

Opracowanie systemu wykorzystującego reguły podejście w procesie wykrywania relacji semantycznych wewnątrz frazy rzeczownikowej

Paweł Kędzia, Marek Maziarz

Celem zadania było opracowanie systemu umożliwiającego wykrywanie relacji semantycznych w tekstach pisanych językiem naturalnym. Narzuconym ograniczeniem było wykrywanie relacji jedynie we frazach rzeczownikowych.

Opracowany system na wejściu przyjmuje tekst, a na wyjściu produkuje ten sam tekst z dodatkowymi informacjami odnośnie wykrytych relacji semantycznych. Proces wykrywania relacji semantycznych jest dwuetapowy:

1. W pierwszym kroku, generowane są potencjalne powiązania między elementami frazy rzeczownikowej, dla których może zajść określona relacja semantyczna.
2. W kroku drugim, dla wygenerowanych par, bądź trójek z kroku 1-go, uruchamiane są reguły wykrywające relacje semantyczne między tymi elementami.

1. Krok 1. Generowanie potencjalnych powiązań

Do realizacji kroku pierwszego opracowany został moduł, umożliwiający wygenerowanie potencjalnych dwójek/trójek. Moduł ten może działać niezależnie od modułu wykrywającego relacje semantyczne. Wyjściem z modułu pierwszego jest pogłębienie płaskiej informacji chunkerowej. Moduł generowania potencjalnych powiązań wykorzystuje informacje o częściach mowy oraz o granicach frazy NP (rzeczownikowej) / AdjP (przymiotnikowej). W pierwszym kroku wykorzystaliśmy korpus KPWr [BrodaMarciniuk i inni, 2012] Dla przykładu, fraza:

projekt polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego w 1931 roku przez inż. Stanisława Praussa w Państwowych Zakładach Lotniczych w Warszawie

opisana jest w sposób następujący:

[projekt]AgP [polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego]AgP [w 1931 roku]AgP [przez inż. Stanisława Praussa]AgP [w Państwowych Zakładach Lotniczych]AgP [w Warszawie]AgP]NP

Cała fraza NP, została podzielona na frazy AgP:

- [projekt]
- [polskiego sześciomiejscowego samolotu pasażerskiego zaprojektowanego]
- [w 1931 roku]
- [przez inż. Stanisława Praussa]
- [w Państwowych Zakładach Lotniczych]
- [w Warszawie]

Podkreśleniem oznaczone zostały głowy fraz AgP, pogrubieniem została oznaczona głowa całej frazy NP. Takie informacje trafiają na wejście modułu generującego potencjalne powiązania, gdzie wynikiem jest tekst z dodatkowymi informacjami o możliwych powiązaniach między elementami. Dla powyższego przykładu wyjście będzie wyglądało następująco:

0:projekt 1:polskiego 2:sześciomiejscowego 3:samolotu
4:pasażerskiego 5:zaprojektowanego 6:w 7:1931 8:roku 9:przez

10:in| 11:. 12:Stanisława 13:Praussa 14:w 15:Państwowych
 16:Zakładach 17:Lotniczych 18:w 19:Warszawie

- 7. Relacja dla pary (OD - DO): < 1 , 3 >:
- 7. Relacja dla pary (OD - DO): < 2 , 3 >:
- 7. Relacja dla pary (OD - DO): < 4 , 3 >:
- 7. Relacja dla pary (OD - DO): < 5 , 3 >:
- 1.1 Relacja dla pary (OD - DO): < 0 , 3 >:
- 3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 6 , 8 >:
- 3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 9 , 10 >:
- 7. Relacja dla pary (OD - DO): < 15 , 16 >:
- 7. Relacja dla pary (OD - DO): < 17 , 16 >:
- 3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 14 , 16 >:
- 3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 18 , 19 >:

Co oznacza, że między tokenem o numerze 1, a tokenem o numerze 3 może zajść jakaś relacja semantyczna, między tokenem 2 a tokenem 3 może zajść jakaś relacja semantyczna, itd. Na całość można spojrzeć jak na drzewo rozkładu gramatycznego, w którym mamy wyróżnione zależności między elementami:



Generator par dla 2. kroku analizy semantycznej fraz można ubocznie traktować jako częściowy parser – analizator składniowy fraz NP oraz AdjP. Warto jednak dodać, że do tej pory wykorzystany został korpus już oznaczony frazami NP/AdjP oraz AgP/PP. Opracowany system do generowania par oraz trójek na wejściu przyjmuje korpus otagowany oraz oznaczony wymienionymi typami fraz. Nic nie stoi jednak na przeszkodzie aby w preprocesie uruchomić chunker oraz tager. Aby umożliwić wykrywanie relacji w tekście nieoznaczonym wykorzystany został tagger *wcrft* [Radziszewski 2013] oraz chunker *iobber* [Radziszewski, Pawlaczek, 2012].

