycznych między grafami reprezentującymi elementy opisu objawu. Proces będzie przebiegał z wykorzystaniem schematów zdaniowych i słow-nika semantycznego.

Schematy zdaniowe

Wydawałoby się, że zdań jest tyle, ile kombinacji wyrażeń, tzn. nieograniczona ilość. Okazuje się, że tak nie jest. Istnienie pewnych zasad i uwarunkowań przy tworzeniu kombinacji wyrazowych powoduje, że można je zapisać w postaci schematów [7]. Kluczowym elementem schematu jest czasownik. Swoimi właściwościami semantyczno-gramatycznymi

decyduje w znacznej mierze o strukturze zdania, w którym występuje. W procesie tworzenia wypowiedzenia wybór wyrazów i ich forma zależą od innych wyrażeń, z którymi wiążą się one gramatycznie i semantycznie. To powoduje, że schematy zdaniowe mogą być bardzo przydatnym narzędziem przy rozwiązaniu takich problemów semantycznych, jak określenie relacji semantycznych, brakujących elementów w zdaniu lub znaczenia wyrazów.

Poszczególne części zdania (podmiot, dopełnienie, okoliczniki itp.) mogą zajmować tylko elementy należące do określonych klas leksykalno-semantycznych. Trzeba także zróżnicować łączliwość obligatoryjną i fakultatywną. Łączliwość obligatoryjna dotyczy składników, które muszą wystąpić przy danym czasowniku. Łączliwość fakultatywna dotyczy składników, które mogą, ale nie muszą, być użyte z danym czasownikiem. Charakterystyka semantyczna składników służy tylko do określenia ograniczeń łączliwościowych.

Algorytm uzgadniania zdania i schematu zdaniowego

Algorytm ma za zadanie znaleźć schemat zdaniowy dla danego zdania, wygenerować oczekiwania w odniesieniu do brakujących składowych i określić znaczenie wyrazów wieloznacznych (w sensie semantycznym) w kontekście danego zdania.

Każdy wyraz przechodzi przez analizę morfosyntaktyczną. Szczególną uwagę zwraca się na czasowniki i rzeczowniki (frazy rzeczownikowe). Czasownik pozwoli dotrzeć do właściwych schematów, a rzeczowniki pomogą wybrać jeden z nich.

Etap pierwszy polega na przypisaniu instancji do klas semantycznych. W tym celu zostanie wykorzystany słownik semantyczny. Dla rzeczowników z pary rzeczownik-rzeczownik w dopełniaczu sprawdza się: jeśli rzeczownik w dopełniaczu ma klasę [element_anatomii], to wiążemy oba rzeczowniki relacją *miejsce_wystąpienia* i rzeczownik w dopełniaczu zapisujemy jako drugi argument relacji.

Po określeniu w zdaniu czasownika, wybierane są ze zbioru schematów zdaniowych te schematy, w których dany czasownik występuje. Dokonuje się wstępnego wyboru schematów z listy, wykorzystując przy tym frazy rzeczownikowe.

Drugi etap polega na rozwiązywaniu niejednoznaczności. Jeśli pierwszy etap zakończył się na wybraniu z listy jednego schematu i wszystkie jego składowe występują w zdaniu, to algorytm kończy działanie. Brak w zdaniu pewnych fraz rzeczownikowych wymaganych przez schemat nie oznacza, że wybraliśmy niewłaściwy schemat. Taka sytuacja może wystąpić, kiedy rzeczownik został już wspomniany w poprzednich zdaniach lub będzie o nim mowa w kolejnych zdaniach. W celu ustalenia, jaki to jest rzeczownik, w pierwszym przypadku trzeba przeprowadzić analizę semantyczną dotychczasowych wyników, w drugim – wygenerować oczekiwanie.

Zbiór czasowników, które mogą pojawić się w opisach JCH i TD jest ograniczony. Dla większości z nich liczba schematów zdaniowych ogranicza się do jednego lub dwóch.

Niektóre czasowniki mogą pojawić się w wielu schematach zdaniowych, np. czasownik *być*. Samo określenie przypadku dla rzeczowników może być niewystarczającym kryterium wyboru właściwego schematu. W takim przypadku pomocny będzie słownik semantyczny.

Połączenie informacji morfo-syntaktycznej z semantyczną umożliwi ujednoznacznienie wyboru schematu zdaniowego, a to spowoduje rozwiązanie problemu wieloznaczności semantycznej.

