*Marek Gajęcki**, *Marek Krężołek*****

AUTOMATYCZNA KLASYFIKACJA RZECZOWNIKÓW DO GRUP SEMANTYCZNYCH NA PODSTAWIE KORPUSU TEKSTÓW

W artykule zaprezentowano metodę klasyfikacji rzeczowników do grup semantycznych. Algorytm oparty na wnioskowaniu statystycznym wykorzystuje słownik fleksyjny oraz korpus tekstów. W tekstach analizowane są związki rzeczownika z przymiotnikami i na tej podstawie określana jest przynależność do grupy. Grupy semantyczne zaczerpnięto ze słownika WordNet. Podział rzeczowników na grupy semantyczne jest małym fragmentem budowy słownika semantycznego dla języka polskiego.

Słowa kluczowe: *przetwarzanie języka naturalnego, wnioskowanie statystyczne, słownik semantyczny*

AUTOMATIC CLASSIFICATION OF NOUNS INTO SEMANTIC GROUPS USING A CORPUS OF TEXTS

This article presents a method of classification of nouns into semantic groups based on statistical inference. The algorithm uses the inflectional dictionary of the Polish language and a corpus of texts to analyse adjective-noun relationships. The semantic groups are consistent with the categorization in the WordNet dictionary. The classification of nouns into semantic groups is a small step towards constructing a semantic dictionary for the Polish language.

Keywords: *natural language processing, statistical inference, semantic dictionary*

1. Wprowadzenie

Ważnym elementem każdego systemu przetwarzania jezyka naturalnego jest słownik, a właściwie słowniki maszynowe. W przypadku języka fleksyjnego, np. polskiego, niezbędny jest słownik fleksyjny zawierający powiązania pomiędzy wyrazem i występującymi w tekście formami fleksyjnymi. Przykładem takim jest słownik fleksyjny języka polskiego [1].

Innego typu słownikiem jest tezaurus zawierający semantycznie i hierarchicznie uporządkowane terminy. Przykładem tego typu słownika dla języka angielskiego jest WordNet [2] oraz EuroWordNet [3] dla niektórych języków europejskich. Słownik EuroWordNet tworzony jak słownik WordNet 1.5, zawiera informacje o rzeczownikach i czasownikach, ułożone w podobne struktury i relacje. Główną cechą odróżniającą go

*Katedra Informatyki, Akademia Górniczo-Hutnicza, mag@agh.edu.pl
**Doktorant Wydziału EAIiE, marek.krezolek@firma.interia.pl

od WordNet jest wielojęzyczność. W związku z tym niektóre relacje zmodyfikowano, a z niektórych zrezygnowano. W tej chwili EuroWordNet tworzony jest dla następujących języków: holenderskiego, hiszpańskiego, włoskiego, angielskiego, francuskiego, niemieckiego, czeskiego, estońskiego.

Dla języka polskiego nie istnieje obecnie taki słownik, ponieważ jego ręczne skonstruowanie wymaga wielu lat pracy wysoko wykwalifikowanych lingwistów. Innym sposobem budowy struktury semantycznej jest wykorzystanie wnioskowania statystycznego wykonywanego na korpusie tekstów.

W artykule przedstawiono eksperymentalny algorytm klasyfikacji rzeczowników do grup semantycznych. Przypisanie rzeczowników do odpowiednich grup jest małym fragmentem budowy słownika semantycznego dla języka polskiego analogicznego jak WordNet.

2. Słownik semantyczny WordNet

Twórcy słownika WordNet założyli, że ich słownik ma być bardziej konceptualny od zwykłego słownika alfabetycznego. WordNet dzieli słowa na pięć kategorii: rzeczowniki, czasowniki, przymiotniki, przysłówki i słowa funkcyjne. Cechą WordNet jest próba organizacji słów nie poprzez formy wyrazu, ale poprzez ich znaczenie. Dostęp do wyrazów w WordNet jest podobny jak w słowniku wyrazów bliskoznacznych. Aby odnaleźć wyrazy o podobnym znaczeniu, należy użyć alfabetycznego spisu słów.

Odwzorowanie pomiędzy wyrazami a ich znaczeniami można przedstawić w postaci tablicy leksykalnej, w której kolumny oznaczają poszczególne słowa, a wiersze – poszczególne znaczenia, jak to przedstawia tabela 1.

