ZESZYTY NAUKOWE WYDZIAŁU ETI POLITECHNIKI GDAŃSKIEJ
Nr 9 Seria: Wytwarzanie Gier Komputerowych 2010

Jacek Rzeniewicz, Julian Szymański

Politechnika Gdańska, ETI, KASK
julian.szymański@eti.pg.gda.pl

GRA SŁOWNA DO POZYSKIWANIA WIEDZY JĘZYKOWEJ

Streszczenie

W artykule opisano implementację gry słownej w pytania, będącej modelem wyszukiwarki kontekstowej oraz narzędziem do pozyskiwania wiedzy o pojęciach języka naturalnego. Zdefiniowano określenie wyszukiwania kontekstowego oraz przedstawiono opis algorytmu znajdującego obiekty na podstawie ich cech. Scharakteryzowano przyjętą reprezentację wiedzy oraz sposób uczenia się w kontekście innych znanych projektów poruszających problem akwizycji wiedzy.

1. WSTĘP

Wyszukiwanie kontekstowe polega na przeszukiwaniu przestrzeni znaczeń w celu odnalezienia obiektu na podstawie jego cech. Ma ono szczególne znaczenie, gdy szukający nie zna nazwy pojęcia bądź znane są jedynie pewne jego właściwości. Zastosowania takiego rodzaju wyszukiwania mogą być szerokie: od wyszukiwarki internetowej, w której zamiast słów kluczowych użytkownik mógłby posługiwać się opisem cech szczególnych poszukiwanego obiektu, poprzez program dopasowujący profesje do cech charakteru osoby, po narzędzia diagnostyczne wspomagające pracę lekarzy.

Prostym modelem wyszukiwania kontekstowego jest gra w pytania. Jest to gra słowna dla dwóch graczy: pierwszy wybiera obiekt należący do ustalonej dziedziny, drugi zadaje pytania, mające pomóc w identyfikacji pojęcia. Na końcu pytający musi zgadnąć, jakie pojęcie było przedmiotem gry.

Podstawowym celem opisanego tu projektu jest implementacja systemu do gry w pytania[1]. System ten stanowi wygodne środowisko do testów algorytmu wyszukiwania kontekstowego oraz umożliwia interakcję z wieloma użytkownikami. Ponadto, gra demonstruje możliwość modelowania elementarnych kompetencji językowych

[1] http://swn.eti.pg.gda.pl/winston/20q

(formułowanie zdań twierdzących oraz zadawanie prostych pytań) wynikających z posiadania wiedzy leksykalnej.

Warunkiem koniecznym do zbudowania systemu AI operującego językiem naturalnym, jest posiadanie leksykalnej bazy wiedzy zdroworozsądkowej – zawierającej podstawowe fakty na temat świata. W ciągu ostatnich dekad problem ten podjęto w wielu projektach: najważniejsze z nich to: WordNet [1] – słownik pojęciowy w postaci sieci semantycznej tworzony ręcznie, oparte na pozyskiwaniu wiedzy w modelu kooperacyjnym ConceptNet [2] i automatycznym MindNet [3] czy CyC [4], łączący oba te podejścia. Bazy zbudowane automatycznie charakteryzują się niską wiarygodnością. Wykazano [5], że niemal 20% faktów, pozyskanych w sposób automatyczny przez system TEXTRUNNER, było błędnych. Wyniki nowszych badań opartych o weryfikację wiedzy pozyskanej automatycznie przez użytkowników Internetu [6] i [7], wskazują, że takie podejście może znacząco poprawić jakość wiedzy systemu. Z pewnością jest to kierunek, w którym w następnych latach podążać będzie wiele projektów.

Implementacja gry w pytania może być tzw. grą z celem (game with a purpose [8]). Ma ona również duży potencjał jako narzędzie do pozyskiwania wiedzy. W trakcie pojedynczej rozgrywki użytkownik, poprzez udzielanie odpowiedzi, wypowiada sie na temat jakiegoś obiektu, dzięki czemu z każdą grą do systemu napływa nowa wiedza. Dlatego celem równorzędnym do stworzenia wyszukiwarki pojęć, jest budowa narzędzia pozyskiwania wiedzy zdroworozsądkowej. Aktualnie system zdobywa wiedzę jedynie na podstawie gier z użytkownikami. W sekcji 4 opisano plan integracji bazy wiedzy ze słownikiem WordNet, który dzięki wprowadzeniu podejścia kooperacyjnego zyska możliwość dynamicznego rozwoju.

