AKADEMIA GÒRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI,
INFORMATYKI I ELEKTRONIKI

KATEDRA INFORMATYKI

ROZSTRZYGANIE WIELOZNACZNOŚCI WYRAZU ZA POMOCĄ RELACYJNEGO OPISU ZNACZENIA

EWA THLON I TOMASZ PĘDZIMĄŻ

INFORMATYKA

PRACA MAGISTERSKA

PRACA NAPISANA POD KIERUNKIEM:

PROF. DR HAB. WIESŁAWA LUBASZEWSKIEGO

Kraków, 2008

SPIS TREŚCI:

SPIS RYSUNKÒW:

SPIS TABEL:

ROZDZIAŁ I

WSTĘP

Natura języka sprawia, że nie istnieje jeden, uniwersalny sposòb wyrażania myśli. Wieloznaczność pojawia się na każdym z poziomòw rozumienia – od wyrazòw i ich związkòw, poprzez zdania, a na całych wypowiedziach kończąc. Z mnogością znaczeń najczęściej mamy do czynienia w słownictwie – poszczegòlne wyrazy, w zależności od intencji mòwiącego, mogą być rozumiane na ròżne sposoby.

Warto zaznaczyć, że zjawisko to i wszystkie zagadnienia z nim związane, rzadko bywają przeszkodą w komunikacji międzyludzkiej. Wieloznaczność stanowi tak integralny element języka, że właściwa interpretacja sposobu rozumienia słowa dokonywana jest przez ludzki mòzg automatycznie i nieświadomie – dla języka angielskiego na przykład, około 40% słòw może być rozumiane na więcej niż 1 sposòb. [1] Wieloznaczność staje się jednak problemem dla aplikacji komputerowych automatyzujących procesy przetwarzania tekstu. Sposobem na jego rozwiązanie jest konstrukcja słownikòw semantycznych i właściwe ich wykorzystanie w celu wskazywania poprawnych w danym kontekście znaczeń przetwarzanych wyrazòw.

Wyròżnić można kilka ogòlnych etapòw pracy hipotetycznego systemu przetwarzającego język naturalny:

1. Rozpoznawanie mowy (opcjonalnie) – etap zamiany dźwięku na zapis słowny,
2. Tokenizacja i segmentacja – wydzielenie z tekstu podstawowych, nierozdzielnych jednostek ,
3. Analiza morfosyntaktyczna – formalny opis tokenòw ze względu na ich właściwości składniowe, rozpoznanie znanych form wyrazowych,
4. Ujednoznacznianie sensu słòw – rozstrzygnięcie niejednoznaczności w przypisaniu znaczenia do poszczegòlnych wyrazòw,
5. Analiza składniowa – przypisanie do poszczegòlnych wyrażeń językowych struktur składniowych ,
6. Analiza semantyczna – przypisanie do wyrażeń językowych wyrażeń języka formalnego,
7. Analiza dyskursu – analiza powiązań znaczeniowych i pełnego znaczenia wypowiedzi. [2]

[1] Y.Ravin, C.Leacock *Polysemy,Theoretical and Computational Approaches*, Oxford University Press 2000,s.1.

[2] D.Jurafsky, J.H.Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall 2000, s. 3-4.

Opracowany algorytm realizuje czwarty z wymienionych powyżej punktòw analizy tekstu, służącym za wejście do bardziej zaawansowanych algorytmòw przetwarzania wypowiedzi języka naturalnego. Tak uzyskana wiedza o analizowanym tekście może zostać wykorzystana w systemach z wielu dziedzin, jak choćby:

- Automatyczne tłumaczenie tekstu – obecnie duża liczba rozwiązań problemu opiera się na podejściu statystycznym. Duży korpus tekstòw pozwala na zmniejszenie problemu wieloznaczności wypowiedzi poprzez statystyczne wybranie odpowiedniego terminu. Zastosowanie naszego algorytmu, poprzez wybranie konkretnego terminu spośròd wielu terminòw, pozwoli na poprawę jakości generowanego tłumaczenia. Istnieją potwierdzone przypadki polepszenia wynikòw działania tłumaczenia opartego na statystyce przez dodanie modułu ujednoznaczniającego tekst. [3]
- Wyszukiwanie pozycji w katalogach zasobòw w poszukiwaniu zadanego terminu (przykładem może być elektroniczna biblioteka książek) – zastosowanie algorytmu pozwala na doprecyzowanie zapytania poprzez wskazanie bardziej zawężonego terminu, bądź powiększenie zapytania o słowa powiązane z wprowadzonymi. Pewną analogią do powyższego algorytmu jest podpowiadanie zapytania użyte w silniku wyszukiwarki Google. Jednak w tamtym przypadku metoda opiera się na podejściu statystycznym, stąd analogia jest jedynie na poziomie użytkowym. [4]
- Kolejną dziedziną, w ktòrej użycie naszego algorytmu może poprawić jakość usługi jest pozycjonowanie reklam dla konkretnego tekstu – dla istniejących rozwiązań, jako przykład można podać Google AdWords, istotne jest umieszczenie reklamy przy treści jak najbardziej się do niej odnoszącej. Rozwiązanie firmy Google opiera się na pozycjonowaniu reklam na stronach, ktòre zawierają zestaw słòw kluczowych z nią powiązanych. [5]

Wymienione wyżej przykłady stanowią najbardziej oczywiste zastosowania opisywanego algorytmu. Należy jednak zaznaczyć, że praktycznie każde zagadnienie ktòre opiera się na przetwarzaniu tekstu, podczas ktòrego występuje zjawisko wieloznaczności może zostać przy jego użyciu ulepszone. Dalszymi, możliwymi zastosowaniami może być automatyczne tłumaczenie tekstòw, wyszukiwarki dla pojęć (a nie słòw, jak jest obecnie), uczenie maszynowe i inne.

Niniejsza praca jest opisem pròby zmierzenia się z problemem występowania wieloznaczności w procesie przetwarzania języka naturalnego i jej celem było opracowanie i implementacja algorytmu pozwalającego na wybòr najbardziej prawdopodobnego ze znaczeń wyrazòw wieloznacznych wypowiedzi, w zależności od kontekstu ich występowania. [6]

---

[3] Y.S.Chang, H.T.Ng, D.Chiang, *Word Sense Disambiguation Improves Statistical Machine Translation*, http://www.isi.edu/~chiang/papers/chan_wsd_in_mt.pdf, s.8.

[4] *Frequency Asked Questions for Google Suggest:* http://labs.google.com/suggestfaq.html.

[5] *Pozycjonowanie w AdWords:* http://adwords.google.pl/support/bin/answer.py?answer=65133&topic=9352.

[6] Kontekst można zdefiniować jako zbiór słów współwystępujących w rozpatrywanej wypowiedzi i dostarczających informacji o jej znaczeniu, do których to ograniczamy przestrzeń poszukiwań.

Istnieje wiele technik zmierzenia się z tym zagadnieniem, takich jak na przykład:

- metody probabilistyczne,
- sieci neuronowe,
- metody regułowe,
- drzewa decyzyjne,
- gramatyki formalne i inne.

W trakcie realizowania pracy wybrano podejście relacyjne, w ktòrym to pojęcia i związki pomiędzy nimi zachodzące uporządkowane są w formie relacyjnej sieci semantycznej.

Podczas implementacji wykorzystano istniejące komponenty stworzone w ramach prac Grupy Lingwistyki Komputerowej Katedry Informatyki Akademii Gòrniczo-Hutniczej oraz Katedry Lingwistyki Komputerowej Uniwersytetu Jagiellońskiego: Słownik Fleksyjny Języka Polskiego w postaci biblioteki CLP[7] oraz Słownik Semantyczny Języka Polskiego[8] ktòry to dostarczył wiedzy na temat wspomnianych wyżej bytòw i ich zależności. Algorytm został zaimplementowany w postaci biblioteki języka C++, działającej zaròwno na platformie Windows jak i Unix/Linux.

Poniżej omòwiona została zawartość poszczegòlnych części pracy.

W rozdziale drugim przedstawiony został kròtki przegląd zagadnień związanych z pojęciem wieloznaczności – potrzeba ujednoznaczniania treści na poziomie semantycznym. Omòwione ponadto zostały ròżne podejścia dotyczące problemu postrzegania znaczeń przez ludzi i wywodzącą się z jednego z nich idea słownika semantycznego.

Rozdział trzeci opisuje sposòb, w jaki słownik semantyczny może być wykorzystany w problemie rozstrzygania znaczeń wyrazu w zależności od kontekstu ich występowania, oraz szczegòły opracowanego algorytmu ujednoznaczniającego.

W rozdziale czwartym zaprezentowano szczegòłowy sposòb implementacji omòwionego wcześniej algorytmu. Przedstawiony został podział na moduły, przybliżone zostały kolejne kroki algorytmu, opis najważniejszych struktur i cel ich zastosowania.

Rozdział piąty zawiera opis zbioròw testowych wraz z otrzymanymi dla nich wynikami. Dla każdego rodzaju testòw dokonano analizy zwròconych rezultatòw.

Szòsty rozdział zawiera podsumowanie i omòwienie wynikòw pracy, a także przegląd możliwości dalszego rozszerzenia zaròwno algorytmu, jak i słownika semantycznego.

[7] Zob. http://winnie.ics.agh.edu.pl/proj_uk/fleksbaz/.

[8] Zob. http://winnie.ics.agh.edu.pl/proj_re/slse/index.html.

ROZDZIAŁ II

SŁOWNIK SEMANTYCZNY

2.1. SŁOWNIK SEMANTYCZNY JĘZYKA POLSKIEGO

Zdolność rozstrzygania wieloznaczności wyrazu na poziomie syntaktycznym znacząco wpływa na polepszenie skuteczności działania algorytmòw z dziedziny przetwarzania języka naturalnego – wykorzystanie odpowiedniej konstrukcji baz odmian słòw języka jest pierwszym etapem działania większości zaawansowanych algorytmòw przetwarzania tekstu. Specyfika języka powoduje jednak obecność wieloznaczności także na poziomie semantycznym. Zjawisko to nazywane jest polisemią, gdy słowa wieloznaczne wywodzą się ze wspòlnego źròdła (przykładem czego jest słowo *język*, w rozumieniu *organu ciała* lub *mowy ludzkiej*) lub – w przeciwnym razie, czyli braku posiadania wspòlnych korzeni – homonimią. [9]

Homonimia może występować w:

- morfologii fleksyjnej – gdy pokrywają się jedynie niektòre formy odmiany, np. słowo *dam*, jako forma rzeczownika *damy* i czasownika *dawać*,
- słowotwòrczej – np. *ranny* jako *zraniony*, lub *ranny* w rozumieniu *poranka*,
- słownictwie – np. *bal* jako *przyjęcie* (z języka francuskiego), lub *bal* jako *kłoda drewna* (z języka niemieckiego),
- składni – np. wyrażenie *zdrada przyjaciela* ktòre może być rozumiane jako *zdradzenie przyjaciela* lub *zostanie zdradzonym przez niego*.

Słownik semantyczny to zbiòr słòw wraz z informacjami o ich znaczeniach. Znaczenie to określane jest poprzez zestaw relacji łączących wyraz definiowany z wyrazami go definiującymi.

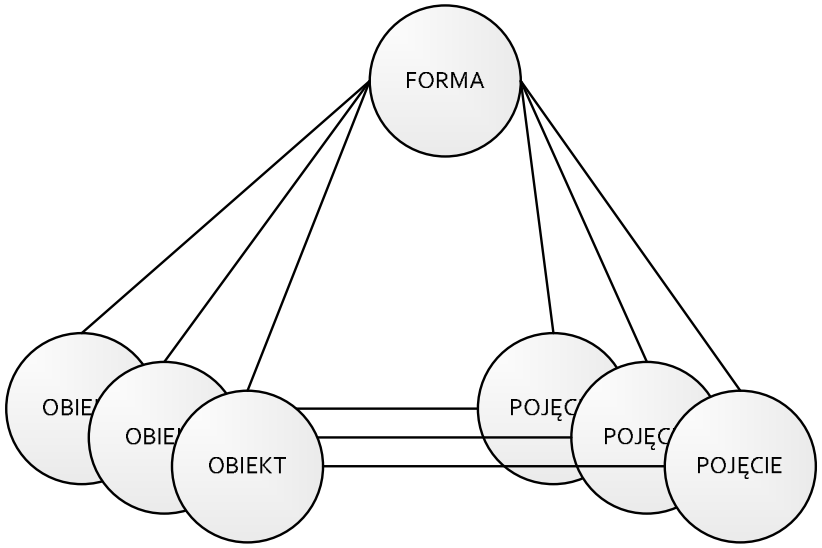
Wykorzystanie słownika semantycznego potrafiącego przekształcić pisemną reprezentację wyrazu w struktury symbolizujące jego znaczenie jest ważne dla zapewnienia skuteczności działania algorytmòw przetwarzania języka naturalnego. Brak tego mechanizmu uniemożliwia przeniesienie

[9] E. Olinkiewicz, K. Radzymińska, H. Styś, *Słownik Encyklopedyczny – Język polski*, Wydawnictwo Europa 1999.

rozumowania z poziomu wyrazòw, na poziom operacji na pojęciach połączonych ze sobą relacjami semantycznymi, co jest sposobem symulacji rozumienia tekstu przez maszynę.

2.2. DEFINICJE W RAMACH SEMANTYKI

Semantyka to dyscyplina badająca relacje pomiędzy nazwami, a przedmiotami ktòrych dotyczą. Semantyka zajmuje się badaniem znaczeń, interpretacją znakòw, zdań i wyrażeń języka. Problem wieloznaczności w semantyce, pojawia się przy odwzorowaniu obiektòw i ich abstrakcji, czyli pojęć w nazwy – często bowiem zdarza się, że ròżnym obiektom odpowiada ta sama nazwa (rysunek 1).



Rys. 1. Tròjkąt semiotyczny C.K. Ogdena i A. Richardsa a wieloznaczność

Poniżej omòwiono podstawowe pojęcia związane z semantyką i zagadnieniami dotyczącymi przetwarzania języka naturalnego.

Najmniejszą, nierozkładalną część znaczenia nazywamy *cechą semantyczną* (inaczej *sem*). Cechy semantyczne znaczenia danej nazwy nie posiadają jednakowej rangi – można je pogrupować w trzy *strefy konotacyjne*:

- I strefa konotacyjna – obejmuje tzw. *kryterialne cechy semantyczne*, cechy wspòlne dla wszystkich przedmiotòw oznaczanych przez dany wyraz, pojęcie czy też znak (*desygnaty* danej nazwy), np. *kot* ma *ogon*,
- II strefa konotacyjna – obejmuje tzw. *konotacje semantyczne*, cechy łączone z danym przedmiotem stereotypowo, na zasadzie doświadczenia kulturowego z daną klasą obiektòw; dla słowa *kot* mogą to być takie cechy jak: *nielubienie psòw*, *łapanie myszy*, *przynoszenie nieszczęścia przez koty o czarnej sierści* itp. – nie można cech tych uznać za kryterialne, gdyż można znaleźć koty niespełniające ich,
- III strefa konotacyjna – obejmuje cechy indywidualne łączone z danym obiektem przez danego użytkownika języka, wynikające z jego osobistych przekonań lub doświadczeń, np. *przynoszenie kapci* przez *kota*. [10]

Zestawienie wszystkich cech semantycznych znaczenia danego znaku językowego tworzy jego *definicję kognitywną* – najbardziej wyidealizowany model, stanowiący idealnie typowy przykład egzemplarza danej kategorii.

Sumę znaczeń danego znaku językowego nazywamy *polem semantycznym* (lub znaczeniowym). Obejmuje ono zaròwno wszystkie konotacje (cechy wspòłoznaczane łącznie przez nazwę, tworzące jej sens i treść) jak i denotacje (zbiòr desygnatòw, czyli cech wspòlnych obiektu) wyrazu. Do pola semantycznego danego znaku językowego należą te znaki, ktòrych kryterialne cechy semantyczne mieszczą się w jego konotacjach.

Na podstawie istnienia cech semantycznych jesteśmy w stanie przyporządkowywać obiekty do określonych *kategorii semantycznych*, czyli zbioròw obiektòw wchodzących w zakres określonego znaczenia. Kategorie mogą być całkowicie rozłączne, zawierać się w sobie, przecinać czy też pokrywać – determinując właściwą dla danego języka i kultury taksonomię świata.

Uważa się, że podział na kategorie jest jedną z podstawowych zasad ludzkiego myślenia (E. Sapir, 1978). Za łączenie w grupy informacji o podobnym charakterze odpowiedzialna jest pamięć semantyczna. Jest ona określana jako rodzaj pamięci skojarzeniowej, odpowiedzialnej za etykietowanie zbioròw informacji o podobnym charakterze.

[10] M.Brzozowska *Etymologia a konotacja wybranych nazw kamien*i w: *Etnolingwistyka. Problemy języka i kultury* nr 12, Lublin 2000.

2.3. WIELOZNACZNOŚĆ SŁÒW

Wyròżnić można trzy głòwne podejścia do semantyki i zagadnienia znaczenia wyrazu:

- składnikowe – opracowane m.in. przez P.E. Goddarda i Noama Chomsky'ego, wywodzące się z filozofii i logiki, a stanowiące rozwinięcie myśli Arystotelesa,
- prototypowe – ktòre opracowali: Marvin Lee Minski, Roger Schank, Charles J. Fillmore, a wywodzące się z psychologii,
- relacyjne – opracowane przez George'a A. Millera.[11]

2.3.1. PODEJŚCIE SKŁADNIKOWE

Zjawisko mnogości znaczenia wyrazu jest od dawna obserwowane i badane. Już w IV wieku p.n.e. Arystoteles w dziele „Organon" porusza zagadnienie bytu i ròwnocześnie wskazuje na wielość jego znaczeń wymieniając cztery podstawowe:

- byt akcydentalny,
- byt wedle kategorii,
- byt jako prawda i fałsz,
- byt potencjalny i aktualny.

Dzieli byty według dziesięciu kategorii:

- substancja albo istota,
- jakość,
- ilość,
- relacja,
- działanie,
- doznawanie,
- miejsce,
- czas,
- posiadanie,
- położenie.

Arystoteles zauważa, że tę samą myśl można wyrazić na wiele sposobòw ròżniących się jedynie stylistyką języka, ale nie kategorią (jak na przykład wyrażenia: *człowiek powraca do zdrowia* i *człowiek jest powracający do zdrowia*).

[11] Y.Ravin, C.Leacock, *Polysemy: Theoretical and Computational Approaches*, Oxford University Press 2000, s.7 – 18.