Tabela 1
Odwzorowanie między formami i znaczeniami

Znaczenia wyrazów	Formy wyrazów				
	F_1	F_2	F_3	...	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
...				...	
M_m					$E_{m,n}$

Jest to odwzorowanie wiele do wielu, czyli każda forma może mieć wiele znaczeń i każde znaczenie może mieć kilka odpowiadających mu form. Każdy kto uczy się danego języka, nabywa wiedzę o tych odwzorowaniach i później dla danego wyrazu w tekście jest w stanie przyporządkować odpowiednie znaczenie, a podczas wypowiedzi dla danego znaczenia jest w stanie wybrać jeden z wyrazów odpowiadający danemu znaczeniu. W WordNet są zdefiniowane zbiory synonimów (synset). Są to zbiory form odpowiadające temu samemu znaczeniu, w naszej tabelce jest to np. $\{F_1, F_2\}$.

Każdemu znaczeniu jest przyporządkowany zbiór synonimów mu odpowiadających. WordNet zorganizowano poprzez relacje semantyczne (relacje pomiędzy znaczeniami). Jeśli mamy zdefiniowaną relację R pomiędzy znaczeniami reprezentowanymi przez zbiory synonimów $\{x, x', \ldots\}$ i $\{y, y', \ldots\}$, to istnieje także relacja R' z $\{y, y', \ldots\}$ do $\{x, x', \ldots\}$. Ponadto jeśli mamy zdefiniowaną relację R pomiędzy dwoma zbiorami synonimów, to R może zostać użyte do oznaczania relacji pomiędzy konkretnymi elementami tych zbiorów.

3. Relacje w słowniku

Definicje pospolitych rzeczowników dają nam o każdym z nich pewne informacje. Są to cechy i określenia, dzięki którym możemy dany rzeczownik odróżnić od innych. Można określić różnego rodzaju relacje, panujące pomiędzy rzeczownikami zawartymi w słowniku [4].

Synonimy to relacje pomiędzy formami wyrazu. Są to bardzo ważne relacje, ponieważ ich określenie jest warunkiem wstępnym do reprezentacji znaczeń w tablicy leksykalnej. Dwa wyrażenia są synonimami w kontekście C, jeśli zastąpienie jednego przez drugie w kontekście C nie zmienia jego prawdziwości. Synonimami dla siebie mogą być tylko formy z tej samej kategorii, czyli dla rzeczownika synonimem może być rzeczownik, dla czasownika – czasownik itp. Zdecydowanie, czy dwa wyrazy są synonimami, czy też nie, jest bardzo trudne, gdyż w jednych wyrażeniach mogą oznaczać to samo, a w innych coś zupełnie innego, dlatego o synonimach mówimy w kontekście danego tekstu lub tematu.

Antonimy są to wyrazy przeciwstawne, np. *biedny–bogaty*. Jest to relacja określona nie pomiędzy znaczeniami (zbiorami synonimów), ale pomiędzy formami. Relacja ta ma duże znaczenie przy organizacji przymiotników i przysłówków w WordNet.

Hiponimy w przeciwieństwie do dwóch poprzednich relacji opisują relację pomiędzy znaczeniami. Znaczenie reprezentowane przez zbiór (synset) $\{x, x', \ldots\}$ jest hiponimem dla $\{y, y', \ldots\}$, jeżeli w danym języku prawdziwe jest stwierdzenie, że x *jest rodzajem* y. Przykładowo *klon* jest hiponimem dla *drzewa*, a *drzewo* jest hiponimem dla *rośliny*. Relacja ta jest niesymetryczna. Hiponim dziedziczy z wyrazu nadrzędnego własności i dodatkowo posiada pewne cechy odróżniające go od innych hiponimów danego wyrazu nadrzędnego. Taka relacja pozwala na budowanie pewnej hierarchii znaczeń w słowniku.

Hiperonimy są relacją przeciwną do hiponimów, jeśli x *jest hiponimem* y, to y *jest hiperonimem* x. Przykładowo *roślina* jest hiperonimem *drzewa*.

Meronimy także opisują relacje pomiędzy znaczeniami. Jest to przykład relacji całość–część. Mówimy, że znaczenie $\{x, x', \ldots\}$ jest meronimem znaczenia $\{y, y', \ldots\}$, jeśli można powiedzieć, że x jest częścią y (przykładowo *gałąź* jest częścią *drzewa*). Jest to także relacja niesymetryczna i może być użyta do budowy pewnej hierarchii wyrazów. Taka sieć połączeń zbudowana w słowniku oraz wiedza, gdzie konkretne słowo znajduje się w danej sieci, daje nam znaczną informację o znaczeniu określonego słowa.