1.1. Reprezentacja wiedzy

Wiedza w systemie reprezentowana jest w sieci semantycznej, opisanej grafem, w której każdy węzeł opisuje znaczenie opatrzone nazwą i krótką definicją. Krawędzie łączące węzły opisują relacje między pojęciowe. Dzięki temu tworzone są trójki wiedzy vwORF [9, 10], reprezentujące fakty podane w formie obiekt–relacja–cecha, przy czym każde pojęcie może występować zarówno po lewej, jak i prawej stronie relacji. Na trójkę wiedzy składają się:

- O – obiekt, którego dotyczy fakt
- R – relacja łącząca obiekt z cecha
- F – cecha obiektu
- v – pewność wiedzy, liczba rzeczywista z zakresu $h0, 1i$
- w – częstość występowania, liczba rzeczywista z zakresu $h-1, 1i$

Parametry v oraz w wprowadzają rozmyty charakter wiedzy; dodatkowo, ich obecność zwiększa ekspresywność trójek. Tablica 1.1 ilustruje możliwe kodowanie przykładowych faktów w notacji vwORF.

Wraz z rozwojem systemu, pozyskiwanych jest coraz więcej informacji dotyczących trójek vwORF. Elementarna jednostka wiedzy na temat faktu (na przykład reprezentująca pojedyncze „klikniecie” gracza) została nazwana przesłanka (*PRM* – premise) i jest określona dwoma parametrami:

- w – częstość występowania, analogicznie do roli w w vwORF
- g – waga

Za każdym razem, gdy w systemie pojawia sie nowa przesłanka, wartości parametrów *v* i *w* odpowiadającej jej trójce wiedzy są obliczane na nowo, by fakt przez nią wyrażany był jak najwłaściwszy w świetle całej posiadanej wiedzy.

Wykorzystanie sieci semantycznej jest alternatywą wobec przechowywania informacji w macierzy *obiekty / cechy*, jak w projekcie 20q.net[2] [11]. Sprzyja ono prowadzeniu wnioskowania na podstawie dostępnej wiedzy, umożliwia łatwy dostęp do grup powiązanych pojęć oraz graficzną eksplorację wiedzy. Ze względu na brak bezpośredniej reprezentacji niektórych informacji (na przykład tam, gdzie cechy są dziedziczone na podstawie relacji IS-A) jest niepraktyczne w trakcie procesu wyszukiwania. Dlatego przeprowadzenie każdej gry rozpoczyna sie od konwersji wiedzy zawartej w sieci semantycznej do formy przestrzeni wektorowej, w której każde znaczenie reprezentowane jest poprzez wektor cech [9, 10]. Narzędzie do przeglądania sieci semantycznej powstającej w trakcie gier dostępna jest w Internecie pod adresem http://swn.eti.pg.gda.pl/winston/network.

Tablica 1.1

Kodowanie faktów trójkami vwORF

Fakt	O	R	F	v	w
Koty najczęściej mają ogony	kot	ma	ogon	0.95	0.85
Trawa nigdy nie jest niebieska	trawa	jest	niebieski	0.98	-0.98
Zwierzęta czasami jedzą inne zwierzęta	zwierzę	je	zwierzę	0.85	0.3

1.2. Przebieg gry

Gra rozpoczyna sie od zadania przez system dwudziestu pytań (lub mniej, o ile jest to wystarczające) o cechy obiektu, który gracz ma na myśli. Na każde z pytań użytkownik może odpowiedzieć: „tak”, „nie” lub „nie wiem”; dostępne są również odpowiedzi pośrednie: „raczej tak”, „raczej nie”, „czasami”, „rzadko”.

Po zakończeniu etapu pytań i odpowiedzi, system „zgaduje”, jaki obiekt był opisywany przez użytkownika. Odbywa się to poprzez obliczenie podobieństwa pojęcia opisanego w grze do znanych już pojęć i wybranie tego, dla którego to podobieństwo jest największe. Użytkownik weryfikuje odpowiedź systemu i przekazuje nazwę opisywanego obiektu, o ile rezultat jest niepoprawny. Ten etap wiąże się z koniecznością odwzorowania literału podanego przez użytkownika, na pojęcie zapisane w bazie wiedzy systemu (o ile takie pojęcie jest już znane) bądź rozszerzenia bazy o nowe znaczenie. W tym etapie nie można działać tylko na poziomie nazw pojęć, gdyż wpisany literał może być homonimem znanego juz obiektu, jak również może zostać podany błędnie np. mogą wystąpić błędy literowe. Dlatego do poprawnego zidentyfikowania opisanego znaczenia wykorzystywane są krótkie definicje pojęć.

Ostatnią fazą jest próba uzyskania od użytkownika, w procesie prostych dialogów, dodatkowej wiedzy na temat przedmiotu gry. Gracz proszony jest o podanie jak najbardziej

[2] http://20q.net/

dystynktywnego faktu dotyczącego opisanego pojęcia. Aby wyeliminować konieczność parsowania zdań w języku naturalnym oraz wymusić na graczy użycie formy nadającej się do zapisu w notacji ORF, wymaga się w tej fazie wypełnienia dwóch pól - jednego przeznaczonego na relację, a drugiego na cechę.