System dla każdego dokumentu w korpusie pobiera jego frazy NP/AdjP po czym z tych fraz pobiera frazy AgP/PP. Dodatkowo podczas wczytywania korpusu system ustala czy zadany token jest predykatem pierwszego rodzaju (lemat słowa znajduje się w słowniku predykatów pierwszego rodzaju) predykatem drugiego rodzaju (lemat znajduje się w słowniku predykatów drugiego rodzaju) jak również odczytuje informacje o głowach fraz (NP/AdjP, AgP/PP). Następnie dokonywana jest analiza fraz NP/AdjP, podczas której ustalane są powiązania między elementami fraz. Analiza ma charakter regułowy. Reguły generowania par/trójek mogą być ciągle rozszerzane

ze względu na opracowany zestaw metod operujących na elementach fraz. Np. metody umożliwiają sprawdzenie czy na prawo od zadanej pozycji we frazie znajduje się predykat pierwszego typu, umożliwiają pobranie najbliższego predykatu pierwszego/drugiego typu, najbliższej głowy, umożliwiają sprawdzenie typu danej frazy (AgP/PP) oraz wiele innych operacji. Opracowany generator par/trójek generuje obiekty, w których:

- zapisana jest oryginalna postać frazy NP/AdjP,
- podane są pozycje elementów frazy, które wchodzi w skład wygenerowanej dwójki/trójki,
- zapisany jest obiekt opakowujący wygenerowane pozycje w sztuczne zdanie (zawiera odpowiednie tokeny z frazy NP/AdjP).

Oprócz wspomnianego obiektu, generowany jest również numer reguły na podstawie której wygenerowana została dwójka/trójka. Następnie, wygenerowana dwójka/trójka zaaplikowana może być do modułu regułowego umożliwiającego uruchomienie operatora WCCL na przekazanych elementach (opisane w **kroku 2**).

Wyniki generatora (sprawdzone pod kątem poprawności składniowej generowanych trójek i dwójek) przedstawia poniższa tabela:

NP	AdjP
$P = 160/184 = 87.0\%$	$P = 100/113 = 88.5\%$
$R = 162/198 = 82.3\%$	$R = 104/143 = 72.7\%$
$F = 84.6\%$	$F = 79.8 \%$

Generator został surowo sprawdzony, tzn. wzięte zostały pod uwagę także takie typy powiązań składniowych, które w ogóle nie zostały uwzględnione w generatorze. Ocena zatem to porównanie do idealnego rozbioru składniowego.

2. Krok 2. Wykrywanie relacji semantycznych

Prace odbywały się nad określonymi relacjami semantycznymi, które wyłonione zostały na podstawie analiz frekwencji danego typu relacji semantycznej w próbkach korpusu KPWr. Lista rankingowa wygląda następująco:

Częstotliwość	Nazwa relacji semantycznej
0.21	Właściwość
0.16	Obiekt
0.14	Subiekt
0.09	? (relacja niedookreślona)
0.06	Miejsce
0.06	IDENT
0.05	Przeznaczenie
0.05	Meronimia / holonomia
0.04	Relacja genetyczna (pochodzenie)

0.04	Wytwór / rezultat
0.04	Sposób/parametr
0.03	Instrument / narzędzie (czynności) / środek czynności
0.02	LIMIT (użycie limitujące)
0.01	Materiał / obiekt materiałowy
0.01	Symilatywność
0.01	Kauzacja / przyczyna
0.00	Hiponimia/ hiperonimia
0.00	Krzyżujące się zakresy
0.00	Czas
0.00	Potencjalność
0.00	Habitualność
0.00	Kwantytatywność i ocena

Po wykryciu par i trójek - potencjalnych instancji relacji semantycznych – uruchamiany jest zestaw operatorów WCCL-a, np. dla przykładowej frazy NP wydobyte zostały następujące relacje semantyczne (na niebiesko relacje wykryte, na brązowo relacje niewykryte):

0:projekt 1:polskiego 2:sześciomiejscowego 3:samolotu
4:pasażerskiego 5:zaprojektowanego 6:w 7:1931 8:roku 9:przez
10:inż 11:. 12:Stanisława 13:Praussa 14:w 15:Państwowych
16:Zakładach 17:Lotniczych 18:w 19:Warszawie

7. Relacja dla pary (OD – DO): < 1 , 3 >:

POCHODZENIE

7. Relacja dla pary (OD – DO): < 2 , 3 >:

HOLONIMIA

7. Relacja dla pary (OD – DO): < 4 , 3 >:

PRZEZNACZENIE

7. Relacja dla pary (OD – DO): < 5 , 3 >:

OBIEKT-PPAS-0

1.1 Relacja dla pary (OD – DO): < 0 , 3 >:

OBIEKT

3.1.1 Relacja dla trojki (OD – prep – DO): < 5 , 6 , 8 >:

CZAS-PP-0

3.1.1 Relacja dla trojki (OD – prep – DO): < 5 , 9 , 10 >:

SUBIEKT-PPAS-przez-acc-0

7. Relacja dla pary (OD – DO): < 15 , 16 >:

SUBIEKT POSIADANIA

7. Relacja dla pary (OD – DO): < 17 , 16 >:

PRZEZNACZENIE

3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 14 , 16 >:

MIEJSCE

3.1.1 Relacja dla trojki (OD - prep - DO): < 5 , 18 , 19 >:

MIEJSCE-PP-0

2.1 Konfiguracja w pliku

Dodatkowo, warto zaznaczyć, że konkretny operator może być podpięty pod konkretną regułę generującą parę/trójkę. Ma to na celu uniknięcie odpalania operatorów WCCLa na z definicji nieodpowiednich dwójkach/trójkach dla tego operatora. Mechanizm ten konfiguruje się poprzez odpowiedni wpis w pliku konfiguracyjnym. Plik ten zawiera dwie sekcje [operators] oraz [rules]:

```
[operators]
particip = resources/operatory/OS_mod7_participium.ccl
not_particip = resources/operatory/OS_mod7_not_participium.ccl
```

Sekcja [operators] w pliku konfiguracyjnym zawiera wykaz operatorów, które podpinane będą do konkretnych reguł. Wpis jest postaci:

```
unikalna_nazwa_operatora = ścieżka/do/pliku_operatora.ccl
```

Po czym, w tym samym pliku konfiguracyjnym, w sekcji [rules] ustawiane jest powiązanie reguł generowania par/trójek z operatorami WCCL:

```
[rules]
1 = not_particip
2 = not_particip
3 = not_particip
4 = not_particip
5 = not_particip
6 = not_particip
7 = particip
```

Wpisy mają postać:

```
numer_reguły = nazwa_operatora_z_sekcji_operators
```

2.2 System uruchomieniowy

W celu uruchomienia wykrywania relacji semantycznych wykorzystując regułowe podejście, opracowany został system operujący na generowanych parach/trójkach oraz operatorach WCCL. System znajduje się w repozytorium npsemrel i uruchamia się go poprzez uruchomienie programu `rb_run.py`. Przykładowe wywołanie:

```
python rb_run.py \
    -w cfg/ops.ini
    -d /home/pkedzia/work/corps/kpwr-disamb-1.1.4 \
```

```
-C index_chunks.txt \
-O \
-q
```

Gdzie poszczególne argumenty wywołania, to:

- `rb_run.py` - system regułowego wykrywania relacji
- `-w` - plik konfiguracyjny z powiązaniem operatorów z numerami reguł (opisany wcześniej)
- `-d` - ścieżka do korpusu, na rzecz którego wykrywane są relacje semantyczne
- `-C` - plik zawierający indeks plików z korpusu (dla wymienionych plików wygenerowane zostaną dvojki/trójki oraz uruchomione zostaną operatory)
- `-O` - przełącznik mówiący o tym, że rozpatrujemy tylko ciągłe frazy
- `-q` - przełącznik oznaczający tryb “cichy” - system nie wyświetla dodatkowych informacji (np., że znalazł predykat pierwszego typu)

Po uruchomieniu, system będzie analizował korpus i zgodnie z formatem podanym wcześniej w przykładach, będzie wyświetlał informacje o wygenerowanych parach/trójkach oraz wykrytych relacjach.

2.3 Dodatkowe zasoby

Dodatkowo, opracowany został szereg słowników, które wykorzystane zostały zarówno w procesie generowania par/trójek, jak również wewnątrz operatorów. Słowniki tworzone były na podstawie Słownosieci, ram walencyjnych Dębowskiego, wykazu miejscowości z gazetteera oraz wykazu odmian wyciągniętych z Morfeusza. Te słowniki to:

1. Wykaz agensów, pobranych ze słownosieci jako hiponimy SUMO-wego agensa
2. `dict-budowla.lex` - hiponimy {obiektu budowlanego (wytw)} do operatora MIEJSCE
3. `dict-charakteryzowanie.lex` - relacja charakteryzowanie ze Słownosieci - do operatora WLASCIWOSC
4. `dict-city.lex` - słownik miast z gazetteera do operatora MIEJSCE
5. `dict-country.lex` - słownik państw z gazetteera do operatora MIEJSCE
6. `dict-derywacyjnosc.lex` - ze Słownosieci
7. `dict-frame_acc.lex` - ze słownika Dębowskiego, bierzemy informację o najczęstszej ramie walencyjnej - służy do rozpoznawania SUBIEKTÓW i OBIEKTÓW
8. `dict-habitualnosc.lex` - do SUBIEKTU, relacja habitualności ze Słownosieci
9. `dict-kolejny.lex` - liczebniki porządkowe, do WLASCIWOSCI (filtrowanie)
10. `dict-noun-domain.lex` - domena pierwszych znaczeń rzeczowników, wykorzystywane w różnych operatorach
11. `dict-potencjalnosc.lex` - do OBIEKTU, relacja potencjalności ze Słownosieci
12. Słownik predykatów drugiego typu: `dict-pred-snd-type.lex` -
 - a. `apel` -> `apelować`
 - b. `argument` -> `argumentować`
 - c. `blaga` -> `blagować`
 - d. ...- ten słownik jest wykorzystywany w algorytmie znajdowania par i trójek.
13. `dict-przym_dziedzina_rel.lex` - operator WLASCIWOSC, podaje, które przymiotniki są relacyjne