Holonimy są relacją przeciwną do meronimów, jeśli *x jest meronimem y*, to *y jest holonimem x*, np. *drzewo* jest holonimem *gałęzi*.

Podstawową relacją opisującą związki między rzeczownikami są hiponimy, czyli hierarchia, w której hiponim jakiegoś wyrazu dziedziczy jego wszystkie cechy. Rozróżniamy kilka rodzajów cech: atrybuty, części (czyli meronimy), funkcje. Jednak w WordNet zaimplementowane są tylko meronimy. Wśród rzeczowników występuje relacja hiponimu oznaczana @->, czyli a@->b oznacza, że *a* jest hiponimem wyrazu *b*. Mówimy, że *a* jest znaczeniem podrzędnym tej relacji, natomiast *b* znaczenim nadrzędnym (gdyż jest to relacja między znaczeniami). W ten sposób powstaje drzewo zależności. Przykładowo: *klon* @-> *drzewo* @-> *roślina*. Wprowadzono także relację odwrotną do @-> i oznaczono ją ~->. Jeśli takie relacje będziemy mieć opisane w bazie, to łatwo będziemy mogli chodzić po tym drzewie zależności.

Dzięki relacji @-> wprowadzonej do bazy możemy zdobyć dość szeroką informację o danym znaczeniu wyrazu, gdyż mamy do dyspozycji cechy dotyczące danego znaczenia i jego znaczeń nadrzędnych. Przykładowo, jeśli ktoś mówi o wilczurze o imieniu Rex, to my wiemy, że jest to pies, żywe stworzenie, które ma cztery nogi. Podobne informacje możemy uzyskać ze słownika zbudowanego w taki sposób.

W WordNet niektóre wyrazy mają szersze znaczenie niż w potocznym użyciu. Przykładowo, wyraz *kot* potocznie jest używany w znaczeniu *kota domowego*, podczas gdy w słowniku *kot domowy* i *dziki kot* są hiponimami wyrazu *kot*.

Hierarchię rzeczowników opartą na hiponimach można zbudować na kilka sposobów. Na przykład na szczycie tego drzewa postawić semantyczną pustkę, a hiponimami dla niej będą niejasne abstrakcje, takie jak: *jednostka, obiekt, pojęcie*, a następnie struktura ta rozrasta się o kolejne rzeczowniki.

Twórcy WordNet przyjęli inną strategię. Stworzyli kilka takich drzew hierarchii, z których każde zaczyna się od jakiegoś znaczenia. Zdecydowali się oni na 25 grup, których nazwy znajdują się w tabeli 2.

Relacja hiponimów jest podstawowym związkiem z punktu widzenia opisywanego w dalszej części algorytmu, gdyż przydział do jednej z 25 grup jest stwierdzeniem, że dany wyraz jest hiponimem pośrednim lub bezpośrednim wyrazu podstawowego dla danej grupy.

Poszczególne hiponimy są rozróżnialne za pomocą cech. Cechami mogą być atrybuty określane przez przymiotniki, części określane rzeczownikami i czynności określane za pomocą czasowników. Cechy te są dziedziczone od wyrazów nadrzędnych. Do każdego znaczenia można dołożyć kolejne cechy.

Często przy określaniu cech pojawia się wiele problemów. Dla atrybutów takim problemem jest to, że to samo określenie nie oznacza tego samego, na przykład mały oznacza co innego w odniesieniu do ptaka, a co innego w odniesieniu do konia. Twórcy WordNet proponują, aby wtedy ta relacja nie była dwustronna, tzn. jeśli chcemy poznać cechy *kucyka* (jako *konia*), to wśród jego cech powinien pojawić się atrybut *mały* (w znaczeniu *mały* jako *koń*), ale gdy będziemy szukać rzeczy, które są *małe*, to na tej liście nie powinien pojawić się *kucyk*.

