Computer Science. Vol. 3. 2001 004.8:025.4:004.424

*Grażyna Szostek**

SYNTAKTYCZNE METODY ROZSTRZYGANIA WIELOZNACZNOŚCI FORM FLEKSYJNYCH

1. Przełom w automatycznym wyszukiwaniu informacji tekstowej

Na Uniwersytecie Harvarda w latach 1961–64 powstał automatyczny system wyszukiwania dokumentów SMART [13]. W przeciwieństwie do innych zautomatyzowanych systemów wyszukiwania informacji tekstowych, SMART nie korzysta z tradycyjnie wyznaczonych słów kluczowych czy indeksów służących do identyfikacji dokumentów ani też nie kieruje się częstością występowania pewnych słów lub zdań zawartych w tekstach tych dokumentów. SMART ma dostęp bezpośrednio do tekstu dokumentu i analizę treści przeprowadza z wykorzystaniem pewnych intelektualnych narzędzi w rodzaju: słowników synonimów, hierarchicznych katalogów przedmiotowych, statystycznych i syntaktycznych metod generowania fraz itp. Wyniki wyszukiwania dokumentów były lepsze w porównaniu do innych systemów, ale dalej nie były w pełni zadowalające.

W systemie SMART mamy do czynienia z analizą językową, która jest składową procesu indeksacji. Analiza przebiega w sposób mechaniczny i jej jakość w dużym stopniu zależy od sposobu budowy słowników. Z góry jest ustalone (dla danej dziedziny tematycznej), jakie słowa są nieważnie, które są pospolite, które są ważne. Takie, a nie inne zakwalifikowanie wyrazów wynika ze statystyki ich występowania w tekstach o ustalonej tematyce i odpowiednich reguł budowy słowników (ustalenie liczby klas pojęć, ogólności lub szczegółowości klas itd.). Nie ma mechanizmów rozstrzygających wieloznaczność wyrazu. Takim wyrazom nawet obniża się wagę w stosunku do słów jednoznacznych. Pomijane są wartości leksykalne związane z danym wyrazem. Wyrazy o wspólnym temacie są równoważne [11].

* Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Akademia Górniczo-Hutnicza, Kraków

2. Wieloznaczność – jeden z podstawowych problemów przetwarzania tekstu

Człowiek wykorzystuje język do wyrażania i zapisywania myśli. Kiedy wiemy, co chcemy powiedzieć lub zapisać, wybór wyrazów i form zdaniowych nie jest dość trudny. Wypowiedź lub tekst są traktowane przez autora jako coś zrozumiałego i jednoznacznego dla wszystkich. Okazuje się, że my jako odbiorcy często stajemy przed dylematem: co autor chciał przez to powiedzieć. Nie mając możliwości dialogu z autorem tekstu, momenty wieloznaczne występujące w nim tłumaczymy na swój sposób uznając go za słuszny. Nasz obraz myślowy powstały po przeczytaniu tekstu jest nieco odmienny od tego, który chciał przekazać autor. W życiu codziennym takie sytuacje grożą konfliktami i nieporozumieniami. Wyjaśnienie następuje w wyniku dialogu. Dzięki zadawaniu pytań otrzymujemy informacje umożliwiające zrozumienie tekstu w sposób jednoznaczny.

Problem wieloznaczności wyrażeń nie występuje w językach formalnych. Tworzenie zdań w tych językach jest opisane za pomocą reguł, a każde zdanie może mieć tylko jedno znaczenie. Przy określeniu znaczenia zdania (wybraniu reguły użytej do jego utworzenia) pomocne są słowa kluczowe. W ten sposób konstruktorzy języków programowania zyskują pewność, że napisany za ich pomocą program komputerowy będzie działał w sposób deterministyczny.

Powstaje zatem problem zbudowania systemu formalnego, który w sposób jednoznaczny będzie interpretował wyrażenia języka naturalnego. Jest to zadanie trudne, gdyż wieloznaczność w języku naturalnym ma kilka różnych przyczyn. Poniżej pokazane są typy wieloznaczności, z którymi taki system musi sobie poradzić. W pracy przedstawimy algorytm (system formalny) rozwiązujący jeden typ wieloznaczności – wieloznaczność syntaktyczną.

2.1. Wieloznaczność leksykalna – problem rozpoznawania wyrazu w tekście

Wieloznaczność leksykalna ma miejsce, kiedy dwa lub więcej wyrazów ma identyczną formę.

*Nasze **drogi** rozchodziły się w różnych kierunkach.*

*Jutro przyjeżdża **drogi** mój przyjaciel.*

W obu przypadkach przy analizie leksykalnej wyrazu *drogi* otrzymamy:

- *droga* – rzeczownik, liczba mnoga, rodzaj żeński itd.,
- *drogi* – przymiotnik, liczba pojedyncza, rodzaj męski itd.

2.2. Wieloznaczność syntaktyczna – problem określenia funkcji wyrazu

Z wieloznacznością syntaktyczną mamy do czynienia, kiedy dwie formy tego samego wyrazu są identyczne.

***Okna** wieżowca błyszczały odbijając słoneczne promienie.*

*W szarych kłębach dymu próbowałem dostać się do **okna**.*

W obu przypadkach analiza leksykalna poda dla formy *okna* dwie informacje: liczba pojedyncza, dopełniacz i liczba mnoga, mianownik.