Warto zaznaczyć, że pomimo zwròcenia uwagi na fakt wieloznaczności bytu, Arystoteles nie określa żadnej relacji pomiędzy nimi – istnienie tej złożonej relacji pomiędzy słowem a jego znaczeniem zostało po raz pierwszy dostrzeżone przez stoikòw. Zauważyli oni, że jedno pojęcie może być opisane przez wiele ròżnych słòw (zjawisko synonimii), a także że pojedyncze słowo może oznaczać wiele pojęć (zjawiska polisemii i homonimii).

Zgodnie z tym podejściem, każdy istniejący obiekt należy do ktòrejś z wymienionych wyżej kategorii. Możemy powiedzieć, że rozpatrywany obiekt należy do pewnej kategorii, jeśli posiada zestaw cech koniecznych (inaczej definiujących) oraz wystarczających (bazowych) dla niej.

Tradycyjnie, przyjmuje się kilka założeń co do tego podejścia:

- kategorie składają się z list właściwości połączonych operatorami logicznymi takimi jak koniunkcja czy alternatywa,
- kategorie uporządkowane są w strukturze hierarchicznej, gdzie pojęcia znajdujące się na tym samym poziomie współdzielą odziedziczone właściwości bazowe pojęć z poziomòw wyższych, ale mają wzajemnie wykluczające się zbiory właściwości definiujących.

Nowoczesna wersja tego podejścia powstała z dokonanej przez Jerrolda Katza i Jerry'ego Fodora pròby kompilacji klasycznego ujęcia teorii znaczenia, z pracami Noama Chomsky'ego opublikowanymi w 1957 roku. [12] Chomsky zauważył, że sama składnia nie tworzy podstawowej struktury języka, a słowa to znaki o określonych właściwościach, ktòre mogą funkcjonować jedynie w odpowiednim kontekście semantycznym. Dowodem jest słynne zdanie jego autorstwa: "Bezbarwne zielone idee wściekle śpią" (ang. "Colorless green ideas sleep furiously") – jest ono poprawne gramatycznie, chociaż bezsensowne. Jego poprawność gramatyczną dowodzi rozkład logiczny, zgodnie z zaproponowanym przez Chomsky'ego fragmentem gramatyki języka angielskiego, w postaci:

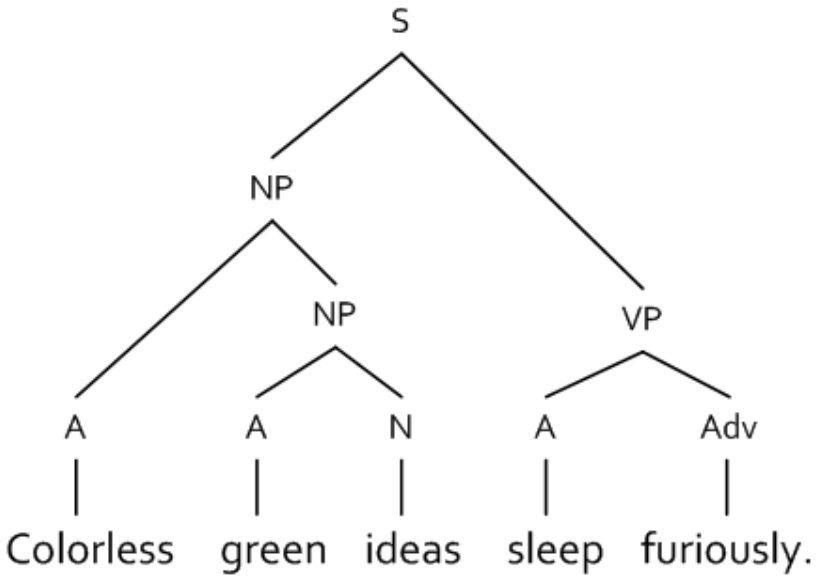
S → NP VP

NP → Det N1

N1 → (AP) N1 (PP)

[12] N.Chomsky, *Syntactic Structures*, Mouton, The Hague 1957.

Rozkład powyższego zdania wygląda następująco:



Rys. 2. Rozkład zdania Chomsky'ego

Celem Katza i Fodora było wyartykułowanie zasad uniwersalnej teorii semantycznej, ktòra tłumaczyłaby reguły rządzące relacjami pomiędzy słowami a ich znaczeniami. Konfrontacja podejścia klasycznego z rzeczywistością okazała się trudna ze względu na rozmycie granic kategorii semantycznych.

2.3.2. PODEJŚCIE PROTOTYPOWE

Problemem dla podejścia klasycznego, ktòry wystąpił przy pròbie podziału na kategorie był fakt, że ròżne znaczenia słowa wykazują podobieństwa poprzez zazębianie i przecinanie się właściwości, na takiej samej zasadzie jak cechy gatunkòw.

W psychologii, teoria kategoryzacji poprzez podobieństwo została przedstawiona w latach siedemdziesiątych przez Eleanor Rosch. Stwierdziła ona, że ludzie nie kategoryzują obiektòw poprzez posiadanie przez nich cech koniecznych i wystarczających, ale na zasadzie podobieństwa ich do prototypòw – typowych przedstawicieli danej kategorii.

Podejście to uznaje zatem istnienie hierarchicznych kategorii pojęć, ale zmienia sposòb przypisywania do nich obiektòw poprzez wprowadzenie reprezentującego ją prototypu. Rosch prezentuje dwa rodzaje prototypòw:

- pojedynczy obiekt posiadający możliwie największą liczbę charakterystycznych cech,
- kilka obiektòw, z ktòrych każdy posiada oddzielny zestaw charakterystycznych cech, niekoniecznie do siebie zbliżonych – to ten rodzaj grupowego prototypu został przyjęty przez językoznawcòw.

W swoich badaniach Rosch wielokrotnie wykazywała, że ludzie poznają nowe pojęcia poprzez przybliżanie ich znanymi sobie, istniejącymi obiektami. Prototypy są więc silnie zależne od cech indywidualnych i kulturowych, a stwierdzenie czy dany obiekt należy do kategorii jest tym prostsze i szybsze, im bardziej obiekt jest podobny do prototypu.

Podejście to zostało zaadaptowane przez wielu lingwistòw, między innymi wymienionego powyżej Charlesa J. Fillmore'a.

2.3.3. PODEJŚCIE RELACYJNE

W podejściu relacyjnym, słowa zgodnie ze swoim znaczeniem zorganizowane są w formę relacyjnej sieci semantycznej. Tak jak w przypadku podejścia prototypowego, model semantyczny pracuje na domenach semantycznych. W idealnym przypadku, znalezienie znaczenia słowa sprowadza się do znalezienia jego lokalizacji w strukturze modelu.

Istnieje wiele typòw relacji, ktòre mogą być uwzględnione w strukturze sieci semantycznej. Możemy je podzielić na:

- relacje paradygmatyczne – dotyczące słòw reprezentujących podobne pojęcia i często pojawiających się w zbliżonym kontekście, ale rzadko razem ze sobą (np. słowa *gładki* i *jedwabisty*), a także pojęcia będące swoimi generalizacjami lub specjalizacjami,
- relacje syntagmatyczne – dotyczące słòw wspòłwystępujących często w wypowiedziach języka naturalnego (np. *gładki* i *jedwab*).

Jako najbardziej podstawowe uznaje się następujące relacje paradygmatyczne:

- synonimię – jedno znaczenie może być wyrażane przez kilka słòw,
- antonimię – przeciwieństwo znaczeń,
- hiperonimię – relacja generalizacji,
- hiponimia – relacja specjalizacji.

Przykładem realizacji podejścia relacyjnego jest słownik semantyczny WordNet omòwiony w następnym rozdziale.

2.4. REALIZACJE SŁOWNIKA SEMANTYCZNEGO

Przykładami słowniköw semantycznych są: WordNet (język angielski), EuroWordNet (język czeski, duński, estoński, francuski, hiszpański, niemiecki i włoski), Multilingual Central Repository, Global WordNet i inne.

WordNet to słownik semantyczny języka angielskiego stworzony przez Georga A. Millera i jego współpracowniköw z Princeton University. Budowę słownika rozpoczęto w 1985 roku. Początkowo projekt traktowany był na zasadzie eksperymentu i miał zbadać granice możliwości relacyjnego opisu pojęć poprzez sprawdzenie, jak wiele wiedzy można w ten sposòb zapisać. Zgodnie z informacjami podawanymi na stronie projektu, w tym momencie baza słòw zawiera 206941 elementòw.

Słownik WordNet grupuje rzeczowniki, czasowniki, przymiotniki i przysłòwki w kognitywne zbiory „prawie synonimòw" nazywane synsetami, z ktòrych każdy wyraża osobne pojęcie. Są one ze sobą połączone za pomocą relacji semantycznych i leksykalnych. Synset jest zapisywany w postaci zbioru jednostek leksykalnych reprezentujących poszczegòlne leksemy np.:

> { anger, choler, ire – (a strong emotion; a feeling that is oriented toward some real or supposed grievance) }
>
> { anger, angriness – (the state of being angry) }

Pomiędzy synsetami zachodzą relacje semantyczne. Założono, że pomiędzy dwoma synsetami relacja taka zachodzi wtedy i tylko wtedy, gdy występuje ona dla wszystkich par elementòw z obu synsetòw.

W podejściu relacyjnym wykorzystywane są oba typy relacji semantycznych, jednak istniejące słowniki w większości przypadkòw (jak np. WordNet) ograniczają się jedynie do relacji paradygmatycznych dzielących pojęcia na kategorie semantyczne:

NAZWA RELACJI	OPIS	PRZYKŁAD
SYNONYMY	Synonimia	*zamek(błyskawiczny) – synonymy – suwak*
IS A PART OF	meronimia, relacja bycia częścią	*zamek(budowla) – is a part of – krajobraz*
CONSISTS OF	holonimia, przeciwieństwo meronimii	*zamek(budowla) – consists of – budynki mieszkalne*
IS A	hiperonimia, przeciwieństwo hiponimii	*budowla – is a – zamek(budowla)*
IS A KIND OF	hiponimia, relacja bycia rodzajem	*zamek(budowla) – is a kind of – budowla*

Tabela 1. Relacje paradygmatyczne słownika semantycznego

Relacje te klasyfikują pojęcia, opisując hierarchiczne zależności pomiędzy nimi. Sama znajomość taksonomii nie jest jednak wystarczająca do wyłuskania kontekstu użycia wyrazu, ponieważ zależności tego typu rzadko występują w zdaniach w języku naturalnym. Dla zapewnienia skuteczności wyszukiwania kontekstu, należy uwzględnić drugi typ relacji, z ktòrymi to mamy w zdaniach najczęściej do czynienia. Mowa tu o relacjach syntagmatycznych, opisujących akcje i stany, w jakich może dane pojęcie się znajdować.

Zaròwno twòrcy WordNet'u, jak i innych słownikòw, zapowiadają rozszerzenie ich o tego typu relacje, jednak jak na razie pozostaje to w sferze planòw. Pozostałymi propozycjami uzupełnienia WordNetu są:

- Braki w połączeniach znaczeń z właściwym ich użyciem,
- Słabe odzwierciedlenie wag znaczeń w języku naturalnym,
- Zapewnienie wzajemnego wykluczania się znaczeń w ramach synsetòw,
- Lepsze osadzenie w języku współczesnym.[13]

[13] P.Hans, http://www.fi.muni.cz/gwc2004/pres/panel/Hanks/hanks-panel.pdf.

2.5. SŁOWNIK SEMANTYCZNY KATEDRY LINGWISTYKI KOMPUTEROWEJ UJ

W trakcie implementacji wykorzystaliśmy słownik semantyczny stworzony poprzez Katedrę Lingwistyki Komputerowej Uniwersytetu Jagiellońskiego. Słownik ten znajduje się w fazie mocno rozwojowej i – jak na razie – zawiera niecałe 52000 pojęć.

2.5.1. RELACJE SYNTAGMATYCZNE W SŁOWNIKU SEMANTYCZNYM UJ

Słownik opracowany na Uniwersytecie Jagiellońskim tym różni się od innych dostępnych słowników semantycznych, że oprócz wymienionych powyżej relacji paradygmatycznych, został on rozszerzony o relacje syntagmatyczne opisujące akcje i stany w jakich może znajdować się rozpatrywany obiekt.

Pełny spis relacji syntagmatycznych wykorzystanych w słowniku przestawiony został w tabeli 2:

NAZWA RELACJI	**OPIS**	**PRZYKŁAD**
ACTION	czynności wykonywane przez pojęcia, co do których nie można określić ich zgodności czy też niezgodności z przeznaczeniem	*mięta – action – pachnieć*
ACTION NEGATIVE	akcja negatywna (czyli niezgodna z przeznaczeniem) z punktu widzenia pojęcia	*admirał – action negative – poddanie się*
ACTION NEGATIVE RT (related to)	pojęcia, których dotyczyć może akcja negatywna	*admirał – action negative rt poddanie się – przeciwnik*
ACTION PASSIVE	akcja bierna, w której podmiot jest odbiorcą czynności	*admirał – action passive – zranienie*
ACTION PASSIVE RT (related to)	pojęcia powiązane z istniejącą akcją bierną	*admirał – action passive rt zranienie – przeciwnik*
ACTION POSITIVE	akcja pozytywna (zgodna z przeznaczeniem) z punktu widzenia pojęcia	*admirał – action positive – zwyciężanie*

ACTION POSITIVE RT (related to)	Pojęcia powiązane z istniejącą akcją pozytywną	*admirał – action positive rt zwyciężanie – przeciwnik*
ACTOR	aktor	*bicie – actor – gracz*
DESTINATION	przeznaczenie, określa do czego służy dany przedmiot	admirał – destination – dowodzenie
DESTINATION RT (related to)	pojęcia powiązane z istniejącą relacją przeznaczenia	admirał – destination rt – ludzie
OBJECT	przedmiot mogący być celem danej czynności	*bicie – object – pionek*
ROLE	rola w jakiej obiekt może występować, specjalizacja	*admirał – role – szef sztabu*
ROLE RT (related to)	pojęcia powiązane z relacją roli	*admirał – role rt – flota*
SIMILAR TO	relacja bycia podobnym do	*zamek(budowla) – similar to – twierdza*
SOURCE	materiał wykonania	*zamek(w drzwiach) – source – metal*
STATE	stan w jakim może znajdować się pojęcie	*admirał – state – czytanie*
STATE NEGATIVE	stan negatywny w jakim może znajdować się pojęcie	*admirał – state negative – tchòrzostwo*
STATE NEGATIVE RT (related to)	pojęcia powiązane ze stanem negatywnym	*admirał – state negative rt tchòrzostwo – przeciwnik*
STATE POSITIVE	stan pozytywny w jakim może znajdować się pojęcie	*admirał – state positive – lojalność*
STATE POSITIVE RT (related to)	pojęcia powiązane ze stanem pozytywnym	*admirał – state positive rt lojalność – przełożony*

Tabela 2. Relacje syntagmatyczne słownika semantycznego

Relacje typu RT (related to) umożliwiają wprowadzenie relacji tròjstronnych, określając zachodzenie pewnej relacji warunkowo, w stosunku do określonego obiektu.

Powyższe relacje odzwierciedlają zależności występujące pomiędzy pojęciami, rozszerzając taksonomię. Dzięki ich istnieniu możliwe jest określenie w sposòb bardziej precyzyjny związkòw występujących pomiędzy pojęciami.

2.5.2. PRÒBKA SŁOWNIKA SEMANTYCZNEGO

Poniżej przedstawiona została pròbka słownika semantycznego dla hasła ADMIRAŁ:

Hasło słownika: admirał (oficer marynarki)

CATEGORY: human
IS A PART OF: marynarka, flota,
IS A:
IS A KIND OF: marynarz
DESTINATION: dowodzić/dowodzenie, walczyć/walka, żeglować/żeglowanie/żegluga
ROLE:
RELATED TO: flota, eskadra, flotylla IS: dowòdca, szef sztabu
ACTION:
POSITIVE:
RELATED TO: przeciwnik IS: pokonać, zwyciężyć/zwycięstwo, zatopić/zatopienie
NEGATIVE:
RELATED TO: morze, woda, ocean IS: utonąć
RELATED TO: przeciwnik IS: przegrać/przegrana, poddać się
RELATED TO: burta IS: wypaść (za)
PASSIVE: choroba, niedyspozycja
REALTED TO: przeciwnik IS: dostać, oberwać, rana, zraniony, ranny, obrażenia, zginąć
STATE:
POSITIVE: odwaga/odważny, służba/służyć
RELATED TO: podwładny IS: autorytet, szacunek/szanowany, popularność/popularny, ceniony
RELATED TO: przełożony IS: posłuszny/posłuszeństwo, lojalny/lojalność
NEGATIVE: rezerwa, emerytura
RELATED TO: przeciwnik, żywioł IS: tchòrz/tchòrzostwo/tchòrzliwość

Hasło słownika: admirał (motyl)

CATEGORY: animal
IS A PART OF: natura
IS A:
IS A KIND OF: motyl
DESTINATION: zapylić/zapylać/zapylanie
ROLE:
ACTION:
POSITIVE:
RELATED TO: kwiat IS: (zbierać) nektar
NEGATIVE: choroba, niedyspozycja
PASSIVE: zginąć
STATE:
POSITIVE: żyć/żywy
NEGATIVE: martwy

Powyższy przykład prezentuje relacje paradygmatyczne i syntagmatyczne dla wyrazu *admirał*. Pierwsze z nich pozwalają na kategoryzację pojęć, drugie odzwierciedlają związki pomiędzy elementami świata rzeczywistego. Ich obecność daje nam możliwość wychwycenia dodatkowych zależności pomiędzy słowami obecnymi w tekście i może posłużyć do rozstrzygnięcia ewentualnych wieloznaczności wyrazòw.

ROZDZIAŁ III

ROZSTRZYGANIE WIELOZNACZNOŚCI

Stworzony algorytm stanowi kolejny element, mający na celu polepszenie procesu maszynowego rozumienia tekstu. Jego użycie pozwala na poprawę jakości działania narzędzi z dziedziny przetwarzania języka naturalnego, obejmujących automatyczne tłumaczenie tekstu, kategoryzację tekstòw językowych, pozycjonowanie reklam kontekstowych, wyszukiwanie informacji na podstawie wyrażeń języka naturalnego i inne.

Idea działania algorytmu opiera się na następujących obserwacjach:

> [1] Dla każdego potrafiącego budować kontekst wypowiedzi słowa, można wskazać zbiòr innych słòw bliskich mu znaczeniowo,
>
> [2] Dodawanie do wypowiedzi słòw bliskich znaczeniowo słowom budującym kontekst skutkuje wzmocnieniem go,

Oraz:

> [3] Właściwe znaczenie słowa wieloznacznego w wypowiedzi determinowane jest przez kontekst, w ktòrym ono występuje, ktòry tworzony jest przez istniejąc relacje syntagmatyczne.