Tabela 2
Grupy semantyczne w słowniku WordNet

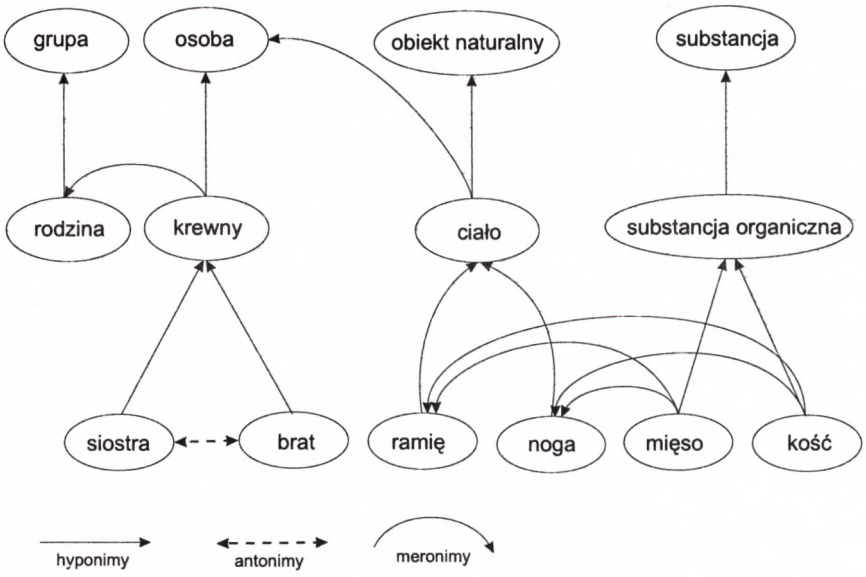
Nr	Nazwa angielska	Nazwa polska
01	act, action, activity	akt, akcja, czynność, działanie
02	animal, fauna	zwierzę, fauna
03	artefact	przedmiot, artefakt
04	attribute, property	atrybut, właściwość, cecha
05	body, corpus	ciało, tułów
06	cognition, knowledge	wiedza
07	communication	komunikacja
08	event, happening	zdarzenie
09	feeling, emotion	uczucie, emocje
10	food	jedzenie
11	group, collection	grupa, kolekcja
12	location, place	lokacja, miejsce
13	motive	przyczyna
14	natural object	obiekt naturalny
15	natural phenomenon	zjawisko naturalne
16	person, human being	osoba, człowiek
17	plant, flora	roślina, flora
18	possession	posiadanie, własność
19	process	proces
20	quantity, amount	liczba, ilość
21	relation	relacja, związek
22	shape	kształt
23	state, condition	stan, położenie
24	substance	substancja
25	time	czas

Problemem przy określaniu części (meronimów) jest to, że ludzie często traktują tę relację jako przechodnią, podczas gdy nie zawsze tak jest, bo *gałąź jest częścią drzewa, drzewo jest częścią lasu*, ale nie można powiedzieć, że *gałąź jest częścią lasu.* Można wyróżnić wiele rodzajów meronimów. W WordNet zostały wprowadzone trzy rodzaje.

Mówimy, że x jest meronimem y:

1) jeśli x jest częścią y;
2) x jest członkiem y;
3) x jest czymś, z czego y jest zrobione.

Przykładowe relacje pomiędzy rzeczownikami przedstawia rysunek 1.



Rys. 1. Relacje pomiędzy rzeczownikami

4. Związek przymiotnika z rzeczownikiem

Rzeczownik posiada dwie kategorie fleksyjne: odmienia się przez przypadki i liczby. Dodatkowo każdy rzeczownik ma przypisany rodzaj gramatyczny. Rodzaj rzeczownika przejęto z pracy Wróbla [5], w której rodzaj przyjmuje jedną z siedmiu wartości: męski osobowy, męski żywotny, męski nieżywotny, żeński, nijaki, *plurale tantum* osobowe i *plurale tantum* nieosobowe. Rodzaj gramatyczny rzeczowników decyduje o końcówkach przy odmianie przez przypadki i liczby, ale przede wszystkim pozwala rzeczownikowi wybrać odpowiednią formę leksemu podrzędnego, np. przymiotnika.

Przymiotnik w stosunku do rzeczownika pełni rolę podrzędną. W zdaniu rzeczownik narzuca przymiotnikowi swoje wartości trzech kategorii gramatycznych: przypadka, liczby i rodzaju, innymi słowy następuje ich uzgodnienie (związek zgody).