2. ALGORYTM WYSZUKIWANIA

Proces wyszukiwania oparty jest na kolejnych iteracjach cyklu pytanie-odpowiedź, w wyniku których powstaje wektor odpowiedzi ANSW opisujący wyszukiwane pojęcie. W każdym kroku obliczane zostaje podobieństwo ANSW do każdego z wektorów reprezentujących znane juz pojęcia. Na tej podstawie, dokonywany jest wybór pytania dla następnego kroku. Po wykonaniu ostatniej iteracji uznaje się, że pojęcie najbardziej podobne do znaczenia określonego przez odpowiedzi użytkownika reprezentuje wyszukiwany obiekt.

2.1. Wektory cech

Ze względu na możliwość niejawnego przechowywania informacji w sieci semantycznej, do wyszukiwania wykorzystuje się wektorową reprezentację znaczeń. Wektor cech (CDV – Concept Description Vector [12]) danego obiektu przechowuje liczbowe wartości określające, w jaki sposób dana cecha sie do niego stosuje. Element *CDV* [*f*] wektora *CDV* oznacza iloczyn $v \cdot w$, gdzie *v* i *w* to pewność wiedzy oraz częstość występowania cechy, zgodnie z definicją z sekcji 1.1.

2.2. Ocena podobieństwa wektorów

W każdej iteracji algorytmu, konieczne jest obliczenie podobieństwa między wektorem ANSW a każdym znanym CDV. Pierwsze przetestowane podejście oparte było na wyznaczaniu odległości pomiędzy punktami w przestrzeni wielowymiarowej, wskazywanymi przez wektory zgodnie z miarą Euklidesowa. To podejście jednak nie sprawdziło się, ze względu na silną wrażliwość na różnice długości porównywanych wektorów. Nowo nabyte przez system pojęcia, charakteryzujące sie niskimi wartościami parametru pewności wiedzy, geometrycznie umiejscowione były daleko od znaczenia definiowanego przez gracza nawet wtedy, gdy kierunki wskazywane przez oba wektory pokrywały się. Dlatego zaimplementowano wzór oparty na mierze cosinusowej, reprezentującej cosinus kąt miedzy wektorami:

$$\cos(CDV, ANSW) = \frac{\sum_i CDV[i] \cdot ANSW[i]}{\sqrt{\sum_i CDV[i]} \cdot \sqrt{\sum_i ANSW[i]}} \tag{2.1}$$

Ponieważ zbiorem wartości funkcji cosinus jest przedział <-1, 1>, podobieństwo *s* między wektorami CDV i ANSW uzyskiwane jest przez odwzorowanie wartości ujemnych na zero:

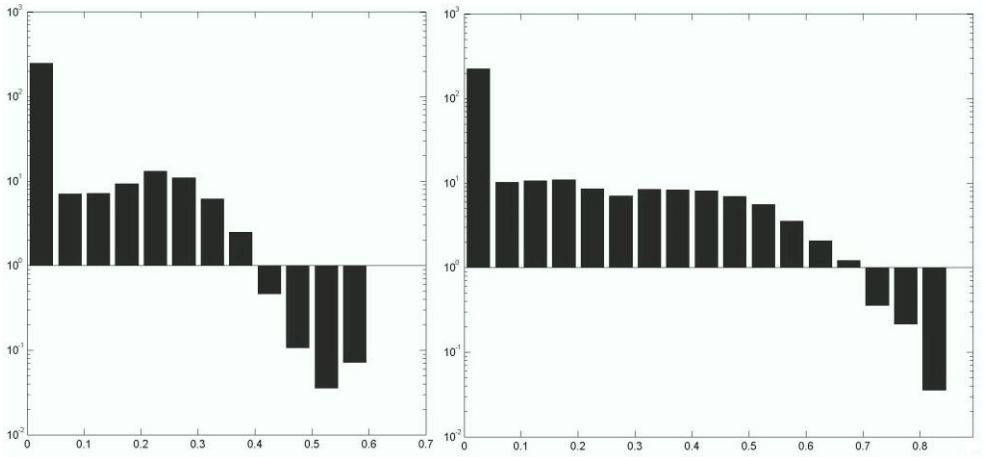
$$s(CDV, ANSW) = \mathrm{clamp}_{\langle 0,1 \rangle}(\cos(CDV, ANSW)) \tag{2.2}$$

gdzie: $\mathrm{clamp}_{<a,b>}(v)$ – funkcja obcinająca wartości *v* do zbioru <a,b>

Obcinanie wartości ujemnych jest praktyczniejsze, niż liniowe odwzorowanie z < −1, 1> na <0, 1>. Gdy cosinus kata jest ujemny, kierunki wskazywane przez wektory są na tyle odległe, że dokładna wartość kąta miedzy nimi nie ma znaczenia; jednocześnie wartości podobieństwa między dobrze pasującymi wektorami nie zostają zacieśnione do małego przedziału.