Zadanie algorytmu, ktòrym jest wskazanie najtrafniej pasującego do kontekstu znaczenia słowa wieloznacznego, sprowadza się zatem do znalezienia dla istotnych słòw wypowiedzi, najbliższego im znaczeniowo homonimu. Odległość jest tu obliczana poprzez ilość i typ relacji semantycznych, za pomocą ktòrych możemy przemieszczać się między słowami. Wiedzę o zachodzących pomiędzy pojęciami relacjach dostarcza słownik semantyczny.

Algorytm w procesie ujednoznaczniania opiera się na trzech zmiennych:

- grafie – w ktòrym zapisane są informacje o powiązaniach semantycznych między słowami na podstawie informacji słownika semantycznego,
- zmiennych liczbowych – określających horyzont przeszukiwania grafu i wagi poszczegòlnych relacji,
- aktualnym wejściowym ciągu znakòw do sprawdzenia.

W pierwszym kroku algorytmu, wyekstrahowane zostają wszystkie występujące w wejściu słowa. Następnie poszukiwane są odpowiadające im wierzchołki. W przypadku słòw wielosegmentowych (np. *energia słoneczna*) znaczenie ma kolejność ich występowania w zdaniu (np. w zdaniu *panna była młoda*, algorytm nie wyszuka wielosegmentowego słowa *panna młoda*). Na tym etapie ma też miejsce podział słòw na jedno- i wieloznaczne, ktòrego kryterium jest ilość znaczeń słowa w grafie. W kolejnym kroku, dla każdego słowa, ktòre zostało uznane za wieloznaczne uruchamiany jest algorytm przeszukiwania grafu wszerz. Poszukuje on skierowanych ścieżek łączących to słowo z pozostałymi znalezionymi. Wagi relacji na poszczegòlnych ścieżkach są sumowane. Otrzymana wartość określa odległość znaczeniową pomiędzy słowami. Algorytm wybiera to, ktòrego suma jest minimalna, a ilość znalezionych słòw największa (algorytm potraktuje tak samo fakt znalezienia jednego słowa w odległości 1, jak dwòch słòw w odległości 2). Dodatkowo zwracany jest procentowy udział tej sumy w ramach sum wszystkich ścieżek, co jest interpretowane jako stopień pewności wyboru danej ścieżki.

Ze względu na trudność określenia faktycznych granic wypowiedzi, aspekt ten został pominięty – zakłada się, że na wejściu programu znajduje się wystarczająca ilość informacji do określenia kontekstu. Przy analizie tego zagadnienia, konieczne jest uwzględnienie składni języka polskiego, w celu określenia granic zdań wtrąconych i zależności pomiędzy poszczegòlnymi ich częściami. Ze względu na naturę języka, odseparowanie wpływu kontekstu wypowiedzi od jej składni jest niemożliwe. Stworzenie pełnego algorytmu obejmującego oba te zagadnienia wykracza poza ramy tej pracy magisterskiej i zostało pominięte w dalszych rozważaniach. [14]

Ujednoznacznienie tekstu przeprowadzane jest dla wejściowej wypowiedzi. Na żadnym z etapòw działania algorytmu nie jest wykonywana operacja korekty pisowni, dlatego zakładamy poprawność zaròwno słownika semantycznego jak i wejściowego ciągu znakòw.

[14] Analiza składniowa odbywa się najczęściej w oparciu o pewną sformalizowaną gramatykę, omówienie istniejących rozwiązań można znaleźć w: A.Mykowiecka, *Inżynieria lingwistyczna, Komputerowe przetwarzanie tekstów w języku naturalnym*, Warszawa 2007.

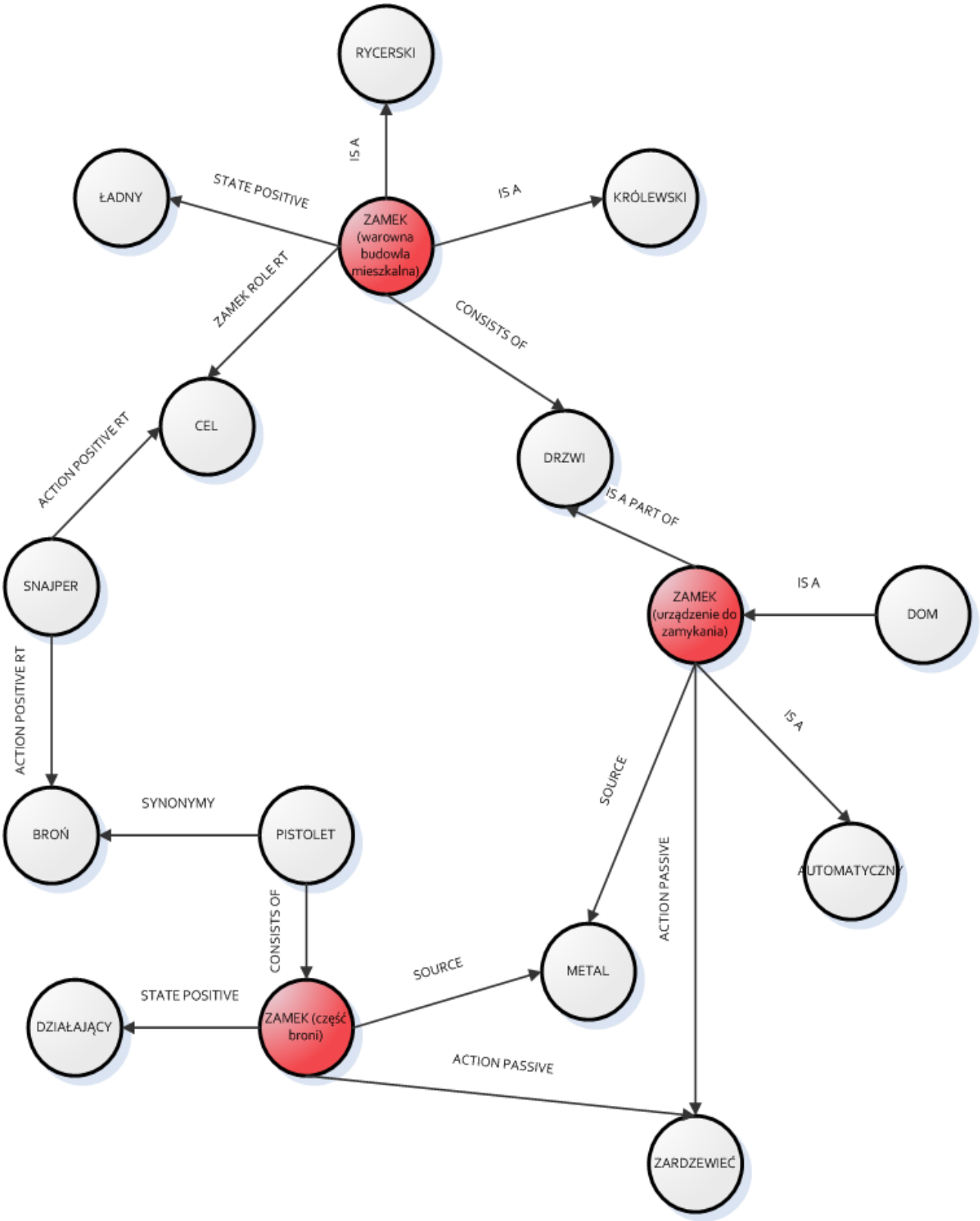
3.1. MODEL

W celu przeprowadzenia procesu ujednoznaczniania słòw, konieczna jest wiedza o wzajemnych zależnościach pomiędzy wyrazami. W tym celu posługujemy się słownikiem semantycznym, ktòry stanowi statyczną reprezentację relacji zachodzących pomiędzy pojęciami języka naturalnego. Algorytm, o ktòry oparto działanie programu jest algorytmem grafowym. Aby możliwe było przeprowadzanie procesu rozumowania i znalezienie zależności pomiędzy słowami, konieczne było odpowiednie odwzorowanie słownika semantycznego w strukturę grafu skierowanego, ważonego.

Poniżej przedstawiono odwzorowanie elementòw pomiędzy modelami:

- wierzchołki (odwzorowujące pojęcia) – reprezentują byty istniejące w modelu; dla każdego pojęcia istnieje odpowiadające mu słowo lub ich ciąg ze słownika semantycznego; na potrzeby modelu dla pojęć nie rozròżniamy części mowy ani wielkości liter,
- skierowane krawędzie (odwzorowujące relacje) – reprezentują istniejące zależności pomiędzy pojęciami; pokrywają się z relacjami istniejącymi w słowniku semantycznym; każdej relacji przyporządkowana jest dodatkowo liczbowa wartość określająca wagę relacji.

Dla potrzeb algorytmu wszystkie nazwy przechowywane w modelu występują w pierwszej formie podstawowej, ktòrą zwraca biblioteka CLP. W przypadku wczytania pojęcia z grafu, ktòre nie spełnia tego warunku jest ono sprowadzane do tej postaci.



Rys. 3. Fragment modelu dla słowa wieloznacznego *zamek*

Stworzony model może być przedstawiony w postaci ważonego grafu skierowanego [15], w ktòrym pojęcia przedstawione są jako wierzchołki, a relacje zachodzące pomiędzy nimi – wyrażane poprzez krawędzie. Relacje są skierowane i łączą ze sobą dwa, niekoniecznie ròżne (np. relacja zwrotna), pojęcia. Dla dwòch dowolnie wybranych słòw ze słownika semantycznego odległością w modelu nazywamy długość ścieżki pomiędzy nimi. Każda relacja posiada określający ją współczynnik ważności, mòwiący o tym, jak mocny jest związek pomiędzy dwoma pojęciami. Istnienie tego współczynnika motywowane jest mnogością relacji i potrzebą rozròżnienia wprowadzanych przez nie zależności przy jednoczesnym zachowaniu prostoty samego algorytmu. Jest to istotne, gdyż poszukiwanie w sąsiedztwie słowa wieloznacznego słòw współwystępujących z nim w zdaniu testowym odbywać się musi w pewnym promieniu, ktòrego wielkość nie powinna być ustalona arbitralnie, ale uzależniona od rodzaju rozpatrywanych relacji. Im waga relacji jest większa, tym wnosi mniejszy narzut na rozmiar promienia poszukiwań. Najsilniejszą relacją jest relacja synonimii. Przejście pomiędzy dwoma pojęciami z użyciem tej relacji ma zerowy koszt.

Słownik semantyczny z definicji zawiera relacje należące do pierwszej i drugiej strefy konotacyjnej, czyli informacje na temat ogòlnie znanych cech obiektòw. Nie dostarcza on zatem ani wiedzy, ani mechanizmu wnioskowania o cechach nietypowych, indywidualnych, skojarzeniach i prawdach relatywnych (np. fantazja literacka, operowania pewnymi relacjami tylko w danym zakresie i pod określonymi warunkami). Kolejnym ograniczeniem słownika jest fakt, że o ile informacje na temat instancji poszczegòlnych obiektòw dobrze wkomponowują się w jego strukturę, to nie ma łatwego sposobu na określenie zależności czasowych ani przestrzennych. Przykładem jest zdanie *Średniowieczne zamki były bardzo awaryjne.* – w procesie ujednoznaczniania słowa *zamek* można pominąć *zamek błyskawiczny*, ktòry został wymyślony w XX wieku.

3.2. OBRÒBKA ZDANIA

Wejściem algorytmu jest ciąg znakòw. W trakcie działania algorytmu wpływ znakòw interpunkcyjnych na kontekst w jakim występują słowa jest ignorowany. Wszystkie operacje związane z analizą interpunkcji, podziałem wypowiedzi na akapity i zdania muszą być przeprowadzane przed działaniem algorytmu. Podyktowane jest to chęcią oddzielenia od siebie zakresu działania poszczegòlnych modułòw biorących udział w procesie przetwarzania wypowiedzi. Jedyną operacją podczas ktòrej interpunkcja jest uwzględniana, jest podział wejścia na słowa. W pierwszym kroku algorytm dzieli wejście na poszczegòlne słowa, ktòre w dalszym opisie będziemy oznaczać jako wektor wejściowy. W celu wydajnego wyszukiwania wyrazòw wielosegmentowych zostało stworzone dodatkowe mapowanie, ktòre pozwala znaleźć je na podstawie pojedynczego fragmentu. Wyraz jest identyfikowany jako wieloznaczny tylko w przypadku, gdy wszystkie fragmenty wchodzące w jego skład znajdują się we właściwej kolejności w zdaniu.

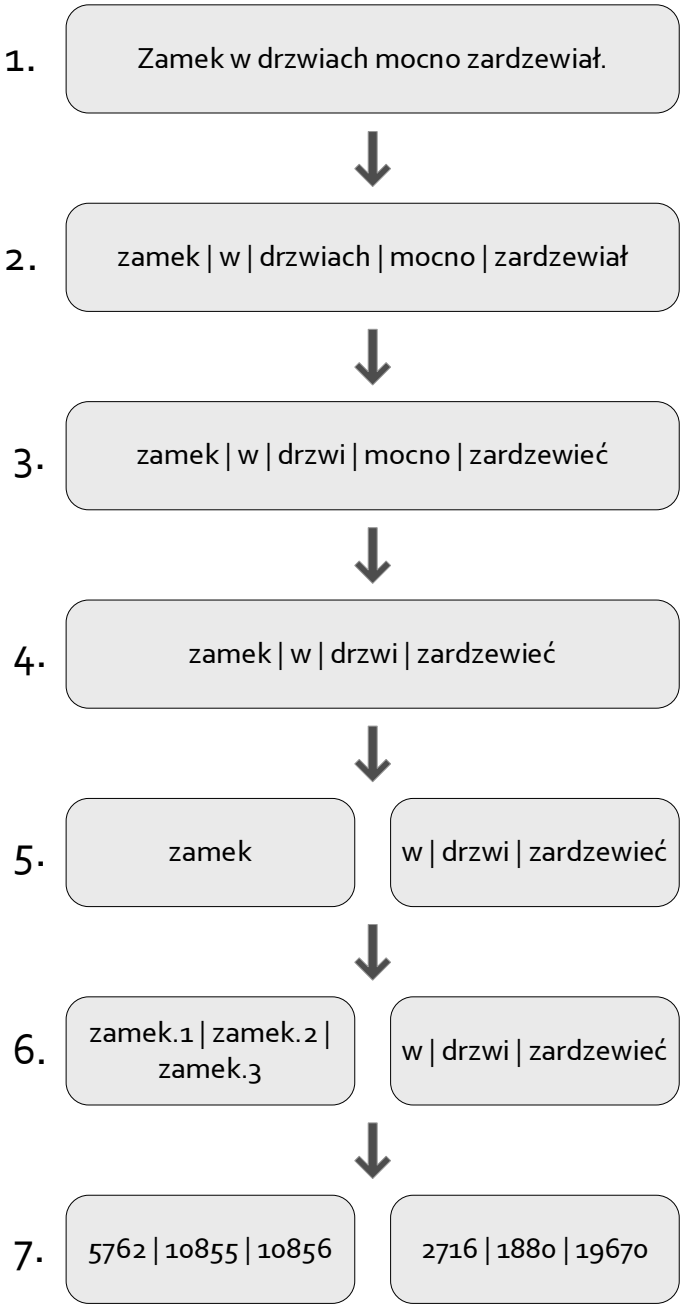
[15] T.Cormen, C.Leiserson, R.Rivest, C.Stein, *Introduction to Algorithms*, MIT Press 2001, s.527-529.

W następnym kroku przeprowadzana jest wstępna identyfikacja wyrazòw wieloznacznych. Algorytm pozwala na identyfikację homonimòw fleksyjnych (np. *jak* jako zwierzę, partykuła itd.). Pozwala to na ujednoznacznienie pojęcia na podstawie jego formy gramatycznej (np. forma *jakiem* bezpośrednio wskazuje na słowo *jak* w znaczeniu zwierzęcia). Niestety, model słownika semantycznego nie zawiera odwzorowania na identyfikator fleksyjny z biblioteki CLP. Powoduje to brak możliwości wykorzystania tego elementu w procesie działania algorytmu. Poza wyszukiwaniem wyrażeń wielosegmentowych, kolejność słòw wypowiedzi nie ma znaczenia dla działania algorytmu. Wybòr zbioru jako struktury przechowywania słòw podyktowany jest nieistnieniem narzędzi umożliwiających uwzględnienie składni w procesie ujednoznaczniania oraz brakiem potrzeby wielokrotnego rozpatrywania tych samych słòw (nieumiejętność prawidłowego określenia powiązań między słowami w zdaniu, granic wypowiedzi, itp.).

Przy pomocy biblioteki CLP, słowa z wektora wejściowego sprowadzane są do formy podstawowej. Podyktowane jest to faktem trzymania wszystkich słòw modelu w formie podstawowej. Algorytm zakłada poprawność ortograficzną wejściowej wypowiedzi. Na żadnym z etapòw algorytmu na zdaniu wejściowym nie jest wykonywana operacja korekty pisowni. Wszelką korektę pisowni należy przeprowadzić przed procesem ujednoznaczniania. Słowa, ktòre nie zostały znalezione przez CLP są traktowane jako błędne i usuwane z dalszej obròbki. Zakłada się, że definicje w bibliotece CLP pokrywają się z definicjami w modelu.

Następnie wyszukiwane w grafie są wszystkie wierzchołki odpowiadające słowom z wektora wejściowego. Jeżeli słowo w nim nie występuje, jest ignorowane (można traktować to jako jego usunięcie z wektora wejściowego). Wprowadzenie takiego ograniczenia powoduje zmniejszenie przestrzeni poszukiwań o słowa, ktòrych nie potrafimy interpretować bez wiedzy o składni wypowiedzi. Pozwala to zachować prostotę algorytmu.

Kolejne etapy algorytmu korzystają z dwòch zbioròw wierzchołkòw. W pierwszym przechowywane są wskazania na te, ktòre interpretowane są jako pojęcia wieloznaczne, w drugim – wskazania na pojęcia jednoznaczne. Dla każdego słowa z wektora wejściowego wyszukiwany jest odpowiadający mu wierzchołek grafu. Jeżeli dla danego słowa w grafie istnieje więcej niż jeden wierzchołek, algorytm automatycznie zakłada, że jest ono wieloznaczne i dodaje je do zbioru do ujednoznacznienia. Jeżeli istnieje tylko jeden wierzchołek, słowo jest dodawane do zbioru jednoznacznych. Algorytm kończy działanie, gdy zbiòr pozycji do ujednoznacznienia jest pusty (wszystkie słowa wypowiedzi zostały ujednoznacznione).



