Ewa Thlon, Tomasz Pędzimąż

Studium anomalii relacyjnego opisu danych

1. Wstęp

Słownik semantyczny to program zawierający informacje o pojęciach języka naturalnego i o zależnościach występujących pomiędzy nimi. Posiadanie takiej bazy okazuje się bardzo cenne, zwłaszcza gdy mamy na względzie wykorzystanie algorytmów NLP – szczególnie tych próbujących odtworzyć sposób funkcjonowania ludzkiego mózgu. Przykładem zastosowania słownika może być ujednoznacznianie pojęć polegające na rozstrzygnięciu, w którym ze znaczeń ono występuje.

Konstrukcja słownika jest procesem długotrwałym i niemożliwym do pełnego zautomatyzowania, co implikuje konieczność tworzenia go przez wielu autorów. Zwiększa to podatność na błędy, a więc negatywnie wpływa na jakość i spójność sieci powiązań.

Miejsca występowania potencjalnych błędów może wskazywać istnienie anomalii – odstępstw od reguł obowiązujących w słowniku. Pełne zidentyfikowanie potencjalnych typów anomalii i implementacja algorytmu wykrywającego je prowadzi zatem do znaczącej poprawy jakości słownika.

2. Słownik semantyczny

Jak wspomniano wyżej, słownik semantyczny to baza wiedzy przechowująca opisy znaczeń wyrazów. Owo znaczenie jest określane poprzez zestaw relacji łączących słowo definiowane z wyrazami definiującymi. Pierwszą zakrojoną na szerszą skalę próbę stworzenia takiego słownika podjął w 1985 roku George Miller w trakcie prac nad WordNetem[1]. W następnych latach powstały analogiczne projekty dla innych języków: EuroWordNet[2], Multilingual Central Repository[3], Global WordNet[4], SłowoSieć[5] i inne[6].

[1] Zob. http://wordnet.princeton.edu/ (dostęp 19.04.2011).

[2] Język czeski, holenderski, estoński, francuski, hiszpański, niemiecki, włoski (zob. http://www.illc.uva.nl/EuroWordNet) (dostęp 19.04.2011).

[3] Zob. http://www.lsi.upc.es/~nlp/meaning/demo/demo.html (dostęp 19.04.2011).

[4] Zob. http://www.globalwordnet.org/ (dostęp 19.04.2011).

[5] Zob. http://plwordnet.pwr.wroc.pl/main/ (dostęp 19.04.2011).

[6] Zob. http://www.globalwordnet.org/gwa/wordnet_table.htm/ (dostęp 19.04.2011).

W ujęciu relacyjnym słowa są zorganizowane w formę relacyjnej sieci semantycznej. Istnieje wiele typów relacji, które mogą być uwzględnione w strukturze takiej sieci. Zgodnie z tradycją językoznawczą da się je podzielić na dwie główne kategorie:

- relacje paradygmatyczne klasyfikujące pojęcia przez opisywanie hierarchicznych zależności pomiędzy nimi;
- relacje syntagmatyczne opisujące pozostałe zależności – stany i akcje występujące w zdaniach języka naturalnego.

W języku występują oba typy relacji semantycznych, jednak istniejące słowniki (np. WordNet) w większości wypadków ograniczają się jedynie do relacji paradygmatycznych. Sama znajomość taksonomii często nie jest wystarczająca do satysfakcjonującego działania wielu typów algorytmów przetwarzania języka, ponieważ zależności tego rodzaju rzadko występują w zdaniach języka naturalnego. Dla przykładu rzadko spotkamy się ze zdaniem typu *Spaniel jest rasą psa*. Częściej natomiast słyszymy wypowiedzi: *Spaniel szczeka* lub *Pies szczeka*. Dla zapewnienia skuteczności algorytmów korzystających ze słownika dobrze jest zatem uwzględnić w nim także drugi typ relacji, gdyż to właśnie z nimi mamy najczęściej do czynienia w zdaniach. Mowa tu o relacjach syntagmatycznych, opisujących akcje i stany, w jakich może się znajdować obiekt reprezentowany przez dane pojęcie.

Przykładem słownika zawierającego oba typy zależności jest *Słownik Semantyczny Języka Polskiego* tworzony w ramach prac Katedry Lingwistyki Komputerowej UJ. Na jego przykładzie zostaną omówione anomalie, które mogą potencjalnie wystąpić.