2.3. Wybór pytań

Ponieważ w bazie wiedzy systemu występują w luki, informacje błędne, jak również użytkownicy mogą udzielać różnych odpowiedzi, dlatego pojęcia nie mogą być nieodwracalnie eliminowane z procesu wyszukiwania. To powoduje, że w ogólnym przypadku nie można skutecznie przeprowadzić wyszukiwania binarnego poprzez zawężanie zbioru obiektów-kandydatów w każdym kroku. W związku z tym, każda iteracja algorytmu rozpoczyna się od wyodrębnienia ze zbioru wszystkich wektorów podprzestrzeni *O*(*A*) zawierającej jedynie znaczenia o realnych szansach na bycie poszukiwanym obiektem, tak, by potem operować tylko w obrębie tej podprzestrzeni.



Rys. 1. Rozkład liczności wektorów w zależności od podobieństwa do ANSW.

Klasyfikacja wektorów do *O*(*A*) jest bardzo istotnym etapem procesu wyszukiwania: jeśli *O*(*A*) nie zawiera szukanego pojęcia, bardzo prawdopodobne jest, że pytanie o najbardziej dystynktywne cechy obiektu w ogóle nie zostanie zadane, przez co cały proces zakończy sie niepowodzeniem. Z drugiej strony, gdy podprzestrzeń *O*(*A*) jest zbyt liczna, mało pasujące wektory powodują szum, skutecznie uniemożliwiający zadanie istotnych pytań. Dlatego przy decyzji, czy dany wektor CDV powinien być włączony do *O*(*A*), dobrze sprawdzają się metody oparte na prawdopodobieństwie. Nie gwarantują, że poszukiwany koncept będzie obecny w rozpatrywanej podprzestrzeni podczas wszystkich dwudziestu kroków, ale nie eliminują go trwale, zachowując ograniczone rozmiary *O*(*A*); jednocześnie zwiększają szanse konceptów, które z różnych względów nie znajdują się w czołówce kandydatów. Prawdopodobieństwo włączenia wektora *CDV* do *O*(*A*) wyznaczane jest przez funkcję opartą na regułach logiki rozmytej:

$$p(CDV) = f(s, s_{\max}, k) \tag{2.3}$$

gdzie: *s* – funkcja obcinająca wartości *v* do zbioru <a,b>

s_{max} – podobieństwo najbardziej pasującego wektora do ANSW w bieżącym kroku
k – nr kroku
f – funkcja rozmyta o zbiorze wartości <0, 1>

Wadą podejścia opartego na prawdopodobieństwie jest „przepuszczanie" do $O(A)$ wektorów o bardzo niskim podobieństwie do ANSW, co ma negatywny wpływ na zachowanie wyszukiwarki, a wynika z dużej liczności grupy najmniej pasujących CDV.

Rysunek 1 prezentuje histogramy ilustrujące średni[3] rozkład obiektów pogrupowanych według podobieństwa do ANSW, odpowiednio po dwóch i osiemnastu iteracjach. Zgodnie z tymi histogramami, wprowadzenie progu odcięcia t_b dla podobieństwa mniejszego od 0.05, spowoduje usuniecie bardzo licznej grupy wektorów. Testy wykazują, że gdy szukany wektor pojawia sie poniżej progu t_b – co bywa możliwe w kilku pierwszych krokach – zawsze wraca ponad ten próg po 2-3 pytaniach, mimo że nie ma go w podprzestrzeni $O(A)$. Testy wykazały również, że w późniejszych iteracjach wartość t_b może być bezpiecznie podniesiona do następujących poziomów:

$$t_b(k) = \begin{cases} 0 & k = 1 \\ 0.05 & k \in \{2,\ldots,9\} \\ 0.1 & k \in \{10,\ldots,18\} \\ 0.15 & k \in \{19,20\} \end{cases} \tag{2.4}$$

Drugą parametrem mającym wpływ na włączenie CDV do $O(A)$ jest próg pewności t_t:

$$t_t = s_{\max} - 0.1 \tag{2.5}$$

zapewniający, że najbardziej pasujące wektory będą przyjmowane do $O(A)$ z prawdopodobieństwem 1.