Rys. 4. Kroki przetwarzania tekstu wejściowego na przykładzie zdania: *Zamek w drzwiach mocno zardzewiał.*

3.3. ROZUMOWANIE W MODELU

Na tym etapie algorytm ujednoznacznia wyrazy homonimiczne i polisemiczne. W procesie ujednoznaczniania, wszystkie słowa ze zdefiniowanych wcześniej zbiorów traktowane są równorzędnie pod względem ważności. W dalszej analizie ignorowany jest wpływ składni języka polskiego na zasięg oddziaływania słów na siebie. Podyktowane jest to trudnościami z prostym opisaniem reguł składni języka polskiego i ich wpływem na proces ujednoznaczniania.

W celu uproszczenia dalszego opisu zostanie wprowadzone pojęcie granicy przeszukiwań przy zadanej wartości. Dla konkretnego wierzchołka grafu, jego granicą przy zadanej wielkości nazywamy wszystkie wierzchołki, które znajdują się w odległości mniejszej bądź równej zadanej wartości, przy czym odległość między wierzchołkami grafu rozumiana jest w sposób standardowy w teorii grafów, czyli jako suma wag najkrótszej skierowanej ścieżki łączącej te dwa wierzchołki.

Parametrami procesu przeszukiwania są:

- wierzchołek startowy – wierzchołek grafu, od którego liczona jest granica przeszukiwań; wierzchołkami grafu są wierzchołki pojęć wieloznacznych,
- waga startowa – określająca granicę przeszukiwania grafu od wierzchołka startowego; w momencie uruchomienia algorytmu aktualna wartość granicy jest równa wadze startowej,
- aktualna waga – określająca aktualną granicę przeszukiwania grafu w otoczeniu wierzchołka startowego,
- przyrost wagi – określający zmianę granicy przeszukiwania w grafie; aktualna waga jest zwiększana w przypadku braku trafień w danej granicy,
- maksymalna waga – wyznacza maksymalną granicę, w której wyszukiwane są powiązania pomiędzy pojęciami; jeżeli aktualna waga przekracza tę wartość, aktualny proces przeszukiwania grafu jest przerywany.

W następnym kroku algorytm ustawia wartość dla granicy na wagę startową. Po czym dla wybranego wierzchołka ze zbioru wieloznacznych przegląda graf w przód w poszukiwaniu ścieżek do wierzchołków z obu zbiorów wskazań, pomniejszonego o wierzchołki danego pojęcia. Podyktowane jest to koniecznością uniknięcia sytuacji, w której wierzchołkiem ujednoznaczniającym jest inne znaczenie danego pojęcia.

Algorytm nie uwzględnia wpływu słów z wektora wejściowego na zmianę wag relacji między słowami. Spowodowane jest to problemem z określeniem, które relacje z grafu mają wpływ na słowa wektora wejściowego.

Jeżeli w zadanej granicy wierzchołki nie zostaną znalezione, jest ona powiększana o przyrost wagi i proces przeszukiwania jest ponawiany. W przypadku znalezienia wierzchołka lub osiągnięcia granicy maksymalnej, proces jest przerywany. Jeżeli na danej głębokości znaleziono więcej niż jeden wierzchołek, informacja o nich jest zapisywana, a odległości do tych wierzchołków są sumowane. Po zakończeniu przeszukiwania, algorytm jest powtarzany na zbiorze krawędzi wstecz, a wierzchołek

źròdłowy staje się wierzchołkiem końcowym ścieżki. Ostatecznie zbiorem wierzchołkòw wskazujących znaczenie zostają te, ktòre znajdują się najbliżej wierzchołka startowego.

W tym momencie proces przeszukiwania grafu dla danego wierzchołka jest zakończony i ponawiany jest dla kolejnego słowa ze zbioru wieloznacznych. Proces przeszukiwania kończy się w momencie gdy dla wszystkich wierzchołkòw ze zbioru wieloznacznych przeszukano graf.

3.4. ANALIZA WYNIKÒW

Zastosowanie powyższej procedury powoduje znalezienie dla każdego ze znaczeń słowa wieloznacznego, wierzchołkòw z wypowiedzi, ktòre znajdują się najbliżej niego w grafie i określenie odległości do nich. Algorytm dokonuje podziału otrzymanych wynikòw na grupy, według słòw wieloznacznych i analizuje je osobno. W przypadku gdy tylko dla jednego z tych znaczeń wartość jest niezerowa, algorytm zwraca je w odpowiedzi. Gdy algorytm wskazał więcej niż jedno znaczenie, dla każdego z nich wyliczana jest suma odwrotności odległości poszczegòlnych wierzchołkòw. Premiuje to te znaczenia, wokòł ktòrych w minimalnej możliwej odległości znajduje się najwięcej wierzchołkòw. Po dodaniu sum algorytm potrafi wskazać procentową część każdego ze znaczeń. Na wyjściu algorytm zwraca wszystkie znalezione znaczenia oraz procentowy udział każdego z nich, przy czym jako najbardziej prawdopodobne wskazuje to, ktòrego suma jest największa. Najgorszym przypadkiem jest sytuacja, w ktòrej słowa znajdujące się w wypowiedzi wejściowej znajdują się w granicach wszystkich znaczeń. Pròbą poprawienia odpowiedzi dla takiego przypadku jest powiększenie grafu o nowe słowa, lub dodanie dodatkowych słòw do wypowiedzi wejściowej (powiększenie kontekstu).

ROZDZIAŁ IV

IMPLEMENTACJA

Stworzenie aplikacji składało się z dwòch etapòw:

- wygenerowania pliku modelu na podstawie otrzymanej kopii bazy danych słownika semantycznego,
- implementacji programu budującego graf sieci semantycznej na podstawie pliku modelu, a następnie ujednoznaczniającego wyrazy w zadanej wypowiedzi.

4.1. GENERACJA PLIKU MODELU

Aby umożliwić algorytmowi operowanie na danych słownika semantycznego, konieczne było wygenerowanie pliku zawierającego jego model. W tym celu stworzono zestaw skryptòw języka SQL wybierających poprawne dane z udostępnionego nam zrzutu bazy, a następnie formatujących je do właściwej postaci.

Plik modelu zawiera zaròwno pojęcia jak i relacje słownika. Każdy element zapisany jest w osobnej linii – najpierw pojęcia, a po nich relacje. Wyrazy jednoznaczne i wieloznaczne rozròżniane są poprzez kolejne numery nadawane im w trakcie formatowania – jako, że w modelu przechowujemy wyłącznie słowo, jest to konieczne dla jednoznacznego ich identyfikowania.

Poniżej wklejono przykładowe linie pliku:

```
akcept○wanie
akces○ria d○ wł○sòw
akcja.1(działanie)
akcja.2(fabuła)
akcja.3(papier wart○ś○ci○wy)
```

Dwie pierwsze pozycje to wyrazy jednoznaczne. Kolejne pojęcie – *akcja* – jest wieloznaczne i słownik zawiera trzy jego definicje. Każde znaczenie pojęcia wieloznacznego określane jest przez

kolejny numer i opis. Algorytm rozpoznaje pojęcie jako wieloznaczne poprzez zakończone sukcesem poszukiwanie numeru porządkowego pojęcia i jego opisu w nawiasach okrągłych.

Po definicjach wszystkich pojęć znajduje się zestaw relacji, czyli wszystkich powiązań między nimi.

```
zamek.1[IS A],aut○matyczny.1
zakup.3[SIMILAR T○],t○war.1
zadra[IS A],drzazga
f○lwark[IS A KIND ○F],wieś
```

Każde powiązanie znajduje się w osobnej linii i składa się z trzech części:

- wierzchołka początkowego – w przypadku pojęcia jednoznacznego, jego nazwy, a dla wieloznacznego, nazwy wraz z numerem porządkowym, już bez opisu,
- nazwy relacji w nawiasach prostokątnych,
- wierzchołka docelowego – także w postaci nazwy pojęcia waz lub bez numeru.

Plik tekstowy w takim formacie jest wczytywany do programu w momencie jego inicjalizacji i przechowywany w pamięci.

4.2. IMPLEMENTACJA UJEDNOZNACZNIANIA

Program ujednoznaczniający został zaimplementowany w języku C++. Jego działanie ma charakter sekwencyjny – w danym momencie może być analizowane pojedyncze zdanie testowe, a całość odbywa się w ramach jednego procesu. Kod programu jest przystosowany do pracy zaròwno w środowisku Windows (użycie kompilatora firmy Microsoft w ramach Microsoft Visual Studio), jak i w środowisku Unix/Linux (przy użyciu kompilatora gcc).

Program może być używany zaròwno jako niezależna aplikacja, jak ròwnież moduł biblioteczny. W przypadku uruchomienia programu jako niezależnej aplikacji zostały zaimplementowane dwa sposoby działania:

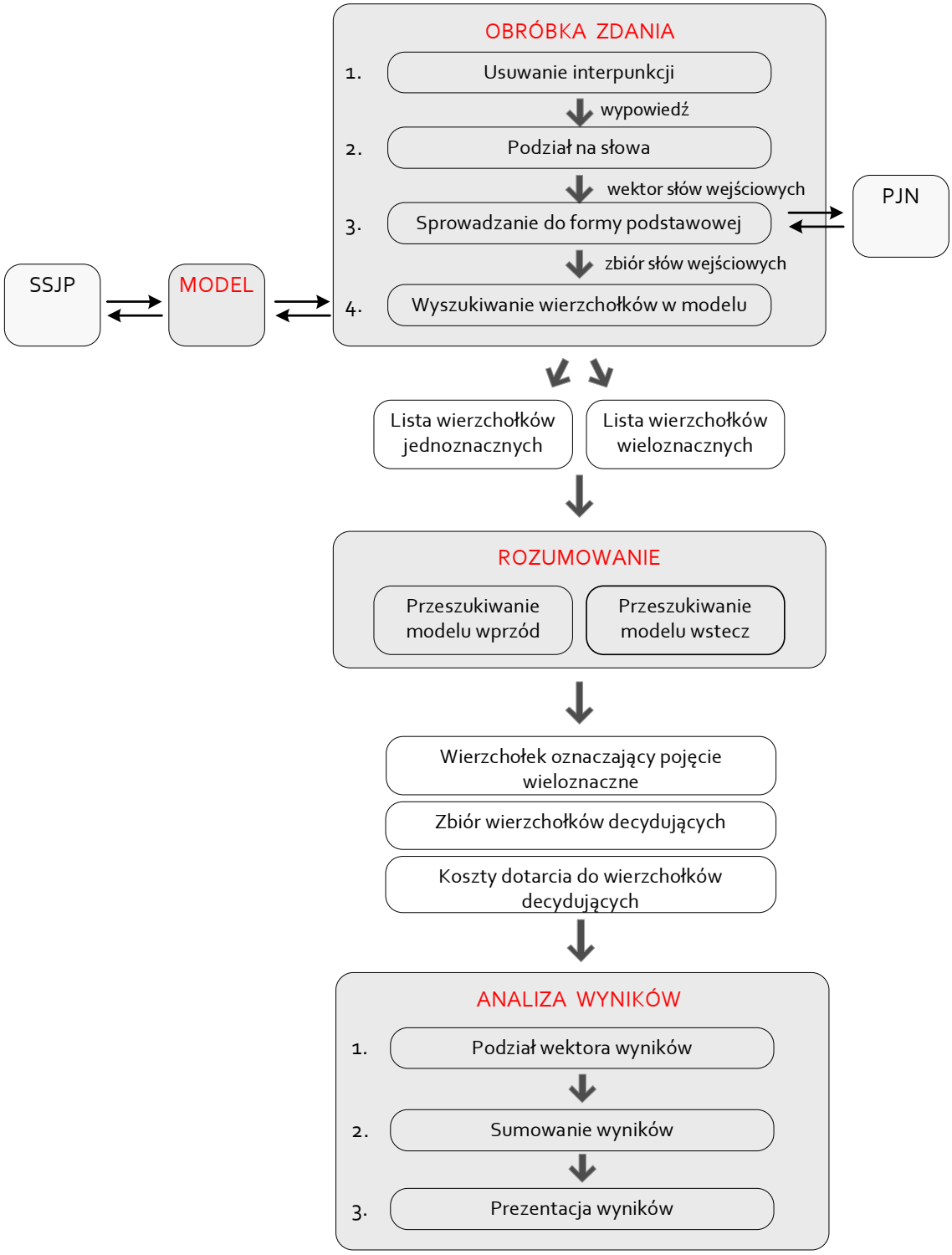
- jako program konsolowy,
- jako webservice.

W ramach aplikacji możemy wyròżnić następujące moduły:

- moduł modelu – odpowiedzialny za wczytywanie, przechowywanie i sprawdzanie spòjności struktur potrzebnych do prawidłowego działania algorytmu,
- moduł odpowiedzialny za obròbkę wejściowego zdania – mapujący słowa wchodzące w skład zdania wejściowego na wierzchołki grafu i identyfikujący w nim słowa wieloznaczne,
- moduł odpowiedzialny za przeglądanie grafu – wyszukujący kolejno powiązania pomiędzy wierzchołkami wieloznacznymi ze zdania wejściowego a pozostałymi,
- moduł odpowiedzialny za interpretację wynikòw – obliczający prawdopodobieństwo jakie algorytm przydziela poszczególnym znaczeniom wieloznacznym słòw.

Poza wymienionymi składowymi programu, do wykorzystywane są także dwa zewnętrzne moduły:

- biblioteka CLP – używana do ujednoznaczniania fleksyjnego słòw,
- Słownik Semantyczny Języka Polskiego – użyty do wygenerowania modelu.



Rys. 5. Moduły aplikacji i ich wzajemne zależności

Moduł odpowiedzialny za przechowywanie danych dla algorytmu korzysta z parametryzowanych struktur biblioteki stl. Podstawową strukturą wykorzystywaną w ramach programu jest klasa string używana zaròwno do przechowywania nazw pojęć i relacji w grafie, jak ròwnież jako klucze dla map odwzorowujących wyrazy w grafie. W ramach programu sieć semantyczna została zapisana w postaci skierowanego, ważonego grafu. Jego struktura jest przechowywana w postaci list sąsiedztwa.[16] Pojęcia sieci semantycznej zostały odwzorowane na wierzchołki, a relacje pomiędzy nimi – na krawędzie. Informacja o nich jest zapisana przy pomocy dwòch map:

- `map<SOURCEIDX, EDGES> _relationsForward;`
- `map<SOURCEIDX, EDGES> _relationsBackward;`

gdzie typy SOURCEIDX, TARGETIDX, RELATIONIDX i EDGES to odpowiednio:

- `typedef map<RELATIONIDX, set<TARGETIDX>> EDGES;`
- `typedef unsigned int SOURCEIDX;`
- `typedef unsigned int TARGETIDX;`
- `typedef int RELATIONIDX;`

Struktura _relationsForward odpowiada za przechowywanie relacji skierowanych od wierzchołka źròdłowego. Struktura _relationsBackward – za relacje wstecz. Informacje zawarte w nich duplikują się. Taka nadmiarowość pozwala na przyspieszenie procesu przeszukiwania grafu. Poszczególne wierzchołki (odpowiadające pojęciom) grafu są reprezentowane przez kolejne liczby naturalne. Użycie mapy podyktowane jest faktem istnienia wierzchołkòw, dla ktòrych nie istnieją powiązane z nimi relacje wejściowe czy wyjściowe.[17] Z każdym wierzchołkiem powiązana jest struktura EDGES, ktòra zawiera informacje o powiązanych z nim skierowanych krawędziach i wierzchołkach końcowych dla struktury _relationsForward i początkowych dla _relationsBackward. Ze względu na potrzebę rozròżnienia typòw relacji, struktura EDGES została zdefiniowana jako osobna mapa, ktòra separuje od siebie poszczególne relacje i grupy wierzchołkòw wyjściowych. Wykorzystanie mapy pozwala przyspieszyć sprawdzanie istnienia określonej relacji i, w razie potrzeby, usunięcie jej z procesu przeszukiwania. Obecność zbioru do przechowywania wierzchołkòw docelowych spowodowana jest potrzebą zmniejszenia narzutu na czas poszukiwania pojedynczego wierzchołka.

[16] T.Cormen, C.Leiserson, R.Rivest, C.Stein, *Introduction to Algorithms*, MIT Press 2001, s.529.

[17] U.Breymann, *Designing Components With the C++ STL (2nd Edition)*, Addison Wesley 2002, s.78-79.

W celu odwzorowania nazw poszczególnych wierzchołków, a co za tym idzie – również pojęć, zostały zdefiniowane następujące struktury:

- `vector<string> _vWordIdxToString;`
- `map<string, CONCEPTIDX> _mUnambiguousWords;`
- `map<string,set<CONCEPTIDX>> _mMultisegmentUniqueConcepts;`
- `map<string, vector<DESCRIPTION>>_mAmbiguousWords`
- `map<string,set<CONCEPTIDX>> _mMultisegmentAmbigConcepts;`
- `vector<string> _vMultisegmentWordsToAnalyze;`

gdzie typy DESCRIPTION i CONCEPTIDX to:

- `typedef pair<string,int> Description;`
- `typedef int CONCEPTIDX;`

Struktura _vWordIdxToString pozwala na wyszukanie nazwy wierzchołka skojarzonego z indeksem. Natomiast struktury _mUnambiguousWords i _mAmbiguousWords realizują odwrotną operację, przy czym pozwalają na dodatkowe określenie czy wierzchołek jest jedno- bądź wieloznaczny. Fakt oddzielenia ich od siebie został podyktowany potrzebą zdefiniowania opisów pojęć wieloznacznych i istnieniem więcej niż jednego wierzchołka dla pojęcia wieloznacznego, która to informacja w przypadku wierzchołka jednoznacznego byłaby nadmiarowa. Wartość liczbowa w opisie określa numer porządkowy znaczenia ze słownika semantycznego. Struktura _mMultisegmentAmbigConcepts pozwala na podstawie fragmentu słowa określić, czy istnieje wielosegmentowe, wieloznaczne słowo je zawierające i w przypadku odpowiedzi pozytywnej zwraca wektor ich wierzchołków. _mMultisegmentUniqueConcepts realizuje podobną funkcjonalność, z tym, że dla słów wielosegmentowych jednoznacznych. Struktura _vMultisegmentWordsToAnalyze używana jest na początkowym procesie działania algorytmu. Umieszczane są w niej wszystkie wielosegmentowe słowa, które w dalszym etapie pozwolą na wypełnienie struktur _mMultisegmentUniqueConcepts i _mMultisegmentAmbigConcepts.