W słowniku istnieją relacje:

1. paradygmatyczne:
   – **synonymy** – identyczność;
   – **similar_to** – podobieństwo;
   – **is_a_part_of** – relacja bycia częścią;
   – **consists_of** – relacja składania się z elementów;
   – **is_a_kind_of** – relacja bycia rodzajem;
   – **is_a** – relacja bycia specjalizacją;
2. syntagmatyczne:
   – **destination** – określa przeznaczenie lub cel istnienia obiektu;
   – **destination_rt** – warunkowa relacja przeznaczenia;
   – **role** – rola, w jakiej obiekt może występować;
   – **action** – akcja neutralna, niezwiązana z przeznaczeniem obiektu;
   – **action_positive** – akcja zgodna z przeznaczeniem obiektu;
   – **action_positive_rt** – warunkowa relacja akcji pozytywnej;
   – **action_negative** – akcja niezgodna z przeznaczeniem obiektu;
   – **action_negative_rt** – warunkowa relacja akcji negatywnej;
   – **action_passive** – relacja, gdzie obiekt jest przedmiotem akcji;
   – **action_passive_rt** – warunkowa relacja akcji pasywnej;
   – **state_positive** – stan zgodny z przeznaczeniem obiektu;
   – **state_positive_rt** – warunkowa relacja stanu pozytywnego;
   – **state_negative** – stan niezgodny z przeznaczeniem obiektu;
   – **state_negative_rt** – warunkowa relacja stanu negatywnego.

Poniżej została przedstawiona próbka słownika semantycznego dla hasła **admirał**.
Hasło słownika: **admirał** (oficer marynarki)
category: człowiek
synonymy:
similar_to:
is_a_part_of: marynarka, flota
consists_of:
is_a_kind_of: marynarz
is_ a:
destination: dowodzić/dowodzenie, walczyć/walka, żeglować/żeglowanie/żegluga
role:
related_to: flota, eskadra, flotylla **is**: dowódca, 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
related_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**: tchórz/tchórzostwo/tchórzliwość

Hasło słownika: **admirał** (motyl)
category: zwierzę
synonymy:
similar_to:
is_a_part_of: natura
consists_of:
is_a_kind_of: motyl
is_a:
destination: zapylić/zapylać/zapylanie
role:
action:
positive:
related_to: kwiat **is**: (zbierać) nektar
negative: choroba, niedyspozycja

passive: zginąć
state:
positive: żyć/żywy
negative: martwy

3. Anomalie występujące w słowniku

Występowanie błędów i niespójności w strukturze słownika to zjawisko niepożądane. Poprawne ich zidentyfikowanie i usunięcie jest więc bardzo ważne dla zapewnienia dobrej jakości słownika i algorytmów z niego korzystających. Ze względu na specyfikę języka w większości wypadków nie można automatycznie określić pojedynczej relacji lub ich grupy jako jednoznacznie błędnych. W słowniku istnieją jednak typy anomalii, których wykrycie znacząco zwiększa ryzyko wystąpienia błędu – są to sytuacje, gdy dany układ relacji bywa poprawny w tak małej liczbie przypadków, że można wskazać go z dużym prawdopodobieństwem jako błędny, wspomagając w ten sposób pracę twórców słownika. Zidentyfikowano cztery podstawowe typy potencjalnych anomalii:

1. niepełność danych – gdy został stwierdzony brak istnienia pewnych relacji;
2. redundancja danych – gdy dane zostały powielone;
3. sprzeczność danych – gdy znaleziono grupę relacji wzajemnie się wykluczających;
4. błąd typu danych – gdy typ danych nie jest prawidłowy dla relacji.

3.1. Niepełność danych

W szczególnych przypadkach dochodzi do zidentyfikowania danych jako niepełnych, np. w relacjach tożsamościowych – relacjach synonymy i similar_to, gdyż wiążą one pojęcia bliskie znaczeniowo. Problem ten może się pojawić podczas dodawania nowych danych do słownika, a także łączenia relacjami pojęć już istniejących.