Po utworzeniu podprzestrzeni *O(A)*, kolejnym krokiem jest znalezienie cechy jak najbardziej separującej ten zbiór – tak, by wyszukiwanie w $O(A)$ zmierzało do dwupodziału. W [9] zaproponowano wykorzystanie wzoru Shannona na zysk informacyjny (*IG* – Information Gain):

$$IG(f) = -\sum_{i=1}^{M} p(o_{if}) \cdot \log p(o_{if}) \tag{2.6}$$

gdzie:

$$p(o_{if}) = \frac{|w_{if}|}{M} \tag{2.7}$$

gdzie: M – liczba wektorów w *O(A)*
w_{if} – wartość cechy f i-tego wektora w $O(A)$

Ponieważ jednak entropia Shannona nie bierze pod uwagę znaku wartości cechy, lepsze rezultaty przynosi wzór (2.8):

[3] Próba 30 gier na dwóch różnych bazach 200- i 500-konceptów.

$$IG_A(f) = \min(p,n) + \frac{|p-n|}{M+1} \tag{2.8}$$

gdzie: p – liczba wektorów CDV spełniających kryterium $CDV[f] > 0.1$
n – liczba wektorów CDV spełniających kryterium $CDV[f] < -0.1$

Szacowanie ilości informacji związanej z zadaniem określonego pytania poprzez zliczanie wartości dodatnich i ujemnych w kolumnie, stosowane jest również w algorytmie 20q.net [11]. Tablica 2.1 zawiera wartości IG_A obliczone dla kilku cech w małym zbiorze wektorów. Wadą wzoru (2.8) jest ignorowanie podobieństwa do ANSW wektorów zawierających znaczące wartości rozpatrywanej cechy. Miara IG_B (2.9) jest modyfikacją tego wzoru i traktuje podobieństwa jako wagi:

$$IG_B(f) = \min(w(p), w(n)) + \varepsilon \cdot \frac{|w(p) - w(n)|}{W} \tag{2.9}$$

gdzie: $w(p)$ – suma podobieństw s(CDV, ANSW) wektorów CDV spełniających $CDV[f] > 0.1$
$w(n)$ – suma podobieństw s(CDV, ANSW) wektorów CDV spełniających CDV [f] < -0.1
W - łączna suma podobieństw s(CDV, ANSW)wszystkich wektorów z O(A)

Przykładowe oceny pytań według miary IG_B znajdują się w ostatnim wierszu tabeli 2.1. Poprawę jakości oceny cech wynikającą z wprowadzenia wzoru (2.9) dobrze demonstruje zmiana względnej informatywności pytania z kolumny 3 tej tabeli. Miara IG_A nie bierze pod uwagę położenia wiersza, w którym występuje wartość, dlatego wysoko oceniła pytanie o bycie drapieżnikiem. Obecna implementacja systemu wykorzystuje obie te miary, w zależności od sytuacji. Jeśli IG_B ocenia wszystkie pytania na 0, wykorzystywany jest wzór IG_A.

Tablica 2.1 Ocena informatywności pytań wg miar IG_A oraz IG_B

cecha / obiekt	1 zwierzę stadne	2 zwierzę domowe	3 drapie-żnik	4 szcze-ka	5 miau-czy	6 ma ogon	7 lata	podob. do ANSW
pies	0.8	0.7	0.6	0.9	-0.8	0.4	-0.8	0.7
kot	-0.7	0.6	0.6	-0.7	0.6	0.5	-0.4	0.6
lis	-0.7	-0.9	0.8	-0.5	0	0.7	0	0.4
owca	0.9	-0.5	-0.6	0	-0.4	0.3	-0.2	0.3
IG_A	2	2	1.4	1.2	1.2	0.8	0.6	
IG_B	1	0.7	0.3	0.7	0.6	0	0	

Taktyka zadawania w każdym kroku pytania o najbardziej informatywną cechę *O(A)* jest wzbogacona poprawką, polegającą na weryfikacji najbardziej pasującego konceptu w sytuacji, w której zdecydowanie „odstaje" on od grupy kandydatów. Jeśli najbardziej pasujący obiekt reprezentuje znaczenie poszukiwane przez gracza, zadawanie pytań o jego dystynktywne cechy szybko potwierdzi ten fakt. Z drugiej strony, jeśli obiekt ten „niesłusznie" ma wysokie podobieństwo do ANSW, wybieranie pytań o jego szczególne

cechy powinno obniżyć podobieństwo i wpłynąć pozytywnie na przebieg szukania – kwalifikacja wektorów do *O(A)* zależy od podobieństwa najbardziej pasującego obiektu do ANSW, przez co utrzymywanie sie niewłaściwego obiektu na szczycie listy jest zjawiskiem niekorzystnym.

Drugim rozszerzeniem jest zawężanie rozpatrywanej podprzestrzeni, w momencie, gdy dużo wektorów charakteryzuje się podobieństwem do ANSW bliskim s_{max}, tzn. gdy *O(A)* jest gęsta u góry. Wtedy korzystny okazuje się wybór pytania maksymalnie różnicującego te najbardziej pasujące pojęcia. Do wykrycia okoliczności, w których warto stosować jedną z tych strategii wykorzystywana jest klasteryzacja jednowymiarowa.