Zgodność kategorii przypadku, liczby i rodzaju, pozwala na automatyczne wyszukiwanie w tekście par rzeczowników i odnoszących się do nich przymiotników. W najprostszym przypadku każde dwa sąsiednie: przymiotnik i rzeczownik (albo rzeczownik i przymiotnik), które zgadzają się co do przypadka, liczby i rodzaju stanowią

taką parę. Oczywiście należy uwzględnić też przypadki, gdy rzeczownik jest określany w tekście przez kilka przymiotników oraz gdy jeden przymiotnik określa kilka rzeczowników.

Po **kamiennych schodach** płynął **cały potok**, przemywał **kamienną podłogę** i wypływał niżej, od strony stawu. Można się jedynie przyzwyczaić do **ponurego, wiecznego szumu** wody, do **niespokojnych snów** [1].

Trudniej jest w sposób automatyczny odszukać pary przymiotnika i rzeczownika, jeżeli pomiędzy nimi znajdują się inne wyrazy. Wymaga to analizy składniowej zdania.

Tylko **sen** jest **prawdziwy**. Wydało jej się, że znowu słyszy ten **ciepły**, pełen miłości **głos** w lewym uchu. Odwróciła się do niego i zobaczyła **wpatrzone** w siebie ciekawie **oczy**.

Ponieważ tego typu pary występują rzadko, przy analizie statystycznej można je pominąć.

5. Macierz powiązań rzeczownik–przymiotnik

Dla zbudowania macierzy powiązań pomiędzy rzeczownikami i przymiotnikami niezbędne są dwa podstawowe narzędzia: słownik fleksyjny i korpus tekstów.

Słownik fleksyjny zawiera wyrazy języka polskiego podzielone na klasy fleksyjne wraz z pełną odmianą każdego wyrazu. W eksperymentach wykorzystano słownik języka polskiego [1] wraz z mechanizmem dostępu w postaci serwera leksykalnego [6]. Baza serwera zawiera blisko 118 000 wyrazów pospolitych, w tym 53 735 rzeczowników oraz 38 052 przymiotników. Zadaniem słownika fleksyjnego jest określanie klasy fleksyjnej analizowanych wyrazów oraz określanie kategorii gramatycznych rozpoznanych rzeczowników i przymiotników.

Korpus tekstów, jaki posłużył do badań, zawiera łącznie ponad 25 mln słów. Połowę tekstów stanowią notatki prasowe PAP [7], reszta to teksty ze zbiorów Polskiej Biblioteki Internetowej [8], artykuły dostępne w Internecie, teksty prac magisterskich.

Przeglądając korpus tekstów, zliczamy znalezione pary rzeczowników i przymiotników, a wyniki umieszczamy w macierzy powiązań przedstawionej w tabeli 3.

Każdy wiersz odpowiada przymiotnikowi, natomiast kolumna rzeczownikowi. Elementami macierzy są liczby wystąpień par przymiotnika z rzeczownikiem. Na podstawie analizy zbioru tekstów o rozmiarze 200 MB wyodrębniono prawie 3,2 mln par rzeczownik–przymiotnik. Pary te utworzyły macierz związków obejmującą 24 809 rzeczowników i 24 146 przymiotników.

Zliczone pary pozwalają w łatwy sposób generować listy przymiotników opisujących w tekstach dany rzeczownik. Lista taka odpowiada jednej kolumnie z tabeli 3, w której pod uwagę bierzemy tylko wiersze o wartościach większych od 0.

[1] Przykłady pochodzą z książki Olgi Tokarczuk „Dom dzienny, dom nocny" wydanej przez Wydawnictwo RUTA.