W celu przechowywania informacji o dozwolonych dla algorytmu przeszukiwania relacjach i ich wagach, stworzone zostały następujące struktury:

- `map<string,RELATIONIDX> _mRelationDescriptionToIdx;`
- `map<RELATIONIDX,AllowedRelationProp> _vAllowedRelations;`

gdzie typ RELATIONIDX to:

- `typedef int RELATIONIDX;`

Natomiast AllowedRelationProp to struktura pomocnicza:

```
class AllowedRelationProp {
    public:
    int _forwardCost;
    int _backwardCost;
};
```

Struktura _mRelationDescriptionToIdx pozwala powiązać nazwę relacji z odpowiadającym jej indeksem. Natomiast AllowedRelationProp – określić koszt relacji przy przechodzeniu odpowiednio w przòd dla _forwardCost i wstecz dla _backwardCost.

Definicje dostępnych dla algorytmu relacji znajdują się w osobnym pliku tekstowym. Opis każdej relacji umieszczony jest w oddzielnej linii w nawiasach okrągłych. Na początku linii znajduje się tekstowa nazwa relacji, za nią zdefiniowane są dwie liczby naturalne określające koszt relacji przy przechodzeniu grafu w przòd i wstecz. Przykładowa postać pliku wygląda następująco:

```
(action,1,1)
(action negative,1,1)
(action negative rt,1,1)
```

Tworzenie modelu sieci semantycznej polega na wczytaniu zawartości pliku z zapisem grafu połączeń dla odwzorowania słownika semantycznego i pliku z dozwolonymi relacjami. W tym momencie wypełniane są struktury: _vWordIdxToString, _mUnambiguousWords, _mAmbiguousWords, _vMultisegmentWordsToAnalyze, _mRelationDescriptionToIdx, _vAllowedRelations, _relationsForward, _relationsBackward. W kolejnym kroku wielosegmentowe słowa zapisywane w strukturze _vMultisegmentWordsToAnalyze i używane są do wypełnienia stuktur _mMultisegmentAmbigConcepts oraz _mMultisegmentUniqueConcepts. W trakcie tego procesu, dodatkowo zapisywane są informacje o możliwych błędach wczytywania modelu.

W skład możliwych błędòw wchodzą:

- deklaracja nie jest wieloznaczna; deklaracja została usunięta z listy wieloznacznych – kod błędu 10001,
- brak opisu wieloznacznego słowa – kod błędu 10002,
- błędna relacja, relacja pominięta – kod błędu 10003,
- definicja słowa wieloznacznego w złej kolejności – kod błędu 10004,
- deklaracja nie jest wieloznaczna, pomimo takiego jej oznaczenia – kod błędu 10005,
- zła kolejność deklaracji; deklaracja występuje przed deklaracją – kod błędu 10006,
- brakująca deklaracja została dodana do modelu – kod błędu 10007,
- w definicji relacji znajduje się niezdefiniowane wcześniej pojęcie – kod błędu 10008,

- podwòjna definicja tej samej krawędzi; drugie wystąpienie zostało zignorowane – kod błędu 10009,
- wieloznaczna deklaracja dodana drugi raz (przyczyna - konwersja do małych liter) – kod błędu 10010,
- deklaracja nadpisuje deklarację; dodane dodatkowe mapowanie – kod błędu 10011,
- jednoznaczne słowo posiada ròwnież wersję wieloznaczną; obie dodane do modelu jako wieloznaczne – kod błędu 10012.

Log zawierający informacje o procesie wczytywania modelu znajduje się w katalogu aplikacji w pliku tekstowym *log.txt*. Każda kolejna linia zawiera pojedyncze wystąpienie błędu wraz z określeniem jego typu, numeru linii, w ktòrej wystąpił i opisu.

Poniżej znajdują się przykładowe linie z błędami wygenerowanymi w trakcie użytkowania programu:

- 10002 WARNING(3332): Brak opisu wieloznacznego słowa: "Bartosiak.1()". Dodany opis o treści: "Brak opisu",
- 10003 WARNING(57459): Błędna relacja: "[0],obiekt fizyczny.1". Relacja pominięta.

Moduł odpowiedzialny za obròbkę zdania wejściowego oparty jest o funkcję biblioteczną strok[18] przy użyciu wykluczającego filtra dla znakòw ~ *`!@#$%^&*()_+{}|:"<>?-=[]\;',./*. Przy jej udziale dokonywany jest podział wejściowego ciągu znakòw na słowa i usunięcie niepotrzebnych lub niedozwolonych znakòw. W kolejnym kroku, przy użyciu biblioteki CLP poszukiwana jest forma podstawowa danego słowa. W wersji programu dla systemu Windows, tworzone jest w tym celu sieciowe połączenie socketowe w oparciu o protokòł TCP/IP.[19] Dedykowany serwer uruchomiony na maszynie uczelnianej zwraca odpowiednią formę dla zapytania, pośredniczący w komunikacji z biblioteką CLP. W wersji pod system UNIX/LINUX forma podstawowa jest uzyskiwana bezpośrednio. W przypadku otrzymania kilku form podstawowych, uwzględniane są wszystkie z nich. Następnie, przy użyciu struktur _mUnambiguousWords, _mAmbiguousWords, _mMultisegmentAmbigConcepts oraz _mMultisegmentUniqueConcepts określone i zapisane strukturach tymczasowych zostają indeksy wierzchołkòw słòw jedno- i wieloznacznych.

Chodzenie po grafie zostało zrealizowane iteracyjną metodą przeszukiwania wszerz, przy użyciu struktur pomocniczych do przechowywania informacji o aktualnie odwiedzonych wierzchołkach i wartościach odległości do poszczegòlnych wierzchołkòw. Ostateczne wyniki zapisywane są w strukturze:

```
class HitCounter {
    public:
    int _travelCost;
    set<PathTVertex> _wordsIndexesWhichCount;
```

[18] Zob. http://www.cplusplus.com/reference/clibrary/cstring/strtok.html.

[19] Zob. http://msdn.microsoft.com/en-us/library/ms740673(VS.85).aspx.

```
};
```

gdzie PathToVertex to:

```
class PathToVertex {
    public:
    vector<CONCEPTIDX> _path;
    TARGETIDX _targetIdx;
};
```

Zmienna _travelCost określa koszt dotarcia do poszczególnego wierzchołka. Zbiór _wordsIndexesWhichCount zawiera informacje o ścieżce (wektor _path) do wierzchołka docelowego i jego indeks (zmienna _targetIdx).

W fazie analizy następuje podział na wyniki cząstkowe ze względu na różne pojęcia wieloznaczne. Do tego celu używana jest struktura pomocnicza FormatResultsStruct.

```
class FormatResultsStruct {
    public:
    int _conceptsCounter;
    int _travelCost;
    float _wholeSum;
};
```

W zmiennej _conceptsCounter zapisana jest ilość wierzchołków w otoczeniu znaczenia, parametr _travelCost określa sumaryczny koszt dotarcia do wszystkich wierzchołków, natomiast _wholeSum zawiera zbiorczą sumę odwrotności kosztów dotarcia dla całego pojęcia. W procesie wypisywania wyników wykorzystywane są struktury _mAmbiguousWords, _mUnambiguousWords i _vWordIdxToString, które pozwalają określić nazwy pojęć dla wierzchołków.

W przypadku uruchomienia programu jako niezależnej aplikacji, program wczytuje z dysku model sieci semantycznej i dozwolone relacje, a następnie dla każdej linii z podanego jako parametr pliku ujednoznacznia znalezione w nim pojęcia wieloznaczne. Wynik działania programu zapisywany jest w pliku wyjściowym. W trakcie działania algorytmu nie są modyfikowane pliki tekstowe modelu ani dostępnych relacji.

W przypadku uruchomienia programu jako webservice, zostaje otwarte połączenie sieciowe przy użyciu biblioteki socketów. Za pomocą tego interfejsu następować może komunikacja ze stroną PHP stanowiącą GUI użytkownika, lub innym, dowolnie wybranym modułem.

W przypadku używania programu jako biblioteki, udostępniony został następujący interfejs:

```
typedef int IDX;
typedef float PROBABILITY;
struct Results {
    vector<pair<IDX,PROBABILITY> _results;
};
struct Desambiguator {
    void init();
    Results getResults(const string & input);
    const string & getConceptName(IDX idx);
    const string & getConceptDescription(IDX idx);
    int getConceptRelativeNumber(IDX idx);
};
```

ROZDZIAŁ V

ANALIZA ZACHOWANIA ALGORYTMU

W języku naturalnym istnieje wiele rodzajòw wypowiedzi. Dla idealnie działającego algorytmu, typ zdania nie powinien wpływać na jakość otrzymanych wynikòw. W celu przetestowania działania programu, wybranych zostało kilka ich typòw, dla ktòrych przygotowano zestawy testowe:

- testy autorskie (pkt 5.1.) – mające pokazać specyfikę działania programu (11 zdań testowych),
- testy na zdaniach mowy potocznej (pkt 5.2.) – mające dostarczyć wynikòw działania dla zdań ubogich w kontekst (6 zdań testowych),
- testy na wyselekcjonowanych fragmentach dzieł literackich (pkt 5.3.) – dzieła literackie stanowią znaczący zasòb ludzkiej wiedzy. Zbiòr zdań z wybranych utworòw literackich prezentuje zachowania dla tej dziedziny (13 wypowiedzi testowych),
- testy na wybranym podzbiorze notatek prasowych PAP zawierających wybrane słowo (pkt 5.4. i 5.5.) – analogicznie jak dla dzieł literackich, przy czym pojedyncza wypowiedź zawiera w sobie terminologię z pojedynczej dziedziny (19 i 181 wypowiedzi testowych).

W celu sprawdzenia działania algorytmu dla szczegòlnych przypadkòw przeprowadzono dodatkowe testy:

- testy na zbiorach słòw sztucznie wygenerowanych ze słownika semantycznego (pkt 5.6.) – mające sprawdzić wyniki działania w oderwaniu od zupełności modelu (450 testowych ciągòw słòw),
- powtòrzenie wybranych testòw dla ograniczonego zbioru relacji (pkt 5.7.) – przeprowadzone w celu sprawdzenia działania algorytmu bez relacji syntagmatycznych (181 zdań testowych).

Pojedynczy test składa się z:

- poddanego analizie zdania,
- zwròconego przez program zbioru znaczeń znalezionych w tekście pojęć wieloznacznych - jeśli znaczenie słowa zostało określone z prawdopodobieństwem 100% (w zadanym

sąsiedztwie nie znaleziono żadnych innych, dopuszczalnych znaczeń) – podawane jest tylko to znaczenie; w przeciwnym wypadku, wskazywane są wszystkie znalezione znaczenia słowa wraz z ich prawdopodobieństwami oraz kluczowymi słowami,

- prawdopodobieństwo każdego z pojęć (jeżeli wskazane zostało więcej niż jedno pojęcie),
- słòw wskazujących dla każdego z wskazanych pojęć.

Poniżej przedstawiony został format przykładowego testu:

ETYKIETA	**WARTOŚĆ**	**ZNACZENIE**
ZDANIE:	*„Zamek w drzwiach zardzewiał.”*	*Zdanie testowe*
ODPOWIEDŹ 1:	*zamek (urządzenie do zamykania)*	*Pierwsza ze zwròconych odpowiedzi*
WAGA 1:	*0. 666667*	*Obliczone przez program prawdopodobieństwo pierwszej odpowiedzi*
SŁOWA KLUCZOWE 1:	*drzwi , zardzewieć*	*Wskazane przez algorytm słowo wskazujące na pierwszą odpowiedź*
ODPOWIEDŹ 2:	*zamek (część broni)*	*Druga ze zwròconych odpowiedzi*
WAGA 2:	*0. 333333*	*Obliczone przez program prawdopodobieństwo drugiej odpowiedzi*
SŁOWA KLUCZOWE 2:	*zardzewieć*	*Wskazane przez algorytm słowo wskazujące na drugą odpowiedź*

Tabela 3. Format przykładowego testu

Większość testòw wykonana została dla następujących pojęć wieloznacznych:

- zamek w znaczeniu:
  - warownej budowli mieszkalnej
  - części broni
  - urządzenia do zamykania
- ślimak rozumiany jako:
  - motyw dekoracyjny
  - składnik ucha wewnętrznego ssakòw
  - element maszynowy w postaci śruby
  - droga w formie linii spiralnej
  - zwierzę z rodziny ślimakòw
- kozak jako:
  - żołnierz lekkiej jazdy
  - śmiały, odważny mężczyzna
  - ludowy taniec ukraiński
  - grzyb jadalny
  - ciepłe, damskie lub dziecięce obuwie zimowe o wysokiej, miękkiej cholewie

Słowa te zostały wybrane, ponieważ są one stosunkowo dobrze opisane w słowniku semantycznym (średnia dla modelu to ok. 2 relacje dla pojęcia, a dla każdego ze znaczeń słowa *zamek* występuje odpowiednio 20, 30 i 60 relacji ich dotyczących), a przy tym są popularne więc często występują w zdaniach języka potocznego.

5.1. TESTY AUTORSKIE

Test został przeprowadzony na uprzednio przygotowanych zdaniach i miał za zadanie zaprezentować działanie algorytmu. Niektòre spośròd przeprowadzonych testòw mają charakter przyrostowy, czego celem było pokazanie wpływu zmian w zdaniu wejściowym na otrzymane wyniki.

5.1.1. ZBIÒR TESTOWY I OTRZYMANE WYNIKI

Poniżej przedstawiony został zestaw testowy wraz z zwròconymi przez system odpowiedziami i wskazanym prawdopodobieństwem ich występowania, oraz słowa kluczowe, ktòre spowodowały ich wybranie:

[1]
ZDANIE: *„Zamek w warszawie był celem wielu atakòw.”*
ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *cel (cel ataku)*

[2]
ZDANIE: *„Kiedy będę wystarczająco bogaty będę mieszkać w zamku.”*
ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *mieszkać (zajmować na stałe jakieś pomieszczenie)*

[3]
ZDANIE: *„Zamek ktòry widzieliśmy był ładny, chociaż zaniedbany.”*
ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *ładny, zaniedbany*

[4]
ZDANIE: *„Zamek się zamyka.”*
ODPOWIEDŹ: *zamek (urządzenie do zamykania)*
SŁOWA KLUCZOWE: *zamykać*

[5]
ZDANIE: *„Kupiony przeze mnie pistolet okazał się mieć niedziałający zamek.”*
ODPOWIEDŹ: *zamek (część broni)*
SŁOWA KLUCZOWE: *pistolet*

[6]
ZDANIE: *„Źle działający zamek spowodował, że ładunek chybił.”*
ODPOWIEDŹ: *zamek (część broni)*
SŁOWA KLUCZOWE: *[działający, działać],ładunek*

[7]
ZDANIE: *„Będąc na Ukrainie zwiedziłam zamek nazywający się Wysokim ZamkIem.”*
ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *wysoki zamek*

[8]
ZDANIE: *„Zamek zardzewiał.”*
ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*
WAGA 1: *0.500000*
SŁOWA KLUCZOWE 1: *zardzewieć*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0.500000*
SŁOWA KLUCZOWE 2: *zardzewieć*

[9]
ZDANIE: *„Zamek w drzwiach zardzewiał.”*
ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*
WAGA 1: *0. 666667*
SŁOWA KLUCZOWE 1: *drzwi , zardzewieć*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0. 333333*
SŁOWA KLUCZOWE 2: *zardzewieć*

[10]
ZDANIE: *„Zamek był metalowy.”*
ODPOWIEDŹ: *(brak wyników)*
SŁOWA KLUCZOWE: *(brak wyników)*

[11]
ZDANIE: *„Zamek był zrobiony z metalu.”*
ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*

WAGA 1: *0. 500000*
SŁOWA KLUCZOWE 1: *metal*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0. 500000*
SŁOWA KLUCZOWE 2: *metal*

5.1.2. ANALIZA

Z rezultatòw przeprowadzonych testòw można wnioskować, że algorytm dobrze radzi sobie ze zdaniami zawierającymi wyrazy wprost wskazujące na konkretne znaczenie wyrazu. Wyrazy: *cel*, *mieszkać*, *ładny*, *zaniedbany*, występujące w testach [1], [2] i [3], jednoznacznie wskazują na *zamek* w rozumieniu *warownej budowli mieszkalnej* i zostało to poprawnie zinterpretowane przez algorytm. Podobnie jest dla zdań [4], [5] i [6] – słowo *zamykać* wskazuje na *zamek* jako *urządzenie do zamykania*, a *zepsuty*, *działający*, *ładunek* i *pistolet* – na *część broni*. Jak widać na przykładzie zdania numer [7], algorytm ròwnie dobrze radzi sobie z występującym w modelu wyrażeniami wielosegmentowymi (tu np. poprawnie zidentyfikowano znaczenie słowa *zamek* występującego razem z nazwą własną *Wysoki Zamek*).

Dodatkowo, test [6] pokazuje działanie algorytmu, gdy na etapie ujednoznaczniania fleksyjnego zwròcono więcej niż jedną poprawną formę wyrazu (w tym wypadku dla słowa *działający* zwròcono *działać* i *działający*). Algorytm wòwczas sprawdza istnienie wszystkich zwròconych form, ale uwzględnia do wyniku tylko jedną – z najniższym kosztem.

Na przykładzie testòw [8] i [9] widać, że wraz ze wzrostem ilości informacji w zdaniu, zwiększa się ilość słòw stanowiących o kontekście całej wypowiedzi, a co za tym idzie – trafność udzielonej odpowiedzi. W zdaniu [8] zwròcono dwie możliwe odpowiedzi – zaròwno *zamek* w rozumieniu *urządzenia do zamykania*, jak i *części broni* może *zardzewieć*. Dostarczenie dodatkowej informacji do zdania (test [9]), czyli jego uszczegòłowienie, powoduje zwiększenie prawdopodobieństwa jednego ze znaczeń.

Problemem uwidocznionym w wynikach otrzymanych dla zdań [10] i [11] jest brak relacji pomiędzy ròżnymi formami pokrewnymi tego samego wyrazu. Pozornie, obejście tego problemu jest stosunkowo łatwe i polega na sprawdzeniu istnienia w modelu każdej możliwej formy rozpatrywanego słowa. Dość prosto można jednak znaleźć kontrprzykłady skuteczności takiego podejścia – po pierwsze, przy braku mechanizmy rozpoznawania wyrażeń wielosegmentowych na poziomie sprowadzania słòw wypowiedzi do formy podstawowej, łatwo o błędy dla nich. Po drugie, przykładowo, złoty kolor przedmiotu nie jest tożsamy z wykonaniem go ze złota, także zastosowanie tego typu uproszczenia w wielu przypadkach mogłoby odbyć się ze stratą dla kontekstu całości. Umieszczenie w słowniku tylko jednej formy z całej rodziny wyrazòw powoduje zatem nadmierne zubożenie słownika. Rozwiązaniem może być uzupełnienie słownika semantycznego o konieczne relacje dla wszystkich poprawnych form wyrazòw, lub wskazanie części mowy zachowujących prawdziwość relacji. Wprowadzenie do słownika informacji o więcej niż jednej formie wyrazu, powoduje konieczność wyszukiwania całej rodziny wyrazòw wokòł danego wierzchołku, co z kolei prowadzi do wzrostu złożoności obliczeniowej procesu przeszukiwania.