Przykład 2

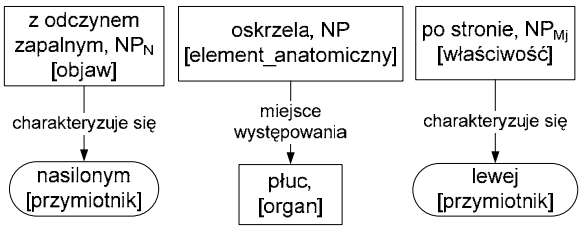
W zdaniu – *Oskrzela płuc są prawidłowe z nasilonym odczynem zapalnym szczególnie po stronie lewej* – występuje czasownik *jest* jako orzeczenie. Poniżej są przedstawione wybrane schematy zdaniowe dla czasownika *jest*:

1. **NP – NP_N+(NP_{Mj})**
   NP→ [element_anatomii]
   NP_N→ [objaw]
   NP_{Mj}→ [właściwość][element_anatomii]
   R: miejsce_wystąpienia(NP_N, NP)
   charakteryzuje_się(NP, NP_{Mj})
2. **NP – NP_N + NP_D**
   NP→ [element_anatomii]
   NP_N → „lokalizacja”
   NP_D→ [objaw]
   R: miejsce wystąpienia(NP_D, NP)
3. **NP – Adj**
   NP→ [objaw]
   NA → [właściwość]
   R: charakteryzuje_się(NP, NA)
4. **NP – NP_N**
   NP → [objaw]
   NP_N → „objaw”

Objaśnienia do schematów:
NP: fraza rzeczownikowa;

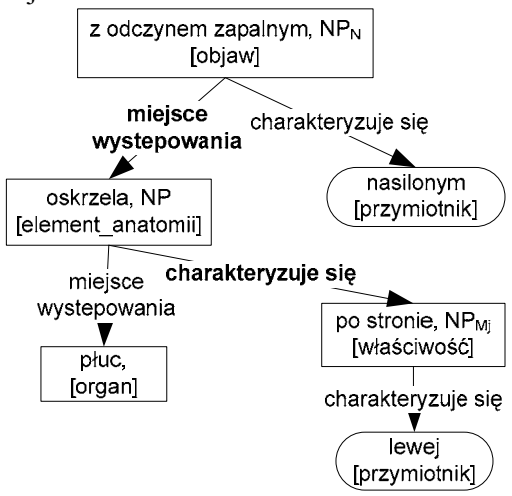
$NP_{D,B,C,N,Mj}$: litery u dołu fraz rzeczownikowych oznaczają ich przypadek (dopełniacz, biernik, celownik, narzędnik, miejscownik);
Adj: przymiotnik;
— pozycja czasownika w schemacie zdaniowym;
+ między składnikami oznacza ich łączenie bez implikacji szyku w aktualnym zdaniu;
( ) fakultatywność składników lub grupy składników (tj. możliwość ich pominięcia);
→ strzałka odsyła do charakterystyki semantycznej;
[ ] dla fraz rzeczownikowych są podane klasy semantyczne;
R: typ relacji.

W korzeniach już wykrytych relacji (w trakcie analizy morfosyntaktycznej) mamy frazy rzeczownikowe: *z odczynem zapalnym, oskrzela, po stronie*.



Rys. 8. Relacje wykryte w trakcie analizy morfo-syntaktycznej

Fraza w narzędniku występuje w schematach 1 i 4, fraza w miejscowniku (może wystąpić, ale nie musi) – tylko w schemacie 1. Podmiot zdania o klasie semantycznej [element_anatomii] pojawia się w schematach 1 i 3. Po sprawdzeniu dopasowania klas semantycznych, wybrany zostaje schemat 1. Uzupełniamy grafy o nowe relacje:



Rys. 9. Graf reprezentujący opis semantyczny zdania z przykładu 2

5. Sieć semantyczna

Produktem procesu wykrywania fraz oznaczających objawy i cech specyfikujących je jest sieć semantyczna stworzona na podstawie tekstu. Każde ze zdań w tekście dostarcza nam pojedynczej gałęzi do struktury sieci. Na podstawie sieci semantycznej będzie zbudowana ontologia – O_{OS}. Z wprowadzonego wcześniej formalnego określenia sieci semantycznej wiadomo, że sieć definiują dwa zbiory. Dla budowanej ontologii O_{OS} będzie to:

$$SN^{O_{OS}} = \left\langle I_{C_{OS}}^{O_{OS}}, I_{R_{OS}}^{O_{OS}} \right\rangle, \qquad (13)$$

gdzie: $I_{C_{OS}}^{O_{OS}}$ – zbiór instancji pojęć ontologii O_{OS} budowanej dla TD lub JCH; zbiór składa się z wyrazów (fraz) występujących we wszystkich opisach analizowanych w ramach danej TD lub JCH; trzeba zaznaczyć, że $I_{C_{OS}}^{O_{OS}} \subseteq I_{C_M}^{O_M}$;

$I_{R_{OS}}^{O_{OS}}$ – zbiór instancji relacji pomiędzy pojęciami C_{OS} ontologii O_{OS}; zbiór zawiera wszystkie relacje wykryte w trakcie analizy zdań TD lub JCH.

6. Tworzenie ontologii

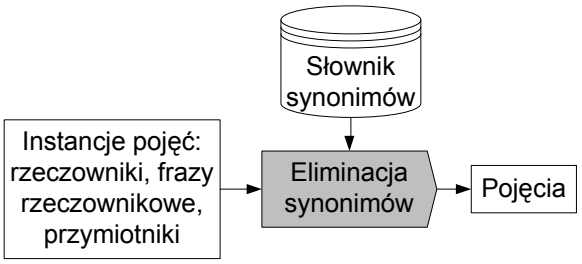
Tworzenie ontologii odbywa się poprzez eliminację synonimów w sieci semantycznej i przekształcenie sieci w ontologię.