Przykładowa lista dla rzeczownika *student* wygląda następująco:

```
# ACT wersja 1.3
# Unicon Version 10.0 beta July 8, 2000
# ILP wersja 0.9.2 Apr 22 2003
# Wyraz: student
1092577     2     12  16.66  świętujący
1090337     2     17  11.76  sytuowany
1097512     7     71   9.85  upieczony
1112020    14    168   8.33  zaoczny
1067386    16    196   8.16  poszukujący
1011437     2     30   6.66  demonstrujący
1086334     2     42   4.76  słuchający
1006206     2     54   3.70  brodaty
1043716     3     81   3.70  najzdolniejszy
1111760     7    216   3.24  zamordowany
1095844     5    165   3.03  uczący
1015044     2     70   2.85  dwudziestoletni
1013078     2     80   2.50  domagający
1085389     2     81   2.46  skomputeryzowany
1008585     8    326   2.45  chudy
1117174     2     91   2.19  zmotoryzowany
1098808     5    228   2.19  utalentowany
1105568     2    102   1.96  wyłoniony
1099346     2    108   1.85  uzdolniony
1031337     4    244   1.63  kończący
1070447     5    318   1.57  protestujący
1038923     7    449   1.55  mieszkający
1067744     3    198   1.51  potrzebujący
1015457     3    203   1.47  dyplomowy
1073425     2    136   1.47  przestały
1061009     2    142   1.40  pobity
1059981     7    540   1.29  pijany
1043523     2    167   1.19  najuboższy
1111706     2    169   1.18  zamieszkały
1009861     2    171   1.16  czarnoskóry
1035041    77   7171   1.07  letni
1042267     2    189   1.05  najbiedniejszy
1088411     2    194   1.03  stacjonarny
1118037     2    201   0.99  zrzeszony
1007723     2    222   0.90  chcący
1095568     2    226   0.88  ubiegający
1116080     5    607   0.82  zgromadzony
1067347     3    475   0.63  poszkodowany
1115191     3    489   0.61  zdolny
1018709     2    357   0.56  fiński
1096582    11   2002   0.54  ukraiński
```

Kolejne kolumny zawierają: numer identyfikujący przymiotnik, liczbę wystąpień przymiotnika z określanym rzeczownikiem, łączną liczbę wystąpień przymiotnika, iloraz dwóch ostatnich wyrażony w procentach oraz sam przymiotnik.

Tabela 3
Macierz związków rzeczowników z przymiotnikami

	abakus	abażur	abdykacja	abecadło	...	żywopłot	żywność	żywot	żyzność
abdominalny									
abdykacyjny									
abdykujący									
abecadłowy									
⋮									
żywotny									
żywszy									
żywy									
żyzny									

Działanie programu generującego listy można zobaczyć na stronie WWW pod adresem http://winnie.ics.agh.edu.pl/badania.

Podstawą algorytmu klasyfikacji rzeczowników jest obserwacja, że rzeczowniki z różnych grup semantycznych są określane różnymi przymiotnikami, na przykład rzeczowniki oznaczające zjawiska naturalne są określane innymi przymiotnikami niż np. przedmioty.

Macierz powiązań z tabeli 3 posłuży do zbudowania pewnej bazy wiedzy. Wiedza ta będzie polegać na określeniu, w jakim stopniu dany przymiotnik jest związany z rzeczownikami danej grupy semantycznej. Na początku należy przypisać do każdej z grup po kilkanaście rzeczowników, np. do grupy **uczucie** przyporządkujemy rzeczowniki: *chęć, emocja, lęk, miłość, nadzieja, nienawiść, pewność, piękno, podziw, przyjaźń, radość, rezygnacja, strach, zło, złość.* Następnie, dla każdego przymiotnika, który wystąpił z rzeczownikami wstępnie przypisanymi do grup, obliczamy współczynnik według wzoru

$$N(G_k, C_j) = \frac{\sum_{i=1}^{n(G_k)} \frac{N(A_i^k, C_j)}{N(A_i^k)}}{N(C_j) \cdot n(G_k)} \tag{1}$$

gdzie:

$N(G_k, C_j)$ – współczynnik występujący dla grupy G_k i przymiotnika C_j,

G_k – k-ta grupa rzeczowników,

C_j – j-ty przymiotnik,

$A_1^k \ldots A_n^k$ – rzeczowniki należące do k-tej grupy,

$N(A_i^k)$ – liczność rzeczownika A_i^k,

$N(C_j)$ – całkowita liczność przymiotnika,

$N(A_i^k, C_j)$ – liczba wspólnych wystąpień rzeczownika A_i^k z przymiotnikiem C_j,

$n(G_k)$ – liczba rzeczowników należących do grupy G_k.

Wartość taką obliczamy dla wszystkich grup, otrzymując dla każdego przymiotnika wektor 25 współczynników.

Wzór (1) uwzględnia fakt, że rzeczowniki z danej grupy mają różną częstość występowania. Wynik dzielimy przez liczbę rzeczowników w grupie, tak aby wartość współczynnika dla grupy nie była uzależniona od liczby rzeczowników w grupie. Dzielenie przez liczność przymiotnika wykonujemy po to, aby współczynnik nie zależał od liczności występowania danego przymiotnika.