14. dict-przymiotniki.lex - słownik wszystkich przymiotników Słownosieci
15. dict-rammy.lex - słownik wszystkich czasowników opisanych w słowniku Dębowskiego
16. dict-rola_adj_v.lex - słownik relacji roli adj
17. Wykaz predykatów pierwszego typu: dict-syn-mpar.lex (absolutyzacja -> cumy, absolutyzowanie -> cumy) oraz dict-gerundium.lex (agregować -> agregowanie, akać -> akanie)
18. dict-syn_mpar_przym.lex - słownik synonimii międzypadarygmatycznego dla przymiotnika - operator WŁASCIWOSCI
19. słownik: val-dic, zawierający szereg słowników dla poszczególnych typów ram walencyjnych, z których każdy słownik zawiera wykaz słów, dla których prawdziwa jest rama, np. słownik: dict-acc-na.lex zawiera ramę dla biernika, po którym występuje słowo "na" np.: *wsadzić na*, *wyjść na* itp.

Jako artefakt uboczny powstało narzędzie umożliwiające tworzenie wykazu wszystkich hiponimów dla zadanego numeru synsetu (opracowany został słownik pomocniczy: pierwsze znaczenie w synsecie na numer synsetu)

3. Definicje relacji oraz ocena ręczna

Niech:

- +m = byt odczuwający, myślący, działający: osoba, instytucja, organizacja, państwo lub zwierzę
- +v = działający zgodnie z własną wolą (zakładając, że zwierzę też ma wolę)
- +c1 = X powoduje, że zachodzi sytuacja Y
- +c2 = sytuacja Y jest spowodowana przez X
- +f1 = X odczuwa coś, ma jakieś emocje, postrzega coś zmysłami
- +p1 = X posiada coś
- +t2 = Y jest użyte do czegoś przez kogoś
- +r = X powstaje w trakcie akcji
- +k = X jest spokrewniony lub spowinowacony z Y
- +loc = Y jest miejscem dla akcji, stanowi arenę czynności, procesów lub stanów
- +temp = Y to czas, w czasie którego coś się dzieje, moment w czasie bądź okres o jakiejś długości

#SUBIEKT - zbiorcza nazwa dla trzech podtypów subiektu (poniżej). Obecnie nie wykrywamy trzech podtypów, tylko jeden typ główny (#SUBIEKT). Nie uwzględniamy również w naszym algorytmie #subiektu posiadania. Nie uwzględniamy w ocenie także subiektu przy niewyrażonym predykacie (tj. zawsze musi być predykat nominalizowany). Subiekt definiowany jest jako:

#SUBIEKT = [+/-m] [+/-v] [+/-c1] [-c2] [+/-p1] [-t2] [+/-r] [+/-k] [+/-loc] [+/-temp]

#agens

- ożywiony wykonawca czynności (odpowiada mu najczęściej pierwszy argument w strukturze P-A), działający w zgodzie ze swoimi pragnieniami i wolą (intencjonalnie)
- w szczególności #agensem może być człowiek, instytucja i zwierzę
- przez czynność rozumiemy sytuację dynamiczną niezmiennostanową (w typologii Laskowskiego są to *L*-czynności) lub zmiennostanową (*L*-działania, *L*-akty, *L*-sytuacje niemomentalne nieteliczne intencjonalne),
- z powyższego zbioru wyłączamy sytuacje (stany i akcje), które odnoszą się do zjawisk psychicznych i mentalnych
- w szczególności #agens jest wykonawcą czynności kauzatywnej (*L*-działania z elementem semantycznym 'powodować'),
- nie zaliczamy do #agensów zjawisk fizycznych

wędrowanie → Piotra

smażenie → przez Piotra

wycie → wilków

połączenie się → firm

(uwaga: wytłuszczonym drukiem oznaczamy predykaty, relacje idą od predykatów)

#experiercer

- byt ożywiony doświadczający jakiegoś zjawiska psychicznego bądź mentalnego (odpowiada mu najczęściej pierwszy argument w strukturze P-A),
- w szczególności #experiercerem może być człowiek bądź zwierzę [+m]
- należą tu m.in. subiekty lubienia, zamiłowania (przy predykatach przechodnich)
- #experiercer może świadomie i dobrowolnie kierować swoim procesem poznawczym [+/-v], nie może jednak być subiektem predykatu kauzatywnego ([-c]),
- nie zaliczamy do #experiercerów organizacji i instytucji

odczuwanie → Piotra

spostrzeżenie → przez Piotra

odczuwanie → zwierząt

#subiekt nieagentywny

- byt nieożywiony bądź ożywiony, zjawisko przyrody występujące w pozycji pierwszego argumentu w strukturze P-A przy predykatie nieintencjonalnym
- należą tu subiekty stanu (nosiciele stanu, cechy, Piotr ← śpiący, czerwoność → zachodu), subiekty zmiany stanu (tzw. procesory, np. wzrost → rododendronu), subiekty podobieństwa, subiekty relacji (dom ← przypominający, wyraz ← [znaczeniem] odpowiadający), subiekty mieszkańców (Piotr ← mieszkający), subiekty lokalizowane (materac ← leżący) itp.
- występuje przy predykatkach należących do klasy *L*-stanów, *L*-zdarzeń, tj. sytuacji dynamicznych (akcji), niezmiennostanowych

wzrost → rododendronu

wybuch → wulkanu

powiewanie → flagi

powiew → wiatru
 dom ← **przypominający**
 wyraz ← [znaczeniem] **odpowiadający**
 materac ← **leżący**
leżenie → Piotra
chrapanie → Janka
śpiący → Piotr
 Piotr ← **mieszkający**