5.2. TEKSTY JĘZYKA POTOCZNEGO

Kolejny test został przeprowadzony na zdaniach języka potocznego. Problemem w analizie tego typu wypowiedzi jest szczątkowa forma lub wręcz zupełny brak kontekstu.

5.2.1. ZBIÒR TESTOWY I WYNIKI

Zestaw zdań wchodzących w skład testu przedstawiał się następująco:

[1]
ZDANIE: *„Zepsuł mi się zamek.”*
ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*
WAGA 1: *0.500000*
SŁOWA KLUCZOWE 1: *zepsuć się*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0.500000*
SŁOWA KLUCZOWE 2: *zepsuć się*

[2]
ZDANIE: *„Kupiłem sobie nowy zamek.”*
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[3]
ZDANIE: *„Zamek ktòry kupiłem okazał się słaby.”*
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[4]
ZDANIE: *„Zamek mi zardzewiał.”*
ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*
WAGA 1: *0.500000*
SŁOWA KLUCZOWE 1: *zardzewieć*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0.500000*

SŁOWA KLUCZOWE 2: *zardzewieć*

[5]
ZDANIE: *„Przeżyliśmy prawdziwą plagę ślimaków.”*
ODPOWIEDŹ: *(brak wyników)*
SŁOWA KLUCZOWE: *(brak wyników)*

[6]
ZDANIE: *„Znalazłem wczoraj dużo kozaków.”*
ODPOWIEDŹ 1: *kozak (żołnierz lekkiej jazdy)*
WAGA 1: *0.500000*
SŁOWA KLUCZOWE 1: *kozak (grzyb jadalny)*
ODPOWIEDŹ 2: *kozak (grzyb jadalny)*
WAGA 2: *0.500000*
SŁOWA KLUCZOWE 2: *kozak (żołnierz lekkiej jazdy)*

5.2.2. ANALIZA

Zgodnie z przewidywaniami, wypowiedzi języka potocznego są ubogie w słowa, co czyni kontekst trudnym, a czasem niemożliwym do określenia. Należy jednak zauważyć, że bez znajomości kontekstu, często jest to problem niemożliwy do rozwiązania także dla człowieka. Otrzymane wyniki wskazują również na braki pojęć w modelu i relacji między nimi.

5.3. TEKSTY DZIEŁ LITERACKICH

Test został przeprowadzony na kilku fragmentach zaczerpniętych z tekstòw literackich.

5.3.1. ZBIÒR TESTOWY I WYNIKI

Fragmenty książek wchodzące w zbiòr testowy:

[1]
FRAGMENT: *„Ogniem i mieczem" – H. Sienkiewicz*
Namiestnik jednak nie dostał się tego wieczora do zamku, bo pan Grodzicki zaprowadził taki porządek, że gdy przed zachodem słońca wybito hasło, nie wpuszczano nikogo z zamku i do zamku i gdyby nawet sam kròl przyjechał, musiałby nocować w Słobòdce stojącej pod wałami fortecy.
ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *forteca*

[2]
FRAGMENT: *„Ogniem i mieczem" – H. Sienkiewicz*
Forteca była istotnie nie do zdobycia, bo pròcz armat broniły jej Dnieprowe przepaście i niedostępne skały pionowo zeskakujące w wodę; nie potrzebowała nawet wielkiej załogi. Toteż w zamku nie stało więcej nad sześćset ludzi, ale za to co najprzebrańszego żołnierza, uzbrojonego w muszkiety i samopały. Dniepr, płynąc w tym miejscu ściśniętym korytem, tak był wąski, że rzucona z wałòw strzała przelatywała daleko na drugi brzeg. Działa zamkowe panowały nad oboma brzegami i nad całą okolicą. Pròcz tego o pòł mili od zamku stała wysoka wieża, z ktòrej osiem mil wokoło widać było, a w niej stu żołnierzy, do ktòrych pan Grodzicki każdego dnia zaglądał. Ci, spostrzegłszy w okolicy lud jaki, dawali natychmiast znać do zamku, a wòwczas bito w dzwony i cała załoga wnet stawała pod bronią.
ODPOWIEDŹ 1: *zamek (warowna budowla mieszkalna)*
WAGA 1: *0.750000*
SŁOWA KLUCZOWE 1: *człowiek, okolica, stać*
ODPOWIEDŹ 2: *zamek (część broni)*
WAGA 2: *0.250000*
SŁOWA KLUCZOWE 2: *[nie działać, działać, działo]*

[3]

FRAGMENT: *„Felicitas” – A. Zajdel*
Słysząc trzask zamka w przedpokoju, przymknąłem szybko wieko walizki i odskoczyłem w stronę biurka.
ODPOWIEDŹ: *zamek (urządzenie do zamykania)*
SŁOWA KLUCZOWE: *walizka*

[4]
FRAGMENT: *„Ferma” – A. Zajdel*
To nowe serce... Okazało się, że ono nie jest w ogòle moje. Ono nie było wszczepione, tylko jakby... zainstalowane. Można je było odłączyć w każdej chwili. Cięcia operacyjne też nie zostały zszyte, tylko jakby... jakby zapięte na zamek błyskawiczny, rozumiesz? Skòra, osierdzie wszystko spojone tak, że w każdej chwili można otworzyć, jak kieszeń!
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[5]
FRAGMENT: *„Ferma” – A. Zajdel*
- Nie mogę. Zamykają nas o zmroku, nie mam klucza.
- Ach, tak... Zaczekaj, spròbuję otworzyć. Sven sięgnął do torby, wyjął kilka drobnych narzędzi i przez chwilę manipulował przy zamku. Bez trudu udało mu się sforsować prosty zamek.
ODPOWIEDŹ: *zamek (urządzenie do zamykania)*
SŁOWA KLUCZOWE: *[zamykać, nie zamykać]*

[6]
FRAGMENT: *„Gra w zielone” – A. Zajdel*
W pewnej chwili, odebrawszy telefon, profesor przeprosił mnie i wyszedł zapewniając, że niedługo wròci. Rozejrzałem się po gabinecie. W głębi, w ścianie, zauważyłem skrytkę, jakby kasę pancerną. Była nie domknięta, pęk kluczy zwisał z zamka sejfu.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[7]
FRAGMENT: *„Bajka o kròlu Murdasie” – S. Lem*
Szeregi elektrykarzy i budowniczych jęły wnosić do zamku druty i szpule, a

gdy rozbudowany kròl wypełnił swoją osobą cały pałac, tak że był jednocześnie z frontu, w piwnicach i oficynie, przyszła kolej na stojące w pobliżu domostwa.

ODPOWIEDŹ: *(brak wynikòw)*

SŁOWA KLUCZOWE: *(brak wynikòw)*

[8]

FRAGMENT: *„Czarna kawa" – A. Christie*

- Doskonale, Tredwell. Możesz teraz zamknąć.

- Tak, proszę pana - odpowiedział lokaj, wychodząc. Zamknął za sobą drzwi i wszyscy usłyszeli odgłos klucza przekręcanego w zamku.

ODPOWIEDŹ: *zamek (urządzenie do zamykania)*

SŁOWA KLUCZOWE: *drzwi*

[9]

FRAGMENT: *„Szatański wirus" – A. MacLean*

Odwròciłem się plecami do drzwi, tak długo skrobałem po nich lufą, aż trafiłem na zamek; i pociągnąłem za spust. Po drugim wystrzale drzwi nagle ustąpiły pod moim ciężarem, co może wiele powiedzieć o stanie mojego umysłu, bo nawet nie obejrzałem przedtem zawiasòw, by sprawdzić, w ktòrą stronę się otwierają do środka czy na zewnątrz.

ODPOWIEDŹ 1: *zamek (urządzenie do zamykania)*

WAGA 1: *0. 600000*

SŁOWA KLUCZOWE 1: *drzwi, spust, otwierać*

ODPOWIEDŹ 2: *zamek (część broni)*

WAGA 2: *0.200000*

SŁOWA KLUCZOWE 2: *lufa*

ODPOWIEDŹ 3: *zamek (warowna budowla mieszkalna)*

WAGA 3: *0.200000*

SŁOWA KLUCZOWE 3: *stanie*

[10]

FRAGMENT: *„Szatański wirus" – A. MacLean*

Drzwi na gòrze ròwnież były zamknięte na klucz, ale to nie moje drzwi i miałem jeszcze pięć nabojòw w magazynku. Zamek ustąpił po pierwszym strzale i wytoczyłem się do hallu.

ODPOWIEDŹ 1: *zamek (część broni)*

WAGA 1: *0.500000*

SŁOWA KLUCZOWE 1: *magazynek, nabòj*

ODPOWIEDŹ 2: *zamek (urządzenie do zamykania)*
WAGA 2: *0.500000*
SŁOWA KLUCZOWE 2: *drzwi*

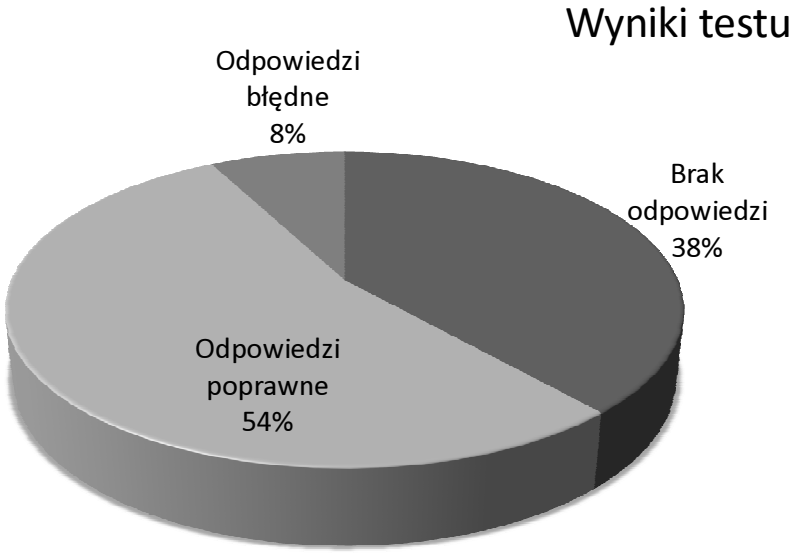
[11]
FRAGMENT: *„Elfia krew" – A. Norton*
- Muszę przynieść swoje narzędzia.
Podniosła się i zniknęła w prywatnych pomieszczeniach, skąd wròciła w oka mgnieniu ze skòrzaną saszetką pełną precyzyjnych narzędzi, ktòrych kobiety-kowale używają przy tworzeniu biżuterii.
- Ten zamek jest bardzo stary i delikatny - powiedziała, sadowiąc się obok Shany, po czym otworzyła saszetkę i wyjęła zestaw do pròbek. - Sądząc po jego wyglądzie, został wykonany przez kobietę. Nie jest może prosty, lecz sama robiłam bardziej skomplikowane.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[12]
FRAGMENT: *„Bòg Imperator Diuny" – F. Herbert*
Doszedł do drzwi swojej kwatery, zbliżył dłoń do zamka zaprogramowanego na jego linie papilarne i zawahał się. Czuł się jak zaszczute zwierzę uciekające do swojego legowiska.
ODPOWIEDŹ: *zamek (urządzenie do zamykania)*
SŁOWA KLUCZOWE: *drzwi*

[13]
FRAGMENT: *„Koniec wiecznosci" – I. Asimov*
Błyskawicznie rzucił się do swoich akt. Gdy palcem jednej ręki przyciskał szyfrowy zamek szafki, druga sięgnęła do teczki. Z biurka wysunął się srebrny język folii, jego perforacja była widoczna nawet gołym okiem.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

5.3.2. ANALIZA

Dla zdecydowanej większości fragmentòw, dla ktòrych program był w stanie udzielić jakąkolwiek odpowiedź – była ona poprawna. Tylko w jednym przypadku rozwiązanie udzielone przez program nie było prawidłowe, wskazano bowiem dwie ròwnie prawdopodobne odpowiedzi. Dużym problemem w działaniu programu jest znaczna liczba przypadkòw, w ktòrych program nie jest w stanie wskazać żadnego rozwiązania. Przyczyną takiego stanu jest brak wiedzy w modelu, w słowniku nie istnieją pojęcia lub relacje występujące w wypowiedziach. Jedynym rozwiązaniem jest uzupełnienie słownika o wymagane zależności.



Rys. 6. Dzieła literackie – wyniki testòw

Świadczy to o brakach w modelu – w słowniku nie istnieją pojęcia lub relacje występujące w wypowiedziach. Jedynym rozwiązaniem jest uzupełnienie słownika o wymagane zależności.

5.4. NOTATKI PAP

Test został przeprowadzony na wyselekcjonowanym podzbiorze notatek prasowych PAP.

5.4.1. ZBIÒR TESTOWY I WYNIKI

Zbiòr testowy stanowiły wybrane notatki prasowe, dotyczące pojęć wieloznacznych:

- *zamek* w znaczeniu:
  - urządzenia do zamykania
  - warownej budowli mieszkalnej
  - części broni
- *ślimak* jako:
  - część ucha wewnętrznego ssakòw
  - zwierzę z rodziny ślimakòw
  - motyw dekoracyjny
  - element maszynowy w postaci śruby
  - droga w formie linii spiralnej
- *kozak* w rozumieniu:
  - żołnierz lekkiej jazdy
  - grzyb jadalny
  - ukraiński taniec ludowy
  - śmiały, odważny mężczyzna
  - ciepłe, damskie lub dziecięce obuwie zimowe o wysokiej, miękkiej cholewie

Poniżej przedstawione zostały notatki testowe wraz z wynikami zwròconymi przez program:

[1]
NOTATKA #007516: *W Centrum Sztuki Współczesnej Zamek Ujazdowski w Warszawie otwarto wystawę prac młodego polskiego artysty Dominika Lejmana, zatytułowaną: "Luksus przetrwania".*
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[2]
NOTATKA #016750:

Pozostałości nieznanego dotychczas średniowiecznego zamku rycerskiego odkryli śląscy archeolodzy we wsi Giebło k. Ogrodzieńca. Na fragmenty starych muròw natrafiono podczas robòt ziemnych na prywatnej posesji. "To prawdziwa rewelacja. Wydawało się, że wszystkie dawne zamki jurajskie są już znane, tymczasem natrafiliśmy na obiekt, o ktòrym nikt nie miał pojęcia" - powiedział PAP Jacek Pierzak, konserwator zabytkòw archeologicznych w Katowicach. Na razie odsłonięto jedynie fragmenty fundamentòw budowli. "Na pewno budynek miał charakter rezydencji rycerskiej, być może była to wieża mieszkalno-obronna. Spodziewamy się znaleźć w pobliżu ròwnież pozostałości zabudowań gospodarczych oraz obronnych, np. wału czy fosy" – oświadczył Pierzak. Archeolodzy ustalają teraz, z jakiego dokładnie okresu pochodzi budowla. Zgromadzony dotychczas materiał archeologiczny – głòwnie ceramika – wskazuje na XIV wiek. Niewykluczone jednak, że zamek postawiono już w wieku XIII. O tym, że zamek był posiadłością rycerską, świadczy m.in. fakt, że w pobliżu znajduje się – jeden z trzech w woj. Śląskim – kościołòw romańskich." Fundatorami takich niewielkich wiejskich kościołòw w tym okresie byli zwykle zamieszkali w pobliżu rycerze, wewnątrz umieszczano dla nich nawet specjalne loże" – wytłumaczył Pierzak. W Jurze Krakowsko-Częstochowskiej znajdują się pozostałości ponad 20 średniowiecznych zamkòw. Najsłynniejsze są ruiny w Ogrodzieńcu, oddalonym od Giebła o kilka kilometròw.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*

SŁOWA KLUCZOWE: *budowla, obronny, rycerski, średniowieczny, zabytek*

[3]
NOTATKA #016166:

Ponad 140 osòb z 27 miast Polski przyjedzie do Krakowa na VI Krajowy Zjazd Bractwa Młodych Miłośnikòw Starych Miast. Spotkanie rozpocznie się w piątek Drużyny z 18 miast będą rywalizować w Ogòlnopolskim Konkursie Historycznym "Wawel – wszystkich Polakòw miejsce najgodniejsze". Jedna trzecia konkursowych pytań ma dotyczyć Biskupstwa Krakowskiego, ktòre w tym roku świętuje jubileusz 1000-lecia, a pozostałe zadania – historii, architektury i zbioròw Zamku Kròlewskiego na Wawelu. W finale zmierzą się tylko trzy drużyny, ktòre wędrując po Wzgòrzu Wawelskim będą szukać odpowiedzi na pytania juroròw. Zwycięzcy konkursu pojadą do Rzymu.

Bractwo Młodych Miłośnikòw Starych Miast powstało w styczniu 1995 r. Jego głòwnym celem jest upowszechnianie wśròd młodzieży wiedzy o zabytkach i historii miast, w ktòrych mieszkają. Co roku w innym miejscu Polski organizowane są zjazdy bractwa. Współorganizatorami zjazdu są: Zamek Kròlewski na Wawelu, Katedra Krakowska, Muzeum Narodowe, Dom Jana Matejki, Muzeum Historyczne Miasta Krakowa, Kopalnia Soli w Wieliczce. Honorowy patronat objął metropolita krakowski ks. kardynał Franciszek Macharski.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *cel, dom, kròlewski, mieszkać (zajmować na stałe jakieś pomieszczenie), muzeum, zabytek*

[4]
NOTATKA #026921:

Gwiazda popu 42-letnia Madonna wyszła w piątek wieczorem za mąż za 32-letniego brytyjskiego reżysera filmowego Guya Ritchie w szkockim zamku Skibo – oświadczyła w sobotę rano miejscowa kobieta-pastor Susan Brown, ktòra udzieliła ślubu. "To się stało" – zapewniła pani Brown dziennikarzy zgromadzonych przed jej domem. Potwierdzenie to ogłoszone zostało po 12 godzinach oczekiwań przedstawicieli mediòw, ktòrzy do soboty rano nie byli pewni, czy w piątek wieczorem ślub Madonny z Ritchie rzeczywiście się odbył.
Ceremonia ślubna zorganizowana została tak, aby nie miało do niej dostępu kilkuset dziennikarzy, ktòrzy przybyli do malowniczej miejscowości Dornoch w pòłnocnej Szkocji, gdzie znajduje się zbudowany w końcu XIX wieku ekskluzywny zamek Skibo. Ślubu udzieliła ta sama pastor-kobieta, ktòra dzień wcześniej ochrzciła czteromiesięcznego Rocco – syna Madonny i Ritchie.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *dom (bliskie sercu miejsce), stać*

[5]
NOTATKA #041103:

Wieczorem, w ramach działań prewencyjnych policjanci wysadzili zamek w drzwiach jednego z samochodòw zaparkowanych koło hotelu Sheraton w Warszawie, w ktòrym zatrzymał się Kasjanow. Policyjny pies sygnalizował, że w aucie może być bomba, ale jej nie było. Wg Aleksieja Wolina, członka delegacji rosyjskiej, Kasjanow ze spokojem odniòsł się do incydentu. "Obecnie też Michaił Kasjanow nie ma zamiaru wprowadzać żadnych zmian do programu swej wizyty w Polsce" – podkreślił Wolin.