Ponieważ uzyskana wartość jest stosunkowo mała, normalizujemy ją, tzn. dzielimy przez sumę wszystkich współczynników dla danego przymiotnika, według wzoru

$$NN(G_k, C_j) = \frac{N(G_k, C_j)}{\sum_{i=1}^{L} N(G_k, C_j)} \tag{2}$$

gdzie:

$NN(G_k, C_j)$ – współczynnik dla grupy G_k i przymiotnika C_j po normalizacji,

L – liczba grup.

Otrzymane wartości współczynników NN dla kilku przymiotników mają następujące wartości:

```
1000128 absurdalny
  0.0386  0.0000  0.0027  0.0095  0.0000  0.1911  0.0691  0.0442  0.0310
  0.0000  0.0000  0.0000  0.3045  0.0000  0.2239  0.0000  0.0000  0.0000
  0.0000  0.0812  0.0000  0.0000  0.0000  0.0036  0.0000
1000193 adaptacyjny
  0.2794  0.0000  0.0000  0.0000  0.0000  0.2174  0.0000  0.0000  0.0000
  0.0000  0.0000  0.2909  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
  0.0316  0.0000  0.0000  0.0000  0.1804  0.0000  0.0000
1000248 administracyjny
  0.0356  0.0129  0.0143  0.0070  0.0914  0.0493  0.0024  0.0105  0.0009
  0.0000  0.0542  0.1236  0.2461  0.0000  0.0000  0.0400  0.0000  0.0215
  0.1762  0.0058  0.0105  0.0816  0.0090  0.0000  0.0062
1000340 aerodynamiczny
  0.0216  0.0000  0.0000  0.0000  0.0000  0.0240  0.0000  0.0000  0.3452
  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
  0.0000  0.0000  0.4130  0.1959  0.0000  0.0000  0.0000
```

Obliczone wektory współczynników posłużą do określania przynależności rzeczowników do grup semantycznych.

6. Algorytm

Aby wskazać przynależność danego rzeczownika do grupy semantycznej, należy przeanalizować przymiotniki, które określały go w tekście. Dla wszystkich takich przymiotników należy zsumować ich wektory współczynników *NN*, w wyniku czego otrzymujemy wektor 25 współczynników. Otrzymane współczynniki należy znormalizować analogicznie jak wartości współczynnika *NN* i posortować od wartości największej do najmniejszej. Współczynnik o największej wartości wyznacza grupę semantyczną, do której należy rzeczownik.

Po zastosowaniu opisanych wcześniej wzorów otrzymano wyniki mające postać przedstawioną w tabeli 4.

Tabela 4

Przydział rzeczowników do grup semantycznych. Objaśnienia w tekście

zupa	273	0,534	0,042	12,79	jedzenie	roślina, flora
erozja	48	0,523	0,041	12,86	zjawisko naturalne	przyczyna
tlen	62	0,574	0,047	12,32	substancja	jedzenie
ziemniak	149	0,584	0,050	11,59	jedzenie	substancja
azot	64	0,537	0,046	11,72	substancja	relacja, związak
obręcz	90	0,545	0,048	11,26	kształt	roślina, flora
milimetr	20	0,608	0,061	10,03	liczba, ilość	czas
tęcza	41	0,468	0,043	10,95	zjawisko naturalne	obiekt naturalny
krzywa	167	0,439	0,039	11,23	kształt	wiedza
flora	62	0,557	0,056	9,92	roślina, flora	zwierzę, fauna
ciasto	287	0,525	0,053	10,00	jedzenie	substancja
hipoteka	81	0,426	0,039	10,94	posiadanie, własność	atrybut, cecha
węglowodan	12	0,580	0,062	9,30	substancja	jedzenie
ucho	13	0,552	0,060	9,13	ciało, tułów	obiekt naturalny
walec	39	0,544	0,061	8,97	kształt	liczba, ilość
bylina	81	0,509	0,057	8,86	roślina, flora	substancja
koń	492	0,436	0,045	9,58	zwierzę, fauna	osoba, człowiek
pokój	86	0,445	0,048	9,32	lokacja, miejsce	substancja
ręka	281	0,447	0,049	9,09	ciało, tułów	roślina, flora
krzak	137	0,439	0,049	8,90	roślina, flora	ciało, tułów
dąb	165	0,440	0,052	8,52	roślina, flora	osoba, człowiek

Kolumny kolejno zawierają: rzeczownik, liczbę wystąpień wyrazu w tekstach, współczynnik rzeczownika dla najlepszej kategorii, współczynnik dla drugiej w kolejności kategorii, stosunek współczynnika najlepszej kategorii do drugiej w kolejności, najlepiej pasującą grupę rzeczowników oraz drugą w kolejności grupę.