Operatory WCCL-a wykorzystują następujące schematy składniowe:

N/V/ → **N_{SUB}(gen)**
wybór dziewczyny ← dziewczyna wybrała
krytyka pośłów ← pośłowie krytykują
placz matki ← matka płacze
czytanie Janka ← Janek czyta

N_{SUB} ← **Adj/V/**
wędrowne ptaki
troskliwa matka
mściwy człowiek

N/V/ → **PP_{SUB}**, **P_{pas}** → **PP_{SUB}**, **Adj/V/** → **PP_{SUB}**
 wczorajsza **kradzież** koca przez chłopca ← wczoraj chłopiej ukradł koc
czytanie książki przez Janka / przez niego ← Janek/on czyta książkę
wybór szefa przez załogę ← załoga wybiera szefa
czytany przez załogę ← załoga czyta
 (PP z przyimkiem "przez")

Wyniki przedstawia poniższa tabela:

#subiekt = {#agens, #experiencer, #subiekt nieagentywny}	Precyzja	Kompletność	miara F
NP	$P = \frac{3}{5} = 60\%$	$R = \frac{3}{8} = 37.5\%$	F = 46%
AdjP	$P = \frac{2}{4} = 50\%$	$R = \frac{2}{6} = 33\%$	F = 40%

Wydaje się, że można jeszcze poprawić precyzję wprowadzając dodatkowe ograniczenia i zwiększyć kompletność przez dodanie nowych operatorów (nowych schematów składniowych).
 Pytanie, czy chcielibyśmy rozróżniać role #agensa, #experiencera i #subiektu nieagentywnego.

W sumie w korpusie KPWr operatory SUBIEKTU uruchomiły się 157 razy.

Operator(y)	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
SUBIEKT	1.5%	14%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Przykład 3.:

0:przejęcie 1:świątyn 2:Kościoła 3:greckokatolickiego 4:przez
5:Rosyjską 6:Cerkiew 7:Prawosławną

3.1.1 Relacja dla trojki (OD - prep - DO): < 0 , 4 , 6 >:

SUBIEKT-N-przez-acc-0

wykryta trójka: przejęcie → przez Cerkiew

Operator poprawialiśmy na całym korpusie. Teraz występuje 61 razy w korpusie, ale z dużo wyższą precyzją - wkrótce sprawdzimy to wg wszelkich zasad sztuki.

Wyniki przedstawia poniższa tabela:

#subiekt	Precyzja	Kompletność	miara F
NP + AdjP	P = 58/61 = 95%	nieznana	nieznana

#subiekt posiadania - niewykrywany i nieuwzględniany przez nas w ocenach - definiujemy następująco: osoba, instytucja występująca w pozycji pierwszego argumentu (posiadacza) w strukturze P-A przy predykanie oznaczającym posiadanie, także osoba, instytucja przy niewyrażonym predykanie (oznaczająca posiadacza)

#OBIEKT

#Obiekt to byt, na który ukierunkowana jest akcja czynności bądź procesu, przy predykatkach relacyjnych jest to drugi element relacji, np. przy predykatkach oznaczających relację podobieństwa - jest to obiekt, do którego coś się przyrównuje, przy predykatkach oznaczających zamiłowanie - jest to obiekt zamiłowania, przy predykatkach oznaczających posiadanie - jest to obiekt posiadany. Odpowiada mu najczęściej drugi element w strukturze predykatowo-argumentowej, występuje przy predykatkach przechodnich. #Obiekt w toku akcji może ulegać zmianom, jednak nie jest wynikiem (#wytworem) akcji, tj. nie powstaje jako byt w toku akcji. Nie uznajemy za #obiekt także takiego argumentu, który oznacza #materiał. Wyróżnić możemy następujące rodzaje #obiektów:

- Byty podlegające zmianom o charakterze przestrzennym:

przesunięcie → pionka,

przerzucenie → wojska,
zrzut → zapasów.

- Byty podlegające zmianom stanu będącym wynikiem działania jakiejś przyczyny bądź agenta:

obieranie → jajka,
spopielenie → szczątków,
zniszczenie → planety.

- Byty, na które działa siła (tzw. “impactee”):

uderzenie → w górę,
uderzenie → góry.

- Byty postrzegane (stimulus procesów i stanów psychicznych):

postrzeganie → znaków.

- Obiekty posiadane:

posiadanie → majątku.