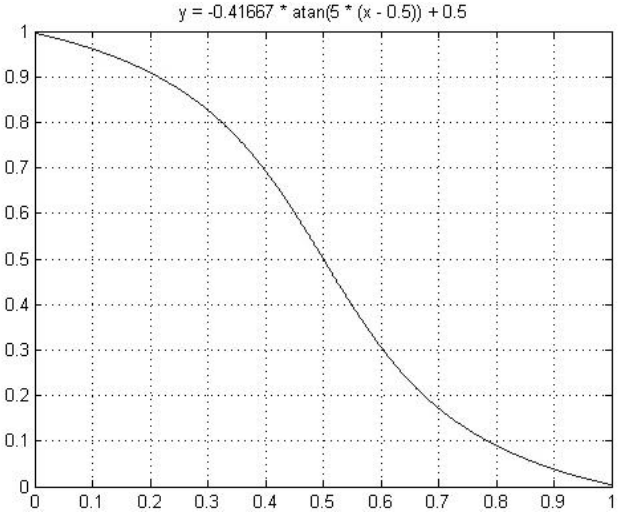
2.4. Modyfikacja sieci semantycznej

Po udzieleniu odpowiedzi na ostatnie pytanie, gracz weryfikuje poprawność wyniku wyszukiwania, w razie niepowodzenia wskazując systemowi opisany obiekt. Trójki wiedzy wiążące ten obiekt z cechami, o które pytano, zostają zaktualizowane bądź utworzone. Dla każdej z tych trójek zachodzi konieczność obliczenia parametrów *v* i *w* pewności wiedzy i częstości występowania cechy, na podstawie wszystkich znanych *PRM*. Częstość występowania *w* obliczana jest jako średnia ważona parametrów *w* wszystkich *PRM* dotyczących danej trójki, natomiast pewność wiedzy jest funkcją łącznej wagi (`abc` – *abundance*) oraz spójności (`coh` – *coherence*) przesłanek:

$$\begin{aligned} \mathrm{abc} &= \mathrm{clamp}_{\langle 0,1\rangle}\left(\frac{1}{5}\sum_i g\left(PRM_i\right)\right) \\ \mathrm{coh} &= -0.41667\cdot\arctan\left(5\cdot\left(stdev-0.5\right)\right)+0.5 \\ v &= \mathrm{abc}\cdot\mathrm{coh} \end{aligned} \tag{2.10}$$

gdzie: $g(PRM_i)$ – waga i-tej przesłanki
stdev – ważone odchylenie standardowe parametrów *w* przesłanek

Rysunek 2 ilustruje zależność spójności przesłanek od odchylenia standardowego parametru częstości występowania przez nie wyrażonego.



Rys. 2. Spójność przesłanek w zależności od odchylenia standardowego parametru *w*.

2.5. Test algorytmu

Przeprowadzono test składający sie z dwóch faz:

1. utworzenie bazy 50 pojęć poprzez wykonywanie kolejnych gier do czasu, aż system poprawnie rozpoznaje każdy z konceptów
2. dodanie do bazy 20 nowych pojęć. Ta faza polegała na wykonywaniu gier wybierając na przemian obiekty jeszcze nieznane oraz istniejące już w bazie, tak by jednocześnie uczyć i testować stabilność wiedzy.

Wszystkie obiekty w przeprowadzonym teście należały do jednej dziedziny - zwierząt. Do zakończenia pierwszej fazy wystarczyły trzy iteracje (w trzeciej wynik był błędny tylko trzy razy), a drugiej tylko dwie. W pierwszej iteracji drugiej fazy wystąpiła jedna pomyłka: system wybrał błędnie nowo poznane zamiast „starego". W drugiej iteracji każda gra zakończyła się wynikiem poprawnym.

W drugiej fazie testu, podczas ponownego wyszukiwania wprowadzonych wcześniej obiektów, algorytm obierał bardzo podobną do pierwotnej ścieżkę zadawania pytań. W każdej z gier więcej niż połowa pytań z pierwszej iteracji powtórzyła się. Dodatkowo, w fazie pierwszej wprowadzono do systemu dużą liczbę cech dobrze opisujących wyszukiwane obiekty. Te czynniki umożliwiły poprawne rozpoznanie obiektów już w drugich grach. Wskazuje to na dobre zdolności przyswajania nowej wiedzy przez system.

Dla kontrastu, w pierwszej fazie druga iteracja niemal w całości składała się z wyników błędnych. Był to efekt braku wiedzy na temat elementarnych cech obiektów – większość gier z pierwszej iteracji kończyła się po kilku pytaniach. Zatem w fazie pierwszej kształtował się dopiero zbiór najważniejszych cech. W drugiej był już on stabilny i pozwolił efektywnie wydobywać najważniejsze informacje na temat obiektów.