7. Otrzymane rezultaty

Analiza otrzymanych rezultatów wykazała, że na poprawność wyboru kategorii mają wpływ trzy wielkości:

1) liczba wystąpień rzeczownika w tekstach (l),

2) współczynnik dla najlepszej kategorii (w),

3) stosunek najlepszego współczynnika do drugiego w kolejności (s).

Przeprowadzone testy wykazały, że im większa liczba wystąpień danego rzeczownika, tym pozostałe dwa współczynniki mogą mieć mniejszą wartość. Na podstawie testów określono następujące warunki, które powinny być spełnione, aby można było uznać, że rzeczownik został prawidłowo zakwalifikowany do określonej grupy:

$$\begin{array}{lll} l > 100 & w > 0{,}1 & s > 1{,}5 \\ l > 25 & w > 0{,}2 & s > 2 \\ l > 10 & w > 0{,}5 & s > 3 \end{array} \tag{3}$$

Nie wszystkie rzeczowniki są tak samo łatwo klasyfikowane. Niektórym wystarczy 15 wystąpień, innym potrzeba 150. Są też takie, dla których nigdy nie będziemy w stanie określić dokładnej przynależności do grupy, gdyż współczynniki przy nich występujące są zbyt małe. Zależy to od tego, czy przymiotniki opisujące dany rzeczownik są charakterystyczne dla danej grupy, oraz od tego, czy dany rzeczownik nie opisuje kilku pojęć, które mogą zostać zaklasyfikowane do różnych grup.

Ostatecznie spośród ponad 24 000 analizowanych rzeczowników około 1800 spełnia warunek (3). Wśród tych rzeczowników, po weryfikacji ręcznej, stwierdzono obecność mniej niż 5% rzeczowników sklasyfikowanych błędnie.

8. Wnioski

Ograniczeniem proponowanej metody jest korpus tekstów. Jak wynika z badań [9], liniowy wzrost liczebności rzeczowników do analizy można uzyskać przy wykładniczym wzroście rozmiaru korpusu. Ten prosty wniosek jest konsekwencją prawa Zipfa [10] określającego zależność częstości wystąpień wyrazu od jego rangi. Należy więc pamiętać o tym, aby posiadać odpowiednio duży i reprezentatywny korpus tekstów.

Przedstawiony algorytm może być zastosowany w celu klasyfikacji do innych, np. mniej szczegółowych grup. Na przykład w słowniku syntaktycznym [11] rzeczowniki mają przypisane cechy semantyczne: abstrakcyjność albo konkretność, żywotność albo nieżywotność, osobowość albo nieosobowość, zbiorowość, żywioł, roślina, informacja, instytucja, narzędzie, płyn, maszyna, materiał, część. W celu dokonania takiej klasyfikacji należy stworzyć tylko początkowe zestawy rzeczowników.

Literatura

[1] Lubaszewski W., Wróbel H., Gajęcki M., Moskal B., Orzechowska A., Pietras P., Pisarek P., Rokicka T.: *Słownik fleksyjny języka polskiego.* Kraków, Wydawnictwo Prawnicze LexisNexis 2001

[2] Fellbaum Ch. (red.): *WordNet. An Electronic Lexical Database.* MIT Press 1998

[3] http://www.illc.uva.nl/EuroWordNet

[4] Miller G. A.: *Nouns in WordNet: A Lexical Inheritance System.* [in:] International Journal of Lexicography, 3(4) 1990

[5] Wróbel H.: *Gramatyka języka polskiego.* Kraków, OD NOWA 2001, 91–92.

[6] Gajęcki M.: *Serwer leksykalny języka polskiego.* Computer Science, vol. 3, 2001, 131–150

[7] http://dziennik.pap.com.pl.

[8] http://www.pbi.edu.pl

[9] Gajęcki M.: *Automatyczne generowanie słownika asocjacyjnego na podstawie korpusu tekstów.* V Krajowa Konferencja Naukowa „Inżynieria Wiedzy i Systemy Ekspertowe", Wrocław 2003

[10] Zipf G. K.: *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Houghton Mifflin 1935

[11] Polański K.: *Słownik syntaktyczno-generatywny czasowników polskich.* Wrocław-Warszawa-Kraków, Ossolineum 1980–1992 (tom I–V)