Wyniki na dwóch próbkach:

frazy NP		
PRECYZJA	KOMPLETNOŚĆ	MIARA F
$P = 16/20 = 80\%$	$R = 16/30 = 53\%$	$F = 64\%$

frazy AdjP		
PRECYZJA	KOMPLETNOŚĆ	MIARA F
$P = 3/4 = 75\%$	$R = 3/12 = 25\%$	$F = 38\%$

łącznie NP + AdjP		
PRECYZJA	KOMPLETNOŚĆ	MIARA F
$P = 19/24 = 79\%$	$R = 19/42 = 45\%$	$F = 57\%$

Wnioski: operatory mają wysoką precyzję, warto byłoby jednak poprawić kompletność. (Ważniejsze są frazy NP - częstsze w korpusie, AdjP to tylko 3% wszystkich fraz).

Operatory napisane są pod następujące schematy składniowe:

N/V/ → **Nob(gen)**

wybór dziewczyny ← wybrano dziewczynę
 krytyka posłów ← krytykują posłów
 pochwała Andrzeja ← pochwalono Andrzej

N_{OB} ← Adj/V/

oprawna książka ← książka, która została oprawiona
 uprawne pole ← pole, które ktoś uprawia
 dostawne krzesła ← krzesła, które dostawia się

N_{OB} ← Ppas, N_{OB} ← Pact

oprawiona książka
 uprawiane pole
 uprawiający pole

W sumie w korpusie KPWr operatory OBIEKTU uruchomiły się 487 razy.

operator(y)	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
OBIEKT	4.5%	16%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Przykład:

0:kasata 1:unii 2:brzeskiej 3:na 4:ziemiach 5:zabranych

7. Relacja dla pary (OD - DO): < 5 , 4 >:

OBIEKT-PPAS-0

wykryta dwójka: ziemiach ← zabranych [ktoś zabrał ziemię]

#WŁAŚCIWOŚĆ - to relacja pomiędzy rzeczownikiem oznaczającym byt a jego cechą, właściwością:

czerwony ← samochód
 nowe ← krzesło
 duży → las

Wyrażają tę relację w szczególności pary rzeczownik → przymiotnik, przy czym przymiotnik jest przymiotnikiem jakościowym. Relacja ta jest dość ogólna, dlatego chcielibyśmy wykrywać bardziej szczegółowe właściwości, np.

czerwony ← WŁ:KOLOR- samochód

nowe ←WŁ:UMIEJSCOWIENIE_W_CZASIE- krzesło
duży -WŁ:WIELKOŚĆ→ las

Relacja właściwości może być również wyrażana za pomocą wyrażen przyimkowych, np.:

samochód -WŁ:KOLOR-> o kolorze czerwonym

Na razie rozpoznajemy wyłącznie schematy składniowe **N + Adj**. Można dołączyć do tego później także schematy z PP.

RELACJA	P
WLASCIWOSC-0	P = 17/22 = 77%

Nie sprawdzaliśmy jeszcze kompletności - operatory można by próbować rozwijać. Mamy też pomysł na wykrywanie konkretnych właściwości (np. KOLORU, WIELKOŚCI, CZASU itp.).

W sumie w korpusie KPWr operator WŁAŚCIWOŚCI uruchomił się 1289 razy.

operator	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
WLASCIWOSC	12%	21%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbką kontrolną - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Przykład 4.

0:po 1:pięć 2:róż 3:czerwonych 4:ukośnie 5:w 6:krzyż

7. Relacja dla pary (OD - DO): < 3 , 2 >:

WLASCIWOSC-0

wykryta dwójka: czerwonych ← róż

#MIEJSCE

Rolę tę definiujemy następująco: miejsce jest to część przestrzeni, która stanowi arenę zdarzenia, stanu, czynności, a także - cel ruchu, lokalizację wyjściową akcji, ruchu, trasę, po której przebiega ruch.

ruch [gwiazd] → po niebie
przebywanie [Franka] → we Francji
wyjazd → z garażu

wycieczka → w góry
uciekający → w góry

Na razie wykrywamy tylko schematy **PRED + PP** (wyrażenie przyimkowe z miejscem). Operator jest na razie bardzo naiwny, tzn. korzysta wyłącznie z informacji o dziedzinie rzeczownika z PP (tj. akceptujemy wyłącznie miejsca i niektóre wytwory). Warto byłoby dołączyć do tego informację o ramie walencyjnej predykatu (często błędy polegały na tym, że obiekt był traktowany jak miejsce). Należy dodać do niego informacje dotyczące konkretnych przyimków (np. z+inst nie ma interpretacji MIEJSCE). Na razie tego nie uwzględniamy.

RELACJA	P
MIEJSCE próbka I	$P = 24/40 = 60\%$
MIEJSCE próbka II	$P = 17/27 = 63\%$

W sumie w korpusie KPWr operator MIEJSCA uruchomił się 136 razy.

operator	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
MIEJSCE	1.3%	6%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Przykład 5.

0:o 1:przygodności 2:naszego 3:życia 4:tu 5:na 6:ziemi

3.1.1 Relacja dla trojki (OD - prep - DO): < 3 , 5 , 6 >:
MIEJSCE-PP-0

znaleziona trójka: życia → na ziemi

#CZAS

Rola czasu lokalizuje nam zdarzenie w wymiarze czasowym, również informuje o długości określonego okresu:

ur. → w roku (1938)
zm. → na wiosnę
mieszkający → (trzeci) rok
opalający się → latem

RELACJA	P
CZAS	$P = 10/13 = 77\%$

Wykrywamy - podobnie jak w przypadku miejsc wyłącznie schematy **PRED + PP** oraz **PRED + N(inst)**. Operator ma małe pokrycie, trzeba by dodać nowe schematy składniowe.