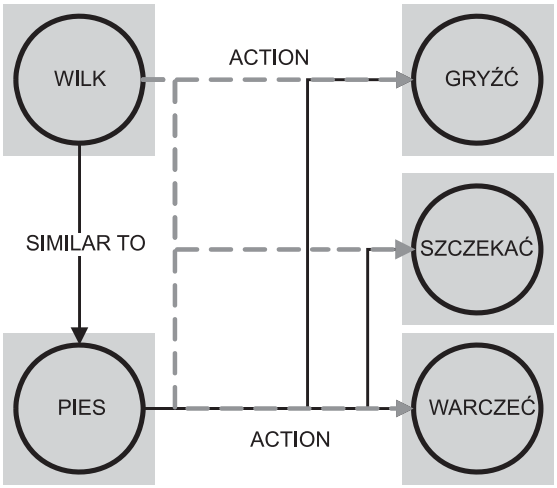
Anomalię niepełności danych da się prosto zidentyfikować dla relacji synonymy – powiązane nią pojęcia powinny mieć identyczny zbiór relacji. Anomalia ta może wystąpić także dla relacji similar_to w następujących przypadkach:

- pojęcia połączone nią mają rozłączny zbiór relacji;
- istnieje kilka pojęć podobnych i większość z nich ma pewną relację, a inne nie.

W przypadku zidentyfikowania anomalii przedstawionej na rysunku 1 zazwyczaj zachodzi potrzeba dodania nowych relacji pomiędzy pojęciami.

3.2. Redundancja danych

Istnieją typy relacji wiążących pojęcia, których wprowadzenie powoduje brak konieczności powtarzania innych zależności – mowa tu o hierarchicznych relacjach is_a_kind_of i consists_of. Jeśli dane pojęcie jest specjalizacją innego, to automatycznie dziedziczy jego wszystkie relacje syntagmatyczne. Analogiczna zależność występuje w relacji bycia częścią – jeśli części składowe mają relacje syntagmatyczne, całość także je ma.



Rysunek 1. Przykład niepełności danych – jeśli *pies* jest podobny do *wilka* i jest w relacji z pojęciami: *gryźć*, *szczekać* i *warczeć*, to być może *wilk* też powinien być z nimi połączony. W przedstawionej na rysunku anomalii zazwyczaj zachodzi potrzeba dodania nowych relacji pomiędzy pojęciami

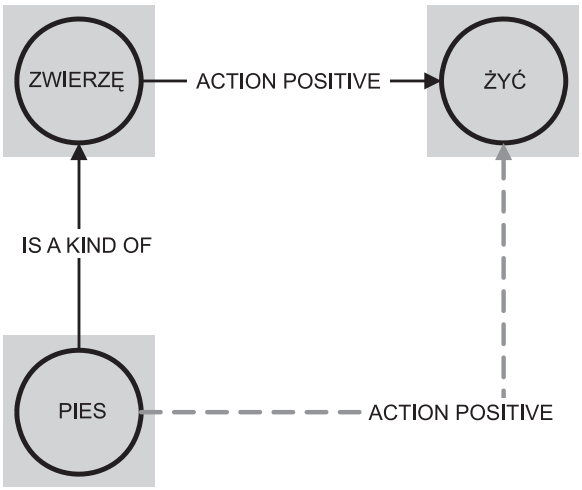
Anomalia powtarzania relacji w ramach pojęć połączonych zależnościami hierarchicznymi powinna zostać zawsze usunięta. Inny przypadek redundancji danych polega na tym, że wszystkie specjalizacje danego pojęcia (w dół hierarchii) zawierają tę samą relację. Istnieje wówczas duże prawdopodobieństwo potrzeby przeniesienia jej w górę hierarchii i powiązania bezpośrednio z generalizacją.

3.3. Problem sprzeczności danych

Problem sprzeczności danych może zostać zidentyfikowany dla obu typów relacji, jednak najprościej dla relacji paradygmatycznych. Wyszczególniamy tu następujące przypadki:

- sprzeczność synonimiczną, gdy dwa pojęcia powiązane relacją synonymy lub similar_to posiadają przeciwstawne[7] powiązanie z innym pojęciem;
- sprzeczność kompozycyjną, gdy dwa pojęcia powiązane relacją is_a_part_of lub consists_of posiadają przeciwstawne powiązanie z innym pojęciem;
- sprzeczność taksonomiczną, gdy dwa pojęcia powiązane relacją is_a_kind_of lub is_a mają przeciwstawne powiązanie z innym pojęciem.