ODPOWIEDŹ: *zamek (urządzenie do zamykania)*
SŁOWA KLUCZOWE: *drzwi*

[6]
NOTATKA #041416:

W Ujeździe koło Sandomierza (woj. świętokrzyskie) odbył się dwudniowy V Turniej Rycerski o szablę Krzysztofa Baldwina Ossolińskiego. Jednym z głòwnych punktòw imprezy była rekonstrukcja bitwy o zamek Krzyżtopòr.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *rycerski*

[7]
NOTATKA #044864:

Zamek Kròlewski na Wawelu od wtorku udostępnia zwiedzającym dwie Komnaty na pierwszym piętrze pòłnocnego skrzydła zamku, w ktòrych prezentowana jest kolekcja XVIII-wiecznych mebli, portretòw, dywanòw, porcelany i sreber. W obu salach przez ostatnie 10 lat mieściły się magazyny.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *kròlewski, zwiedzający*

[8]
NOTATKA #049001:

Kętrzyn (woj. warmińsko-mazurskie) najechało w czwartek 200 rycerzy i żołnierzy z kraju i zagranicy. Rozpoczął się tam II Mazurski Turniej Rycerski. Imprezy potrwają do niedzieli, a zakończy je najazd na zamek w Barcianach. Na turniej rycerski do Kętrzyna przyjechali członkowie bractw polskich i Czesi prezentujący wojsko z czasòw cesarstwa austro-węgierskiego za panowania cesarza Franciszka Jòzef.

ODPOWIEDŹ: *zamek (warowna budowla mieszkalna)*
SŁOWA KLUCZOWE: *rycerski*

[9]
NOTATKA #020725:

Setki tysięcy ślimakòw, ktòre wypełzły na płytę lotniska w Nicei sprawiły, że przez pięć godzin jeden z dwòch pasòw startowych nie przyjmował samolotòw. Plagę ślimakòw wykryto podczas rutynowej kontroli pasa. Najprawdopodobniej skorupiaki przedostały się w nocy na betonową nawierzchnię, uciekając przed ulewnym deszczem. Ślimaczy śluz, ktòrym pokryty był trzykilometrowy pas, wykluczał starty i lądowania. Zarząd lotniska zdecydował się na rozwiązanie radykalne – nieproszonych gości pozamiatano, a pas umyto. Do najbliższej składnicy odpadòw odstawiono około pięciu metròw sześciennych ślimakòw.

ODPOWIEDŹ: *ślimak (zwierzę z rodziny ślimakòw)*
SŁOWA KLUCZOWE: *uciekać przed*

[10]
NOTATKA #028749:

Kilkaset eksponatòw przyrodniczych i ekologicznych wysłało do Muzeum Narodowego w Szczecinie dwoje naukowcòw, uczestnikòw wyprawy jachtem "Maria" dookoła świata "Szlakiem Magellana". Wśròd eksponatòw są m.in. zbiory krabòw, ślimakòw, małży, koralowcòw, gąbek, rozgwiazd z wysp na Pacyfiku znalezionych u wybrzeży Chile.

ODPOWIEDŹ: *(brak wynikòw)*

SŁOWA KLUCZOWE: *(brak wynikòw)*

[11]
NOTATKA #030828:

Oddział dziecięcy ośrodka rehabilitacji laryngologicznej Akademii Medycznej otwarto w Poznaniu. Ośrodek opiekuje się m.in. pacjentami z wszczepami implantòw ślimakowych.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[12]
NOTATKA #036992:

Powodzeniem zakończyła się operacja wszczepienia implantu pniowego do mòzgu 51-letniego poznaniaka. W piątek – 24 godziny po zabiegu – pacjent został wybudzony. Poznaniak utracił słuch na skutek niezłośliwego guza na obu nerwach słuchowych. Z tego powodu nie można było pacjentowi zastosować wszczepienia implanta ślimakowego. Operacja polegała na usunięciu niezłośliwych guzkòw i wszczepieniu implantu pniowego z 21 elektrodami do zachyłka bocznego komory w IV pniu mòzgu.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[13]
NOTATKA #037915:

Na Warmii i Mazurach rozpoczął się zbiòr ślimakòw. Zbieranie ślimakòw w wojewòdztwie o największej stopie bezrobocia to obok zbierania poroży czy grzybòw i jagòd jeden ze sposobòw zarobkowania, zwłaszcza dla mieszkańcòw na terenach popegeerowskich.
ODPOWIEDŹ: *(brak wynikòw)*
SŁOWA KLUCZOWE: *(brak wynikòw)*

[14]
NOTATKA #038966:

Rozporządzenie wojewody warmińsko-mazurskiego w sprawie zbioru ślimakòw – winniczkòw obowiązuje w tym roku od 19 kwietnia do 15 maja.

Jak powiedziała PAP Maria Mellin - Wojewòdzki Konserwator Przyrody, ślimaki można zbierać wszędzie z wyjątkiem parkòw krajobrazowych, rezerwatòw oraz strefy przygranicznej. Nie wolno zbliżyć się do granicy na mniej niż 15 metròw. Ślimaki muszą mieć co najmniej 30 mm długości.

ODPOWIEDŹ: *ślimak (zwierzę z rodziny ślimakòw)*

SŁOWA KLUCZOWE: *zbierać*

[15]

NOTATKA #011939:

Pierwszą w Polsce operację wszczepienia drugiego implantu ślimaka ucha wewnętrznego przeprowadzono w Poznaniu. Operacja wszczepienia powiodła się, implant działa. Za miesiąc pacjent odzyska słuch w obu uszach.

ODPOWIEDŹ 1: *ślimak (zwierzę z rodziny ślimakòw)*

WAGA 1: *0.500000*

SŁOWA KLUCZOWE 1: *ucho (narząd słuchu)*

ODPOWIEDŹ 2: *ślimak (składnik ucha wewnętrznego ssakòw)*

WAGA 2: *0.500000*

SŁOWA KLUCZOWE 2: *działać*

[16]

NOTATKA #012846:

W lasach Gòr Świętokrzyskich rozpoczął się – nie notowany zazwyczaj w lipcu – obfity wysyp grzybòw jadalnych: borowikòw, czerwonych kozakòw, maślakòw oraz kani.

ODPOWIEDŹ: *kozak (grzyb jadalny)*

SŁOWA KLUCZOWE: *borowik, czerwony, grzyb jadalny, las*

[17]

NOTATKA #019167:

Jeśli uwielbiamy maślaka żòłtego, szukajmy go wyłącznie pod modrzewiem. Wielbiciele kozaka brunatnego - kierujcie się pod brzozę, a czerwonego - pod osikę. Tajemnicą tej współpracy jest symbioza grzybòw i roślin - mikoryza. Mikoryza to symbiotyczna współżycie korzeni roślin ze strzępkami grzybòw, podczas ktòrej z gleby do rośliny za pośrednictwem grzyba wędrują związki mineralne, a z korzeni do grzybni transportowane są produkty fotosyntezy. Niektòre grzyby wiążą się tylko z jednym drzewem – np. kozak brunatny z brzozą, czerwony – z osiką, a kurka – z sosną. Inne żyją w symbiozie z wieloma gatunkami drzew. I tak olszòwkę, czyli krowiaka podwiniętego, możemy znaleźć pod olchą, sosną, świerkiem i brzozą.

ODPOWIEDŹ 1: *kozak (grzyb jadalny)*

WAGA 1: *0.600000*

SŁOWA KLUCZOWE 1: *czerwony, roślina(pojęcie kategorialne), żyć*

ODPOWIEDŹ 2: *kozak (żołnierz lekkiej jazdy)*

WAGA 2: *0.200000*

SŁOWA KLUCZOWE 2: *żyć*
ODPOWIEDŹ 3: *kozak (śmiały, odważny mężczyzna)*
WAGA 3: *0.200000*
SŁOWA KLUCZOWE 3: *żyć*

[18]
NOTATKA #033110:

Światowej sławy irlandzki spektakl "Spirit of the Dance" będzie mogła obejrzeć polska publiczność podczas jedynego koncertu w Polsce – 22 marca w Spodku w Katowicach. Eklektyczne widowisko, przygotowane przez Irlandzki Międzynarodowy Zespòł Tańca, łączy wiele gatunkòw tanecznych: irlandzkie stepowanie, flamenco, klasyczny balet, taniec ukraińskich kozakòw.

ODPOWIEDŹ 1: *kozak (ludowy taniec ukraiński)*
WAGA 1: *0.666667*
SŁOWA KLUCZOWE 1: *taniec*
ODPOWIEDŹ 2: *kozak (śmiały, odważny mężczyzna)*
WAGA 2: *0.333333*
SŁOWA KLUCZOWE 2: *zespòł*

[19]
NOTATKA #047864:

W piątek w Warszawie złożono akt erekcyjny pod budowę pomnika Tarasa Szewczenki. Stanisław Żurowski, podsekretarz stanu w Ministerstwie Kultury oraz pełnomocnik rządu ds. polskiego dziedzictwa kulturowego za granicą, uznał piątkową uroczystość za wydarzenie "o znaczeniu symbolicznym". Budowę pomnika określił jako "podwaliny pod wzajemne relacje" między Polską a Ukrainą. Pomnik Tarasa Szewczenki, wyrzeźbiony przez artystę z Kijowa, Anatolija Kuszcza, będzie miał 2,8 m wysokości. Stanie na ponad dwumetrowym cokole, zaprojektowanym przez architekta Baltazara Brukalskiego. Na pomniku ma znaleźć się napis: "Podajże rękę Kozakowi i serce swe ku niemu przychyl".

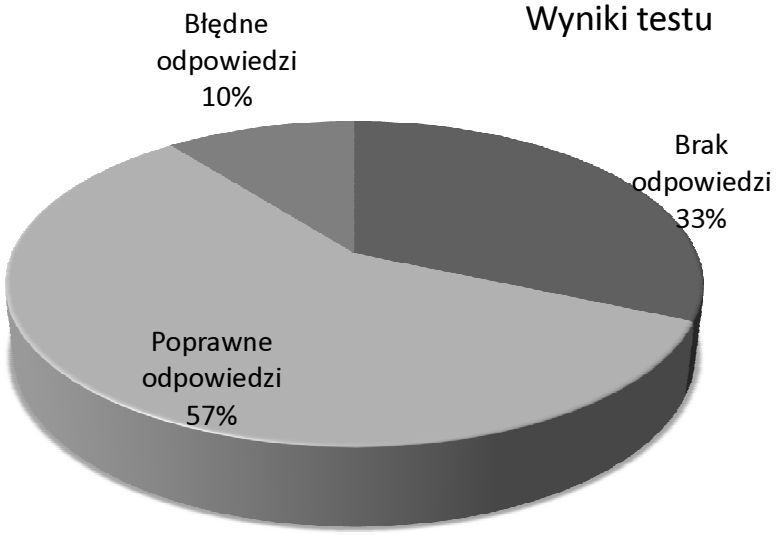
ODPOWIEDŹ 1: *kozak (ludowy taniec ukraiński)*
WAGA 1: *0.400000*
SŁOWA KLUCZOWE 1: *kultura*
ODPOWIEDŹ 2: *kozak (żołnierz lekkiej jazdy)*
WAGA 2: *0.200000*
SŁOWA KLUCZOWE 2: *mieć*
ODPOWIEDŹ 3: *kozak (grzyb jadalny)*
WAGA 3: *0.200000*

SŁOWA KLUCZOWE 3: *mieć*
ODPOWIEDŹ 4: *kozak (śmiały, odważny mężczyzna)*
WAGA 4: *0.200000*
SŁOWA KLUCZOWE 4: *mieć*

5.4.2. ANALIZA

Podobnie jak dla fragmentòw tekstòw literackich, jeśli jakikolwiek wynik został zwròcony – w zdecydowanej większości przypadkòw był on poprawny. Problemem jednak znowu jest duża ilość wypowiedzi, dla ktòrych nie został zwròcony żaden wynik, co potwierdza istnienie sporych brakòw w modelu. Zwròcone dwie błędne wypowiedzi są wynikiem istnienia niepoprawnej relacji w słowniku.

Notatki [11], [12] i [13] potwierdzają istnienie problemu omòwionego w punkcie 5.1.2, a mianowicie braku zdefiniowanych w słowniku semantycznym zależności pomiędzy ròżnymi formami tego samego wyrazu. W tym wypadku, pomiędzy występującym w słowniku słowem *ślimak*, a wyrazami: *ślimakowych* lub *ślimakowego* pochodzących od słowa *ślimakowy* (nieistniejącego w słowniku), oraz *zbierać* a *zbieranie*.



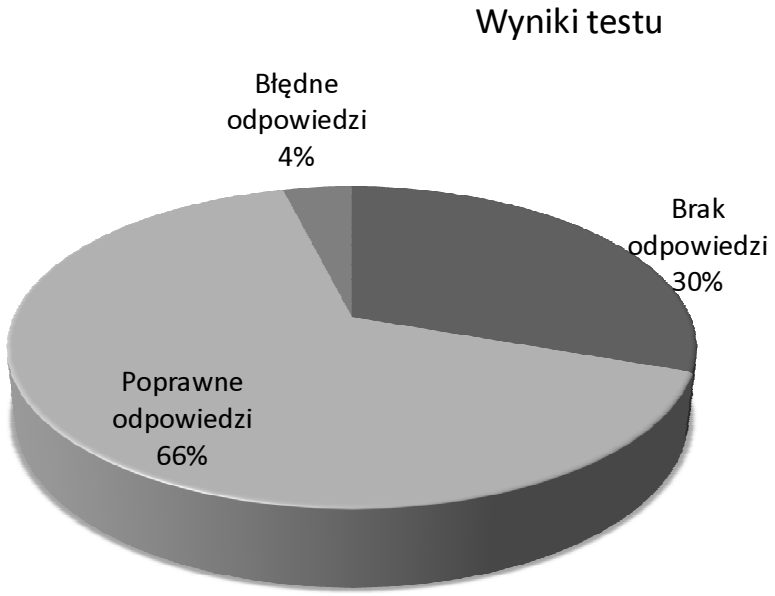
Rys. 7. Wybrane notatki PAP – wyniki testòw

5.5. WSZYSTKIE NOTATKI ZAWIERAJĄCE WYBRANE SŁOWO

Test polegał na sprawdzeniu skuteczności działania algorytmu, dla wszystkich notatek PAP zawierających ktòrąś z form słowa *zamek*. Po przeprowadzeniu testòw, porònano najbardziej prawdopodobne według programu wyniki, z poprawnymi odpowiedziami.

5.5.1. ZBIÒR TESTOWY I WYNIKI

Zbiòr testowy stanowiły wszystkie notatki PAP zawierające słowo *zamek*, czyli 181 notatek.



Rys. 8. Wszystkie notatki PAP – wyniki testòw

5.5.2. ANALIZA

Dla 181 notatek, znaczenia 119 słòw zostały przyporządkowane poprawnie. Spośròd błędnie sklasyfikowanych notatek, dla 7 wskazano błędne odpowiedzi, a dla pozostałych 55 wypowiedzi nie zaproponowano żadnej odpowiedzi, co świadczy o brakach odpowiednich relacji czy pojęć w słowniku semantycznym, lub zbyt dużej objętości testowanych notatek.

Szczególnie często mylone było słowo *zamek* w znaczeniu *warownej budowli mieszkalnej*, z *częścią broni*. Powodowane jest to zbyt obszernym kontekstem. W celu zapewnienia poprawnego działania algorytmu, konieczne byłoby wcześniejsze wykonanie analizy zdań w celu określenia faktycznych granic wypowiedzi.

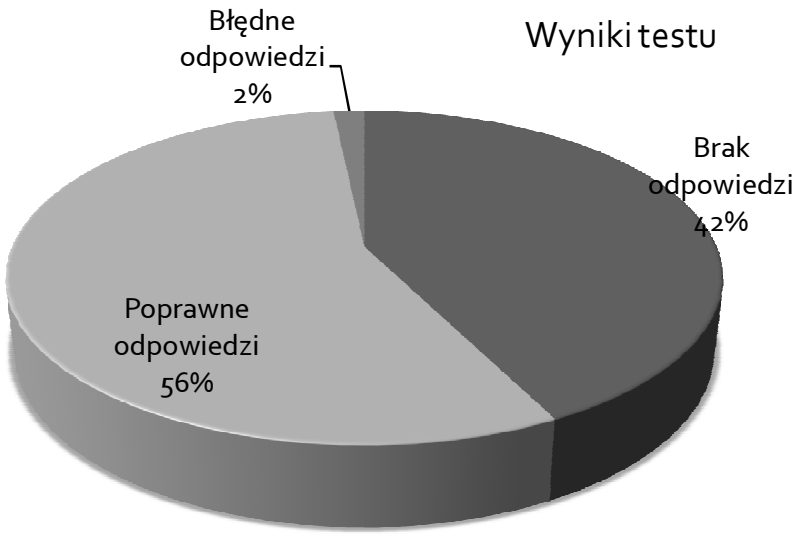
5.6. OGRANICZONY ZBIÒR RELACJI

Test polegał na sprawdzeniu działania algorytmu dla zbioru relacji ograniczonego jedynie do relacji paradygmatycznych (zgodnie z tabelą 1), a następnie poròwnaniu otrzymanych wynikòw, z uzyskanymi w poprzednich testach (czyli dla pełnego zbioru relacji).

5.6.1. ZBIÒR TESTOWY I WYNIKI

Testowi poddano wszystkie notatki PAP zawierające słowo *zamek*, czyli 181 notatek (podobnie jak dla testu nr 5).