3. Blog Winstona Harrisa

Równolegle do rozwijania algorytmu wyszukiwania kontekstowego tworzona jest strona internetowa[4] awatara nazwanego Winston Harris, na której umieszczono grę. Celem tworzenia tego bloga jest pozyskanie zainteresowania użytkowników Internetu grą, poprzez umieszczenie jej w interesującym kontekście.

Intensywny wzrost popularności portali społecznościowych w ostatnich latach otworzył nowe możliwości docierania do graczy. Automatyczna interakcja z użytkownikami poprzez serwisy społecznościowe, daje możliwość ciągłej obecności awatara w Sieci oraz może czynić projekt bardziej atrakcyjnym.

Na drzewie rosnącym w jednym z brukselskich parków zamontowano mierniki CO_2 i pH, wiatromierz, kamerę, termometr oraz inne urządzenia, których zadaniem jest zbieranie danych na temat drzewa i jego otoczenia. Na ich podstawie generowane są wiadomości publikowane na profilach rośliny na portalach Facebook i Tweeter, jak np.: „it's been pretty dry lately. Keeping branches crossed for some rain". Jest to projekt Talking Tree[5] - kampania stworzona przez agencję reklamową Happiness Brussels, udowadniająca, że wirtualny twór może uczestniczyć w internetowej przestrzeni społecznej na takich samych zasadach jak ludzie, skutecznie naśladując ich zachowania. Stanowi on inspirację dla społecznościowych aspektów opisywanego tu projektu.

W chwili obecnej trwają prace nad integracją bloga z portalem Facebook. Aktualnie możliwe jest śledzenie dokładnej historii swoich gier dzięki mechanizmowi uwierzytelniania dostarczonemu przez Facebooka.

4. Dyskusja i dalsze plany

Przeprowadzony test opisanego tu algorytmu został wykonany w obrębie wąskiej dziedziny i na niewielkim zbiorze obiektów. Poza tym, po zakończeniu każdej gry dodawany był wyraźnie deskryptywny fakt na temat przedmiotu gry, a test przeprowadzono jednoosobowo, przez co zaznaczane odpowiedzi nie były ze sobą sprzeczne (choć nierzadko sie różniły). Takie warunki będą nieosiągalne w docelowym środowisku, kiedy odpowiedzi udzielać będzie liczna grupa osób nieposiadających wiedzy eksperckiej przez co ich odpowiedzi dotyczące tego samego obiektu mogą się różnić. Dlatego przeprowadzenie testu w większej skali, jest kolejnym, niezbędnym krokiem w pracy nad projektem.

Testy prowadzone na bazach danych bez ograniczenia domeny wskazują, że algorytm jest znacznie mniej efektywny w takich warunkach. Ma to miejsce dlatego, że dla obiektów należących do innych dziedzin, cechy mające sens w domenie, do której należy szukane pojęcie, są niezdefiniowane. Uniemożliwia to poprawny wybór pytania. Przedstawione w [11] podejście przypisywania wag odpowiedziom w zależności od numeru kroku (początkowe pytania maja wysokie wagi) powinno kierować wyszukiwarkę w odpowiednie obszary sieci. Alternatywny sposób rozwiązania tego problemu to wydzielenie w sieci semantycznej osobnych mikroteorii i oznaczenie ich cechami najbardziej odróżniającymi od innych grup. Pierwsze pytania gry służyłyby wybraniu właściwej poddziedziny, tak, by później wyszukiwanie mogło odbywać się jedynie w jej obszarze. To podejście ma również tę zaletę, że zmniejsza rozmiar danych, na jakich operuje algorytm.

[4] http://swn.eti.pg.gda.pl/winston/
[5] http://www.talking-tree.com/

Tempo nauki poprzez gry jest ograniczone i może zostać przyspieszone automatycznym importem wiedzy z innych sieci semantycznych. WordNet jest tworzony ręcznie i zawiera wiedzę pewną oraz uporządkowaną. W kolejnym etapie rozwoju projektu planowana jest integracja ze słownikiem WordNet. Wiedza zaimportowana z WordNeta ma stanowić najniższą warstwę sieci semantycznej, nad którą powstawać będzie warstwa tworzona poprzez interakcje z użytkownikami: wzbogacająca ubogi zasób relacji zdefiniowanych w WordNecie i zmieniająca charakter wiedzy na bardziej zdroworozsądkowy.