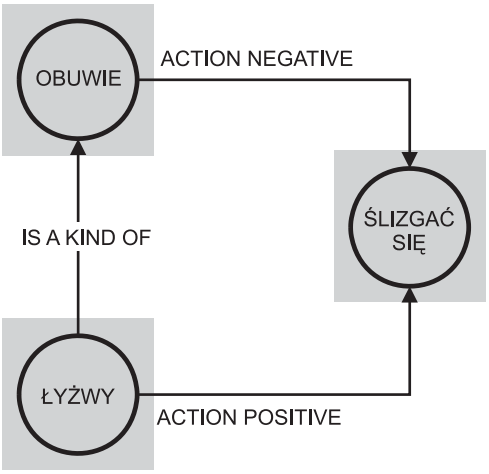
[7] Relacje przeciwstawne to relacje informujące o występowaniu zgodności lub niezgodności z przeznaczeniem lub celem istnienia obiektu. Relacjami tymi są pary: action_positive i action_negative, state_positive i state_negative oraz action_positive_rt i action_negative_rt, a także state_positive_rt i state_negative_rt, pod warunkiem że relacja warunkowa dotyczy tego samego obiektu.



Rysunek 2. Przykład redundancji danych (relacja typu action positive pomiędzy pojęciem *pies* a *żyć* jest zbędna, gdyż jest już odziedziczona po pojęciu *zwierzę*)

Jest to oczywisty przykład anomalii, istnieją jednak wypadki, gdy celowe wprowadzenie takiego układu relacji pozwala na wyrażenie odstępstwa od reguły (przykład na rysunku 3), więc jej istnienie nie może być jednoznacznie uznawane za błąd.

Innym przykładem sprzeczności jest mogący wystąpić w relacjach nietożsamościowych problem cyklu. Warunek konieczny wystąpienia tej anomalii to kierunkowość

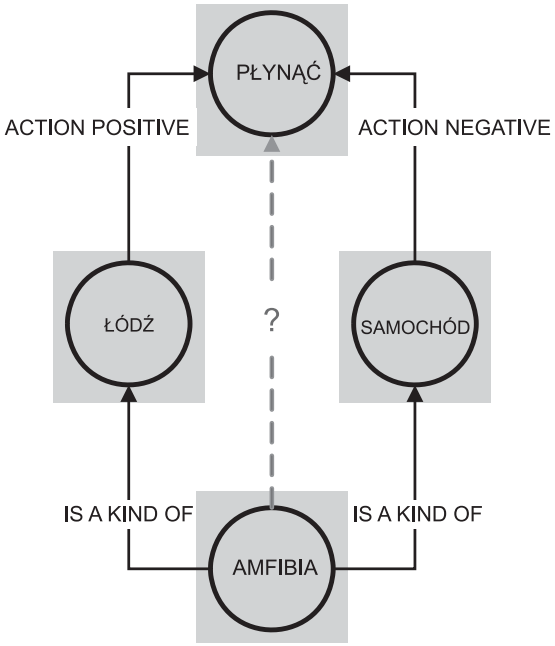


Rysunek 3. Przykład uzasadnionej sprzeczności

relacji – synonymy i similar_to są relacjami tożsamościowymi[8], więc problem cyklu nie wystąpi. Cykl zalicza się do zjawisk niepożądanych i występuje w przypadku, gdy wychodząc od danego pojęcia, można osiągnąć to samo pojęcie, przechodząc przez relacje tego samego typu.

Jeszcze innym przykładem sprzeczności jest problem kraty. Występuje on, gdy istnieje pojęcie powiązane z dwoma innymi relacją umożliwiającą dziedziczenie połączeń (omówione w punkcie 3.2 – is_a_kind_of lub consists_of) i gdy dziedziczy ono sprzeczne relacje.

Aby rozwiązać anomalię kraty, należy przeciążyć relacje sprzeczne, dodając tę zależność jeszcze raz i celowo wprowadzając anomalię redundancji danych.