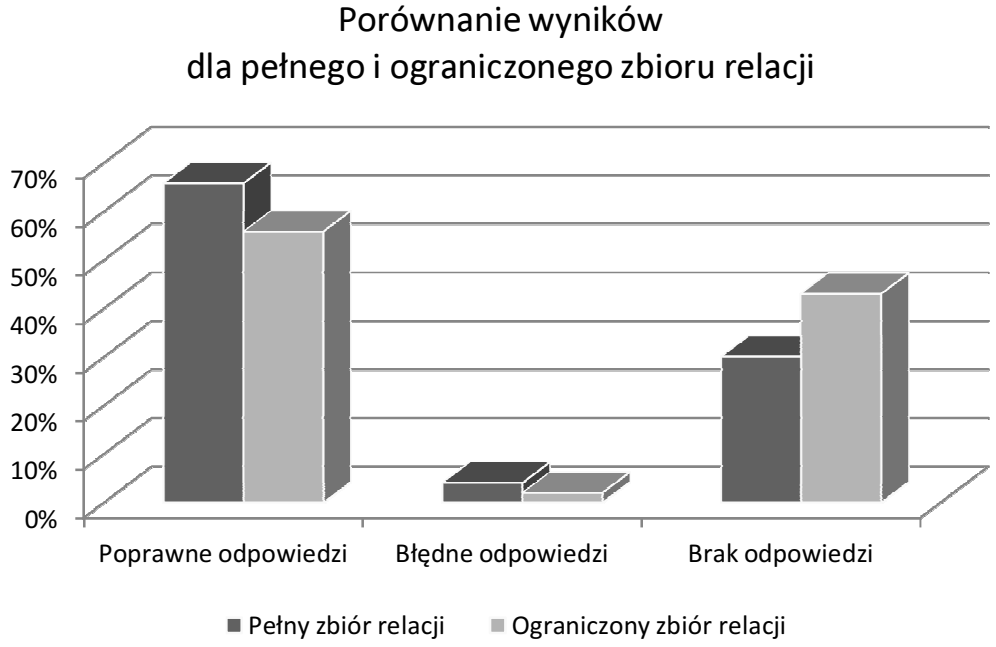
Otrzymane wyniki przedstawiają się następująco:



Rys. 9. Ograniczony zbiòr relacji – wyniki testòw

5.6.2. ANALIZA

Okazało się, że dla większości zdań ograniczenie zbioru relacji jedynie do paradygmatycznych poskutkowało obniżeniem skuteczności działania algorytmu. Znaczenie zostało określone dla 105 z 181 zdań (poprzednio było to 126 znaczeń), czyli 56%. Dla 42%, czyli 76 notatek nie znaleziono rozwiązania w ogòle.



Rys. 10. Ograniczony zbiòr relacji – poròwnanie wynikòw

Pogorszenie skuteczności działania programu spowodowane jest tym, że zależności paradygmatyczne rzadziej występują w zdaniach języka naturalnego. Prawidłowe odpowiedzi mogły zatem zostać udzielone jedynie dla wypowiedzi typu definicyjnego i takich, w ktòrych w jakiś sposòb przemycone zostały treści tego typu.

5.7. SZTUCZNIE WYGENEROWANE CIĄGI SŁÒW

Test został wykonany na zbiorach słòw sztucznie wygenerowanych z bazy słownika semantycznego i miał zweryfikować działanie algorytmu dla istniejących danych, a więc w oderwaniu od problemu kompletności słownika.

5.7.1. ZBIÒR TESTOWY I WYNIKI

Jako zbiòr testowy wybrano zestawy trzech pojęć połączonych ze sobą relacjami. W każdym z zestawòw, przynajmniej jedno słowo było jednym z trzech zawartych w słowniku semantycznym znaczeń słowa *zamek*, czyli:

- urządzenie do zamykania
- część broni
- warowna budowla mieszkalna

Testy przeprowadzono dla 450 spreparowanych zestawòw słòw, jak na przykład:

[1] zamek – IS A PART O – drzwi – IS A KIND OF – obiekt fizyczny [zamek jako urządzenie do zamykania]

[2] broń – IS A – kusza – CONSISTS OF – zamek [zamek jako część broni]

[3] pistolet – CONSISTS OF – zamek – STATE NEGATIVE – zepsuty [zamek jako część broni]

[4] dom – IS A – zamek – CONSISTS OF – budynki mieszkalne [zamek jako warowna budowla mieszkalna]

[5] zamek – SIMILAR TO – twierdza – STATE NEGATIVE – ruiny [zamek jako warowna budowla mieszkalna]

Dla każdego z zestawòw zwracano najbardziej prawdopodobną odpowiedź – istnienie dwòch, ròwnie prawdopodobnych odpowiedzi traktowano jako jej brak.

5.7.2. ANALIZA

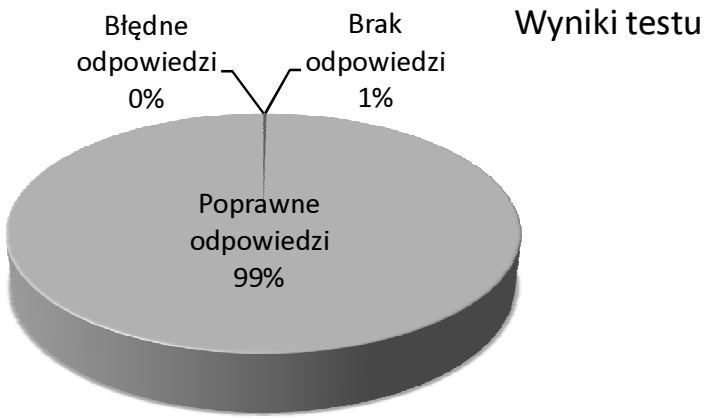
Jak się okazało, jedynie dla 4 z 450 zestawòw słòw nie zwròcono odpowiedzi. Nastąpiło to dla poniższych zestawòw:

[1] Hohenzollern – IS A kind Of – zamek – CATEGORY – obiekt fizyczny [zamek jako warowna budowla mieszkalna]

[2] Hohenzollern – IS A kind Of – zamek – consists of – budynki mieszkalne [zamek jako warowna budowla mieszkalna]

[3] Hohenzollern – IS A kind Of – zamek – consists of – gospodarcze [zamek jako warowna budowla mieszkalna]

[4] Hohenzollern – IS A kind Of – zamek – action negative rt – niechronienie [zamek jako warowna budowla mieszkalna]



Rys. 11. Sztucznie wygenerowane ciągi słòw – wyniki testòw

Dla wszystkich zdań dla ktòrych algorytm nie zwròcił odpowiedzi, pierwsze ze słòw, czyli *Hohenzollern* nie przeszło etapu ujednoznaczniania fleksyjnego, zatem nie było brane pod uwagę w trakcie procesu ujednoznaczniania semantycznego.

Dla testu [1] pozostałe ze słòw, czyli: *obiekt fizyczny*, powiązane było z każdym ze znaczeń słowa *zamek*, dlatego nie można było rozstrzygnąć, o ktòre znaczenie faktycznie chodziło.

W testach [2] i [3] słowa przechowywane w słowniku semantycznym nie są w formie podstawowej, co uniemożliwiło ich zidentyfikowanie.

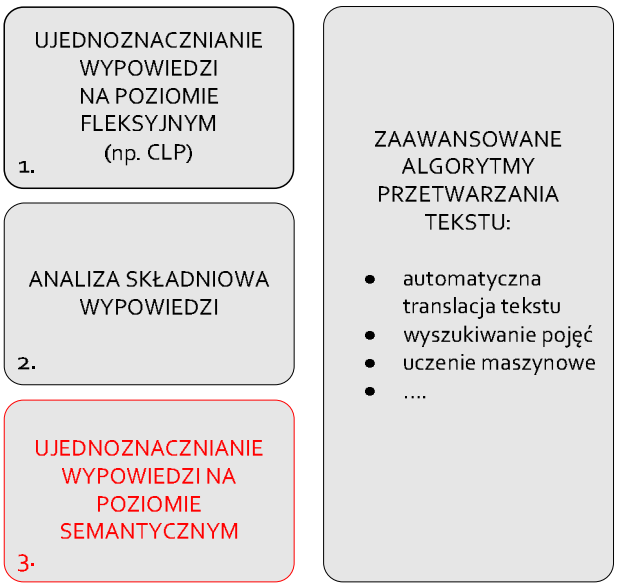
W teście [4] słowo niechronienie nie przeszło etapu ujednoznaczniania fleksyjnego, więc nie mogło pomòc określić właściwego znaczenia wyrazu *zamek*.

Można zatem powiedzieć, że skuteczność algorytmu jest duża i zidentyfikowane błędy wynikają z niekompletności słownika lub pomyłek w nim znalezionych.

ROZDZIAŁ VI

PODSUMOWANIE

Celem pracy było stworzenie i przetestowanie na wybranym zbiorze tekstòw algorytmu rozstrzygającego wieloznaczność wyrazu za pomocą relacyjnego opisu znaczenia. Postawione zadanie zostało zrealizowane. W ramach pracy sprecyzowany został zakres stawianych programowi wymagań. W następnym kroku powstał teoretyczny opis algorytmu, ktòry następnie posłużył jako punkt wyjściowy do implementacji zaproponowanego rozwiązania. Jako bazę wiedzy dla rozwiązania użyto słownika semantycznego stworzonego przez Katedrę Lingwistyki Komputerowej Uniwersytetu Jagiellońskiego, dla ktòrego wykorzystania stworzony został odpowiedni importer danych. W kolejnym kroku został zaproponowany zestaw testòw i procedur ich uruchamiania. Na ich podstawie określono skuteczność algorytmu i jego zachowanie.



Rys. 12. Miejsce opracowanego modułu w algorytmach przetwarzania języka

Opracowany i zaimplementowany algorytm stanowi element pośredni etapu analizy tekstu. Jego istotą jest zmierzenie się z integralną cechą języka naturalnego jakim jest wieloznaczność wyrazu w wypowiedzi. Wykorzystane go wraz z modułami analizy fleksyjnej i składniowej w bardziej złożonych algorytmach przetwarzania tekstu, pozwala na odseparowanie w postaci niezależnego narzędzia informacji o wieloznaczności pojęć w wypowiedzi.

Po analizie odpowiedzi generowanych w odpowiedzi na zdania testowe można zauważyć, że skuteczność działania stworzonego algorytmu jest silnie zależna od analizowanej wypowiedzi (kontekstu, jaki dostarcza) oraz ilości i jakości pojęć zgromadzonych w słowniku semantycznym, które dostarczają informacji o kontekście słòw w zdaniach. Słownik Semantyczny Katedry Lingwistyki Komputerowej UJ znajduje się w fazie tworzenia i w tym momencie zawiera niecałe 52000 pojęć języka polskiego, z ktòrych wiele wymaga poprawy. W poròwnaniu do ogòłu pojęć języka polskiego, jest to stosunkowo niewielka liczba, jednak wystarczająca do przetestowania skuteczności opracowanego algorytmu. Uzyskanie dobrych wynikòw działania już na tym etapie prac napawa optymizmem, należy jednak podkreślić, że ewentualne zastosowanie rozwiązania na większą skalę wymaga pogłębionego wysiłku i nakładu pracy.

Pomimo, że otrzymane wyniki wskazują na wysoką skuteczność algorytmu, nie jest możliwe wykazanie jego skalowalności, czyli ekstrapolacji jego zachowania na kilkukrotnie powiększony model. Wykorzystywany słownik posiada ok. 52000 pojęć i 90000 relacji pomiędzy nimi, co oznacza, że każde pojęcie stanowi węzeł początkowy dla średnio dwòch relacji. Przy tak małej liczbie powiązań, niewielką liczbę stanowią pojęcia łączące się z więcej niż jednym znaczeniem słowa wieloznacznego. Powoduje to, że w większości przypadkòw obecność w wypowiedzi wieloznacznego pojęcia i pojęcia, ktòre z nim się łączy determinuje to znaczenie jako wynik. Powoduje to, że algorytm w większości wypadkòw poszukuje jakiegokolwiek pojęcia mogącego wskazywać znaczenie. Natomiast sytuacja, w ktòrej wybòr znaczenia jest determinowany wagą jaką wnoszą poszczegòlne znalezione pojęcia i wagi relacji między nimi jest rzadko spotykana. Analizując przykłady pojęć opisanych w słowniku można przyjąć, że dla każdego z nich jesteśmy w stanie wskazać szacunkowo, nie mniej niż 100 powiązań z innymi pojęciami. Nie jesteśmy także w stanie wskazać jaką ilość słòw powinien zawierać słownik, aby prawidłowo odwzorowywać rzeczywistość, gdyż nie umiemy wskazać liczby słòw języka polskiego – cytując profesora Jerzego Bralczyka:

„Nie można też stwierdzić, ile słòw jest w całym języku. Mamy wątpliwości, czy przyjąć, że wszystkie słowa zostają w języku na zawsze, czy raczej uważać, że mogą z niego znikać. Nie mamy ustalonych kryteriòw, pozwalających uznać, że oto dane słowo już na stałe weszło do języka. Nie wiemy, jakie słowa powinniśmy uważać za polskie, a jakie są tylko chwilowymi, przypadkowymi gośćmi w polszczyźnie. (...) Nie wiemy, ile jest słòw, bo granica między słowem i połączeniem słòw nie jest wyraźna. Wszystkim wiadomo, że często połączenie dwòch, a czasem i trzech słòw daje nową wartość, zasługującą na osobne miejsce w słowniku." [20]

[20] J.Bralczyk, *Słownik 100 tysięcy potrzebnych słów*, Wydawnictwo Naukowe PWN 2008.

Nawet ograniczone zbliżenie się do takiego stanu modelu doprowadzi do sytuacji, w ktòrej akcent działania algorytmu przesunie się na wybòr znaczenia na podstawie wielu pojęć. Obecne testy wskazują, że w takiej sytuacji o jakości działania algorytmu będzie decydować ilość relacji pomiędzy pojęciami i ich wagi. Zachowanie algorytmu dla takiego przypadku wymaga dalszych prac.

Można wskazać ròwnież inne elementy mogące polepszyć działanie programu. Do najważniejszych można zaliczyć:

- określanie granic wypowiedzi – ograniczenia wymaga rozmiar wypowiedzi, ponieważ w odpowiednio dużym kontekście prawdopodobieństwo pojawienia się dowolnego słowa wzrasta,
- uwzględnienie w ramach słownika i algorytmu informacji o instancjach poszczegòlnych pojęć i powiązanie ich z nimi; pozwoli to przykładowo na przypisanie artykułu o *Zamku Kròlewskim w Warszawie* do pojęcia *zamek* w znaczeniu *budowla* – problemem dla takiego zagadnienia jest ogromna ilość informacji, ktòre wymagają obròbki dla otrzymania zadowalającej bazy wiedzy,
- powiększenie słownika semantycznego doprowadzi do sytuacji, w ktòrej pojawi się konieczność zmian w metryce relacji, sposobu w jaki relacje oddziałują na odległość pojęć między sobą,
- uwzględnienie reguł składni zdania – pozwoli to na etapie analizy na określenie granic kontekstu niezależnych fragmentòw wypowiedzi,
- na etapie pobierania informacji o istniejących pojęciach w wypowiedzi uwzględnienie często występujących w języku zwrotòw, kalek językowych,
- ujednoznacznianie na poziomie leksykalnym poprzez powiązanie pojęć słownika semantycznego z odpowiednimi pojęciami słownika fleksyjnego – wymaga synchronizacji pracy przy tworzeniu słownika fleksyjnego i semantycznego.

BIBLIOGRAFIA

Anusiewicz Janusz, *Lingwistyka kulturowa. Zarys problematyki*, Warszawa 1995.

Bartmiński Jerzy , *Definicja kognitywna jako narzędzie opisu konotacji słowa*, Lublin 1988.

Beckwith Richard , Miller George A., Tengi Randee, *Design and Implementation of the WordNet Lexical Database and Searching Software*, The MIT Press 1998.

Bralczyk Jerzy, *Słownik 100 tysięcy potrzebnych słòw*, Wydawnictwo Naukowe PWN 2008.

Breymann Ulrich, *Designing Components With the C++ STL (2nd Edition)*, Addison Wesley 2002.

Brzozowska Małgorzata, *Etymologia a konotacja wybranych nazw kamieni* w: *Etnolingwistyka. Problemy języka i kultury* nr 12.

Chomsky Noam, *Syntactic Structures*, Mouton, The Hague 1957.

Cormen Thomas H., Leiserson Charles E., Rivest Ronald L., Stein Clifford *Introduction to Algorithms*, MIT Press 2001.

Jurafsky Daniel, Martin James H., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall 2000.

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

Ogden Charles K., *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*, London 1946.

Olinkiewicz Elżbieta, Radzymińska Katarzyna, Styś Halina, *Słownik Encyklopedyczny – Język polski*, Wydawnictwo Europa 1999.

Qouine Willard Van Orman, *Word and Object*, The MIT Press 1960.

Ravin Yael, Leacock Claudia, *Polysemy: Theoretical and Computational Approaches*, Oxford University Press 2000.

Sapir Edward, *Kultura, język, osobowość*, Warszawa 1978.

Wiesław Lubaszewski, Henryk Wròbel, Marek Gajęcki, Barbara Moskal, Alicja Orzechowska, Paweł Pietras, Piotr Pisarek, Teresa Rokicka, *Słownik Fleksyjny Języka Polskiego*, LexisNexis 2001

WordNet a lexical database for the English language: http://wordnet.princeton.edu/

Princeton University Cognitive Science Laboratory: http://cogsci.princeton.edu/

Introduction to Information Extraction Technology – Douglas E. Appelt, David Israel: http://www.ai.sri.com/~appelt/ie-tutorial/

Gramatyka języka polskiego – Grzegorz Jagodziński: http://grzegorj.w.interia.pl/gram/gram00.html

Wstęp do kognitywistyki – Włodzisław Duch: http://www.fizyka.umk.pl/~duch/Wyklady/Cog_plan.html

Word Sense Disambiguation Improves Statistical Machine Translation – Yee Seng Chang, Hwee Tou Ng, David Chiang: http://www.isi.edu/~chiang/papers/chan_wsd_in_mt.pdf

Frequency Asked Questions for Google Suggest: http://labs.google.com/suggestfaq.html

Pozycjonowanie w AdWords: http://adwords.google.pl/support/bin/answer.py?answer=65133&topic=9352

Strona Słownika Fleksyjnego, KI AGH: http://winnie.ics.agh.edu.pl/proj_uk/fleksbaz/

Strona Słownika Semantycznego Języka Polskiego, KI AGH: http://winnie.ics.agh.edu.pl/proj_re/slse/index.html

WordNet: What is to be Done? Hans Patrick, http://www.fi.muni.cz/gwc2004/pres/panel/Hanks/hanks-panel.pdf.

Sockety dla systemu operacyjnego Windows: http://msdn.microsoft.com/en-us/library/ms740673(VS.85).aspx

Opis funkcji bibliotecznej strtok: http://www.cplusplus.com/reference/clibrary/cstring/strtok.html