W sumie w korpusie KPWr operator CZASU uruchomił się 53 razy.

operator	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
CZAS	0.5%	0.0%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Przykład 6.

0:były 1:szwedzki 2:biegacz 3:narciarski 4:uczestniczący 5:w
6:zawodach 7:w 8:latach 9:20 10:. 11:i 12:30

3.1.1 Relacja dla trojki (OD - prep - DO): < 4 , 7 , 8 >:
CZAS-PP-0

wykryta trójka: uczestniczący → w latach

#INSTRUMENT

Zwany również środkiem czynności, narzędziem. #Instrument to narzędzie, urządzenie i maszyna, za pomocą których wykonywana jest czynność, czyli przedmioty oznaczane przez rzeczowniki konkretne mające funkcje pomocnicze w akcji. Do instrumentów zaliczamy także pojęcia oderwane - typu 'argument' w 'argumentować'.

RELACJA	P
WLASCIWOSC-0	$P = 17/22 = 77\%$

Wykorzystujemy na razie schemat składniowy:

PREDYKAT + N(inst)

tj. predykat + rzeczownik konkretny w narzędniku. Operator jest b. naiwny.

W sumie w korpusie KPWr operator INSTRUMENTU uruchomił się tylko 15 razy.

operator	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
INSTRUMENT	0.1%	3%

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

Operatory zrobione przy okazji:

#PRZEZNACZENIE

Przeznaczenie - cel wykonania czynności, np. osoba, ku której skierowana jest akcja. W literaturze zachodniej tej roli odpowiada *Goal*:

“Individual toward which the event is directed” - Larson, Segal 1996: 479

Wykorzystujemy tu na razie tylko jeden schemat składniowy:

dla + N₁(gen)

Operator jest b. prosty - można go jeszcze rozwijać.

RELACJA	P
PRZEZNACZENIE-0	P = 24/37 = 65%

Łącznie operator przeznaczenia uruchomił się na korpusie 68 razy, co daje 0.6% potencjalnych relacji.

#POKREWIEŃSTWO

Wyraża pokrewieństwo pomiędzy osobami. Wykorzystujemy tu schematy składniowe:

N₁ + N₂(gen)

jego/jej/ich + N₁

Adj_{POSS} + N₂

gdzie N₁ to nazwa krewnego lub innego członka rodziny [ze Słownosieci], a N₂ to nazwa osoby [dziedzina “os” lub imię, lub ppron3], a Adj_{POSS} to zaimek dzierżawczy [lematy “mój”, “twój”, “swoj”, “nasz”, “wasz”].

RELACJA	P
WLASCIWOSC-0	$P = 12/13 = 92\%$

Relacja ta jest rzadka, ale wydaje się, że może być rozpoznawana z dużą precyzją, a może mieć istotne zastosowania. W korpusie wykryta została tylko 13 razy..

operator	POKRYCIE OPERATORA [przybliżone*]	PRÓBKA KONTROLNA - POKRYCIE [relacje opisane ręcznie]
WLASCIWOSC	0.1%	nieznane

*) Dzielimy liczbę tych przypadków, dla których uruchomił się operator, przez liczbę wszystkich relacji wykrytych w kroku 1. (generowanie dwójek i trójek). Próbka kontrolna - to ręcznie zliczone wystąpienia danej relacji semantycznej (podajemy ją tutaj dla porównania).

#MERONIMIA-MIEJSCA

Relacja ta zachodzi pomiędzy dwoma miejscami, z których jedno jest częścią drugiego. Wykorzystujemy tu schemat składniowy z dopełniaczem.

RELACJA	P
MERONIMIA-MIEJSCA	$P = 30/45 = 67\%$

Operator uruchomił się 67 razy na korpusie.

#KOLEJNOSC

Podaje, który z kolei jest byt określany przez rzeczownik. Przyjmujemy, że łącznie traktujemy kolejność w sensie przestrzennym, czasowym, jak i kolejność wynikającą z rankingu. Wykorzystujemy tu schemat składniowy:

pierwszy, drugi, trzeci, ... + N_1

RELACJA	P
KOLEJNOSC	$P = 25/35 = 71\%$

Za błędy uznawaliśmy częste połączenia *pierwsza + liga*.

W korpusie operator uruchomił się 105 razy, co daje ~1% wszystkich potencjalnych relacji.

#ILOSC

Relacja #ilości łączy rzeczownik z określeniem liczności. #Ilość mówi nam, jak liczny jest zbiór (5 *krzesel*) albo jaka jest ilość czegoś niepoliczalnego (*hektolitry wody*), do tej relacji zaliczamy także nieokreślone określenia ilości (*wiele ludów*, *liczba stron*). Wykorzystujemy schemat składniowy:

Num + N_i(gen)
N_i("il") + N_i(gen)

Gdzie "il" oznacza dziedzinę Słownosieci "il", Num to liczebnik.