Innym zagadnieniem wciąż wymagającym opracowania, jest motywacja użytkowników do udzielania jak najwłaściwszych odpowiedzi oraz filtrowanie odpowiedzi błędnych. W grze opisanej w [13] gracz dostaje punkty jedynie, jeśli odpowie tak samo, jak losowo dobrany partner – dlatego stara sie odpowiadać jak najwłaściwiej. Z kolei w [6], co najmniej dwóch użytkowników musiało ocenić tak samo prawdziwość zdania wygenerowanego automatycznie na podstawie danych z ConceptNetu i Wikipedii, by zdanie zostało oznaczone jako prawdziwe. Implementacja podobnego mechanizmu jest konieczna do podniesienia jakości wiedzy pozyskiwanej przez system.

Planowane jest też rozszerzenie ostatniej fazy gry, w której użytkownik może dodawać do systemu nowe cechy. Na początku będzie ono służyć do weryfikacji wniosków przyjętych przez system przy próbie generalizacji wiedzy. Przykładowo, na podstawie faktów „kaczka ma skrzydła”, „skowronek ma skrzydła”, „kaczka jest ptakiem” i „skowronek jest ptakiem” wysnuty może być wniosek „ptak ma skrzydła”. Rozszerzenie ostatniej fazy polegać będzie na zadawaniu użytkownikowi dodatkowych pytań „czy prawdą jest, że. . . ?”. Taka forma dialogu może być dobrym punktem wyjścia do ewolucji systemu w stronę interaktywnego dialogu miedzy użytkownikiem a systemem.

Obecnie na stronie internetowej z gry w pytania istnieje możliwość zalogowania się za pomocą konta na portalu Facebook. Użytkownik musi przyznać aplikacji Winston Harris prawo dostępu do podstawowych informacji o sobie. Planuje się, aby był to punkt rozpoczynający znajomość z użytkownikiem Winston Harris. Dokumentacja Facebooka [14] wyraźnie wskazuje na istnienie silnej korelacji odwrotnej między liczbą uprawnień, o które prosi aplikacja, a liczbą użytkowników zgadzających się na przyznanie tych uprawnień. Z drugiej strony, użytkownicy nie przywiązują zbyt dużej wagi do potwierdzania relacji znajomości z innymi użytkownikami. Dlatego w momencie autoryzacji aplikacji, użytkownik Winston Harris będzie wysyłał graczowi zaproszenie do grona znajomych. Otworzy to możliwości bogatej interakcji (np. składanie życzeń przez awatara na urodziny gracza), dającej szanse na zwiększenie popularności systemu.

BIBLIOGRAFIA

[1] G.A. Miller, R. Beckitch, C. Fellbaum, D. Gross, K. Miller. *Introduction to wordnet: An on-line lexical database*. 1993.

[2] H. Liu, P. Singh. *Conceptnet – a practical commonsense reasoning tool-kit.* BT Technology Journal, 22:211–226, Październik 2004.

[3] L. Vanderwende, G. Kacmarcik, H. Suzuki, A. Menezes. *Mindnet: an automatically created lexical resource.* Proceedings of HLT/EMNLP on Interactive Demonstrations, str. 8–9. Association for Computational Linguistics, 2005.

[4] D.B. Lenat, R.V. Guha, K. Pittman, D. Pratt, M. Shepherd. *Cyc: toward programs with common sense.* Communications of the ACM, 33(8):30–49, 1990.

[5] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, O Etzioni. *Open information extraction from the web*. In Proceedings of the International Joint Conference on Artificial Intelligence, 2670–2676, 2007.
[6] A. Herdagdelen, M. Baroni. *The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose*, 2010.
[7] H. Lin, J. Davis, Y. Zhou. *Integration of computational and crowdsourcing methods for ontology extraction*. Proceedings of the 5th International Conference on Semantics, Knowledge and Grid(SKG2009), 2009.
[8] L. von Ahn. *Games with a purpose.* IEEE Computer Magazine, 39(6):92–94, 2006.
[9] J. Szymanski, W. Duch. *Context search algorithm for lexical knowledge acquisition.* 2010.
[10] J. Szymanski, W. Duch. *Information retrieval with semantic memory model.* Cognitive Systems Research, 2011.
[11] R. Burgener. Artificial neural network guessing method and game. Patent EP 1 710 735 A1. 2006.
[12] W. Duch, J. Szymanski, T. Sarnatowicz. *Concept description vectors and the 20 question game.* 2005.
[13] K. Thoring, R. M. Muller. *Semantic dimensions: A web-based game to evaluate the meaning of form.* 2010.
[14] Facebook Developers,*Core Concepts > Authentication*, Źródło internetowe, dostęp: 30 maja 2011. http://developers.facebook.com/docs/authentication/

AN WORD GAME FOR LEXICAL KNOWLEDGE ACQUISITION

Summary

This article describes a word game implementation – a model of context searcher and a lexical commonsense knowledge acquisition tool. Significance of context search was stated here followed by an exact description of the algorithm finding objects by their features. Knowledge representation and approach to knowledge acquisition was presented here along with how they relate to other known projects dealing with this subjects.