Treść o tym samym znaczeniu można zapisać na wiele sposobów, używając przy tym wyrazów lub fraz o zbliżonym znaczeniu. Etap eliminacji synonimów polega na zastąpieniu takich fraz na pojęcia ze słownika synonimów.

$$\forall i \in I \subset I_{C_{OS}}^{O_{OS}}, V_{C_{OS}}^{O_{OS}}(i) \in C_{OS}, \qquad (14)$$

gdzie: $V_{C_{OS}}^{O_{OS}} : I_{C_{OS}}^{O_{OS}} \to C_{OS}$;

$I_{C_{OS}}^{O_{OS}}$ – zbiór instancji wszystkich pojęć C_{OS} zdefiniowanych w ontologii O_{OS};
I – zbiór instancji pojęć w analizowanym zdaniu.



Rys. 10. Schemat eliminacji synonimów

Sieć semantyczną zredukowaną do sieci pojęć przekształca się w ontologię poprzez eliminację wielokrotnych powtórzeń tych samych związków semantycznych.

7. Podsumowanie

Automatyczna budowa semantycznego modelu objawów chorobowych na bazie korpusu słownego jest narzędziem do zbudowania ontologii technologii diagnostycznej lub jednostki chorobowej. Celem konstrukcji ontologii TD jest akwizycja danych diagnostycznych o pacjencie. Ontologia posłuży do budowy interfejsu, poprzez który możliwe będzie wprowadzanie i semantyzacja tych danych. Ontologia JCH jest budowana w celu ujednolicenia opisu jednostek chorobowych. Dzięki zastosowaniu jednolitego formatu do opisu JCH i stanu pacjenta będzie możliwe przeprowadzenie procesu diagnozowania z wykorzystaniem wszelkich możliwych danych. Dalsze prace będą związane z poprawieniem wyników analizy morfosyntaktyczniej, rozbudową systemu o możliwość automatycznej identyfikacji schematów zdaniowych. Do rozważenia jest opracowanie metody wstępnej selekcji zdań ze względu na interesujące nas informacje. Umożliwiłoby to odrzucenie zdań, które nie zawierają informacji istotnej z punktu widzenia opisu objawów.

8. Bibliografia

[1] C. Burgess, „Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling”, in: Gorfein, D.S. (Ed.), *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, APA Press, 2001.

[2] H. Chen, K.J. Lynch „Automatic construction of networks of concepts characterizing document database”, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 22, No. 5, 885–902 (1992).

[3] Z.S. Harris, „Mathematical Structures of Language”, *Interscience Publishers*, New York, 1968.

[4] M.A. Hearst, „Automatic Acquisition of Hyponyms from Large Text Corpora”, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, 1992.

[5] K. Lund, C. Burgess, „Producing high-dimensional semantic spaces from lexical co-occurence”, *Behavior Research Methods, Instrumentation and Computers*, 28, 203–208 (1996).

[6] M. Piasecki, M. Derwojedowa, P. Koczan, A. Przepiórkowski, S. Szpakowicz, M. Zawisławska, „Półautomatyczna konstrukcja Słowosieci” URL www.plwordnet.pl/main. Strona domowa projektu (2007).

[7] K. Polański (red.) *Słownik syntaktyczno-generatywny czasowników polskich,* t. 1–7, Kraków, 1980–1993.

[8] J. Rohmer, ”The Case for Using Semantic Nets as a Convergence Format for Symbolic Information Fusion in NATO”, *RTO-MP-IST-040 Information Systems Technology Panel (IST) Symposium on Military Data and Information Fusion*, Prague, Czech Republic, 2003.

[9] Słowosiec. Witryna WWW projektu. URL http://www.plwordnet.pwr.wroc.pl/main, (2007).

[10] P. Velardi, P. Fabriani, M. Missikoff, „Using text processing techniques to automatically enrich a domain ontology”, in: *Proceedings of the international Conference on Formal Ontology in Information Systems*, FOIS '01, ACM, 270–284, New York, 2001.

Automatic construction of a semantic model of disease symptoms based on text corpus

G. SZOSTEK, M. JASZUK, A. WALCZAK

The research described in article refers the medical data. Descriptions of diagnostic technologies results and descriptions of diseases form the text corpus. The corpus is the basis for building a semantic model of symptoms. A specific symptom can be written in the natural language in many ways, which is a problem for further processing of such information. There is a need to record symptoms in a uniform format. Such format allows for application of the same methods and mathematical tools to support the process of diagnosis. The paper presents method of generating a semantic model based on text corpus. Construction of the model is a part of the research, which aims to make the fusion of data from different sources (heterogeneous data) into homogeneous form.

Keywords: semantic network, ontology, natural language processing.