Jest to dość powszechna relacja w korpusie, łącznie pojawiła się 437 razy w korpusie, co daje 4,1% wszystkich potencjalnych relacji w korpusie.

RELACJA	P
ILOSC	P = 34/45 = 76%

#KWANTYFIKATOR-A/#KWANTYFIKATOR-E

Operator wskazuje na możliwą kwantyfikację. Wykrywamy występowanie kwantyfikatora ogólnego (A) i szczegółowego (E). Wykorzystujemy schematy

kwantyfikator ogólny (A):

każdy + N_i

wszystkie + N_i

kwantyfikator szczegółowy (E)

"któryś", "jakiś", "jeden", "niektóry", "którykolwiek", "jakikolwiek" + N_i

jeden + z + N_i

RELACJA	P
KWANTYFIKATOR-A	P = 62/62 = 100%

RELACJA	P
KWANTYFIKATOR-E	P = 20/24 = 83%

#KWANTYFIKATOR-A-neg

Wykrywamy tu kwantyfikator ogólny razem z negacją (*żaden z x-ów*, *żaden x*). Wykorzystujemy tu schemat składniowy:

żaden/nikt + z + N_i

żadne + N_i

RELACJA	P
KWANTYFIKATOR-A-neg	P = 22/22 = 100%

#DESKRYPCJA_OKREŚLONA

Sprawdzamy występowanie zaimków wskazujących jako wyznaczników deskrypcji określonej. Wykorzystujemy schemat składniowy:

ten/ów + N_i

RELACJA	P
DESKRYPCJA_OKREŚLONA	P = 111/111 = 100%

Łącznie w korpusie operator uruchomił się 229 razy, co daje 2% wszystkich potencjalnych relacji.

Pokrycie operatorów na całym korpusie KPWr z dnia 8.03.2013

Sprawdziliśmy, jak często uruchamiał się dany operator na potencjalnych relacjach. Wyniki umieściliśmy w poniższej tabeli.

RELACJA - kpwr_all	liczność	pokrycie operatora	P*
WLASCIWOSC	3248	11.6%	77%
OBIEKT	1568 ->1792	5.6%	79%
MIEJSCE	1312	4.7%	61.5%
ILOSC	1074	3.8%	76%
DETERMINATOR (DESKRYPCJA)	550	2.0%	100%
KWANTYFIKATOR	340	1.2%	83-100%
SUBIEKT	310 ->428	1.1%	95%
PRZEZNACZENIE	230	0.8%	65%
MERONIMIA MIEJSCA	186	0.7%	67%
CZAS	176	0.6%	77%
INSTRUMENT	32	0.1%	77%
POKREWIEINSTWO	28	0.1%	92%
ALL	28118	---	---

P* - precyzja na próbce tego samego korpusu.

RELACJE POZOSTAŁE

Różnego typu relacje opisaliśmy w dokumencie “relacje w obrębie frazy NP”, tu podajemy tylko te, które nie mają operatorów, ale które trzeba było zdefiniować, żeby właściwie ocenić próbki.

#Przyczyna

Bliską semantycznie subiekтови-agensowi rolą jest rola przyczyny (*Cause*). Definiuje się ją często za pomocą cechy ‘powodować (jakiś stan, zmianę stanu)’. Przyczyna jest to byt, zjawisko nieożywione, działające w sposób czynny, powodujące jakąś zmianę lub jakiś stan, występuje przy predykatkach z elementem semantycznym ‘powodować’ (‘CAUS’).

spowodowane → przez wiatr

#Rezultat

Rezultat definiujemy następująco - jest to zdarzenie mające swoją przyczynę lub efekt czynności przyczynowej. Jest rezultatem przedmiot, który powstaje w wyniku działania przyczyny na obiekty fizyczne (materiały). Rezultat występuje przy predykatkach z elementem semantycznym ‘powodować’ (‘CAUS’).

zniszczenia ← spowodowane (przez wiatr, człowieka)

Rezultatem jest także wytwór, tj. przedmiot fizyczny powstający w toku akcji. Nie jest wytworem przedmiot fizyczny, który ulega zmianie w toku akcji (taki byt nazywamy obiektem), tylko byty fizyczne powstające w toku akcji mogą być nazywane wytworami, występuje przy predykatkach z elementem semantycznym ‘powodować’ (‘CAUS’).

młot ← wykuty (przez kowala)

#Material

Obiekt materiałowy - materiał, z którego dany przedmiot (wytwór) ma być wykonany.

drewniany → stół
stół ← z drewna

4. Literatura

1. Radziszewski, Adam, *A tiered CRF tagger for Polish*, Springer Verlag, 2013
2. Radziszewski, Adam; Pawlaczek, Adam, *Large-scale experiments with NP chunking of Polish*, Proceedings of TSD 2012, 2012
3. Broda, Bartosz; Marcińczuk, Michał; Maziarz, Marek; Radziszewski, Adam; Wardyński, Adam, *KPWr: Towards a Free Corpus of Polish*, Proceedings of LREC'12, 2012