Rysunek 4. Przykład anomalii kraty – problem z określeniem relacji pomiędzy pojęciem *amfibia* a *płynąć*

3.4. Błąd typu danych

W tak zorganizowanym słowniku można rozumieć typy pojęć jako sam szczyt hierarchii w ramach relacji is_a_kind_of – byty i czynności. Przy takim rozumieniu typów da się stworzyć obostrzenia na nie, określając, co może się znajdować po każdej ze stron poszczególnych relacji (por. tabela 1).

[8] Możemy powiedzieć, że relacja jest tożsamościowa, jeśli stwierdzenie, że „A jest w relacji R z B" implikuje fakt, że „B jest w relacji R z A" (zależność jest prawdziwa bez względu na kierunek relacji). W przyjętym modelu słownika relacjami tożsamościowymi są jedynie relacje synonymy i similar_to.

Tabela. 1. Ograniczenia typu argumentów relacji

Rodzaj relacji	Lewa strona relacji	Zależność related to	Prawa strona relacji
synonymy	byt	–	byt
	czynność		czynność
similar_to	byt	–	byt
	czynność		czynność
is_a_part_of	byt	–	byt
	czynność		czynność
consists_of	byt	–	byt
	czynność		czynność
is_a_kind_of	byt	–	byt
	czynność		czynność
is_a	byt	–	byt
	czynność		czynność
destination	byt	–	czynność
destination_rt	byt	byt	czynność
		czynność	
role	byt	–	byt
action	byt	–	czynność
action_positive	byt	–	czynność
action_positive_rt	byt	byt	czynność
		czynność	
action_negative	byt	–	czynność
action_negative_rt	byt	byt	czynność
		czynność	
action_passive	byt	–	czynność
action_passive_rt	byt	byt	czynność
		czynność	
state_positive	byt	–	byt

state_positive_rt	byt	byt	byt
		czynność	
state_negative	byt	–	byt
state_negative_rt	byt	byt	byt
		czynność	

Pojawienie się odstępstwa od przedstawionej reguły świadczy o istnieniu błędu.

4. Podsumowanie

W słowniku semantycznym występują anomalie. Zidentyfikowanie ich typów, implementacja algorytmu automatycznego ich wykrywania oraz kontrolne wykonywanie go wraz z rozwojem słownika może pozytywnie wpłynąć na jakość przechowywanych w nim danych poprzez wskazywanie miejsc potencjalnych błędów. Pomimo że pewne typy anomalii jednoznacznie świadczą o występowaniu błędów, to pełne zautomatyzowanie tego procesu nie jest możliwe i konieczny okazuje się udział lingwisty rozstrzygającego, czy istnienie danej anomalii to rzeczywista pomyłka, czy celowo wprowadzone odstępstwo od reguły.

BIBLIOGRAFIA

Beckwith R., Miller G.A., Tengi R. (1998). *Design and Implementation of the WordNet Lexical Database and Searching Software*, [w:] Ch. Fellbaum, *WordNet an Electronic Lexical database*. Cambridge, MA: The MIT Press.

Duch W. *Wstęp do kognitywistyki*, http://www.fizyka.umk.pl/~duch (data dostępu 19.04.2011).

Hanks P. (2004). WordNet: What is to be Done? http://www.fi.muni.cz/gwc2004/pres/panel/Hanks/hanks-panel.pdf (data dostępu 19.04.2011).

Jurafsky D., Martin J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics and Speech Recognition. Upper Saddle River, NJ: Prentice Hall.

Lubaszewski W. (red.) (2009). *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. Kraków: Wydawnictwo Akademii Górniczo-Hutniczej.

Mykowiecka A. (2007). *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: Polsko-Japońska Wyższa Szkoła Technik Komputerowych.

Richens T. (2008). *Anomalies in the WordNet verb hierarchy*. COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics, 1. Manchester, UK: Association for Computational Linguistics.

Relational Description of Data – Study of Anomalies

This article presents the idea of a semantic dictionary and concentrates on the problem of anomalies which might appear in it. The existence of anomalies indicates the points of potential errors in the semantic network, thus the implementation and consequent application of an anomalies discovery algorithm, should have a positive impact on the quality of the semantic network. The presented problem and its solution are general, and might be applied to any knowledge source, but this work concentrates on the Polish Semantic Dictionary developed in the Computational Linguistics Department at the Jagiellonian University.