2.3. Wieloznaczność semantyczna i uzupełnianie informacji

Struktura semantyczna zdania składa się z kilku elementów, takich jak: akcja, sprawca akcji, obiekt podlegający akcji, źródło i cel akcji (kierunek, w którym akcja jest zorientowana OD – DO) [14, 16], np.:

Janek przyjechał z Krakowa do Warszawy.

sprawca akcji		Janek
akcja		jechać
obiekt		Janek
kierunek	DO	Warszawa
kierunek	OD	Kraków

Poniżej podane są dwa przykłady wieloznaczności powstałej w wyniku braku niektórych z tych elementów.

Określenie sprawcy akcji

Prowadziła *nas do samej granicy.*

Sprawcą akcji może być istota żywa (człowiek albo zwierzę) albo obiekt nieożywiony i konkretny np.: droga lub ścieżka.

Ciekawym przykładem jest powiązanie czasownika w następującym zdaniu

Pod wieczór ***zachmurzyło się****.*

Moglibyśmy rozważyć dwa przypadki:

1) czasownik *zachmurzyło się* jest bezosobowy, więc nie może być mowy o sprawcy akcji;
2) szukamy w sąsiednich zdaniach rzeczownika rodzaju nijakiego w liczbie pojedynczej.

Okazuje się, że ten problem jest urojony. Nic innego nie może się zachmurzyć – tylko *niebo* (nie rozważamy frazeologizmów).

Określenie obiektu akcji

Zobacz jak one jedzą.

Nie określony w zdaniu obiekt może być jednym z dwóch typów:

1) pokarm – obiad, kanapkę, siano itd.;
2) coś żywego – mnie, nas, konia, krowę.

2.4. Wielkość kontekstu a problem wieloznaczności

Analizując pojedyncze zdanie staramy się wydobyć i wywnioskować jak najwięcej informacji dotyczących poszczególnych słów i ich powiązań. Staramy się ustalić znaczenie dla słów wieloznacznych, powiązać zaimki z obiektami, wstępnie ustalić, które wyrazy są ważne lub nieważne itd. Często rozwiązanie niektórych z powyższych problemów nie jest możliwe bez analizy kontekstu w postaci poprzedniego lub następnego zdania.

Np. mając do czynienia ze zdaniem

Myślał o trzech ***technikach***.

nie jesteśmy w stanie stwierdzić, w jakim znaczeniu występuje wyraz *technikach*. Może to być forma wyrazu *technikum*, *technik* lub *technika*. Dopiero analiza sąsiednich zdań, a jeśli zaistnieje konieczność – to i analiza kontekstu akapitu, a nawet całego tekstu, może rozwiązać ten i inne problemy.

W sytuacji kiedy po analizie kontekstu problem wieloznaczności nadal pozostaje nierozwiązany, pomocna staje się analiza treści na podstawie zgromadzonej wiedzy o świecie (tj. wiedzy o obiektach, zdarzeniach) poprzez ustalenie celów, planów, intencji uczestników zdarzeń. Jeśli i teraz nie mamy jednoznacznej odpowiedzi, to stawiamy kilka hipotez: jedne bardziej prawdopodobne, drugie – mniej. W wyniku dialogu z użytkownikiem możemy uzyskać dodatkowe informacje, które w sposób jednoznaczny umożliwią ustalenie znaczenia wypowiedzi.

3. Syntaktyczne metody rozstrzygania wieloznaczności

Wydawałoby się, że zdań jest tyle, ile kombinacji wyrażeń, tzn. nieograniczona ilość. Okazuje się, że rzeczywistość jest inna. „Budując określone wypowiedzenia, jesteśmy ograniczeni w wyborze wyrażeń. Kombinacje wyrazowe są z jednej strony uwarunkowane czynnikami ściśle semantycznymi, takimi jak sensowność czy bezsensowność określonych połączeń i zgodność z potrzebą wyrażania odpowiedniej treści, z drugiej zaś strony zależą od zasad łączliwości wyrazowej" [9]. Istnienie pewnych zasad i uwarunkowań przy tworzeniu kombinacji wyrązowych powoduje, że można je zapisać w postaci schematów. Kluczowym elementem schematu jest czasownik. Swoimi właściwościami semantyczno-gramatycznymi decyduje w znacznej mierze o strukturze zdania, w którym występuje. *Słownik syntaktyczno-generatywny czasowników polskich* pod red. K. Polańskiego zawiera wszystkie czasowniki i powiązane z nimi schematy zdaniowe, w których mogą one występować. Budowa schematów opiera się na koncepcji gramatyk formalnych, których rozwój zapoczątkował N. Chomski. To, że „wybór wyrazów i ich forma w procesie tworzenia wypowiedzeń w znacznej mierze zależą od innych wyrażeń, z którymi wiążą się one gramatycznie i semantycznie", powoduje, że słownik może być bardzo przydatnym narzędziem przy rozwiązaniu takich problemów semantycznych, jak określenie brakujących wyrazów w zdaniu lub znaczenia wyrazów wieloznacznych.

3.1. Budowa słownika i objaśnienia

Słownik zawiera wszystkie czasowniki polskie. Dla każdego z nich podane są schematy zdaniowe, w których może on występować.

3.1.1. Schematy zdaniowe

Poszczególne pozycje w zdaniu (podmiot, dopełnienie, okoliczniki itp.) mogą zajmować tylko elementy należące do określonych klas leksykalno-semantycznych. Ale to nie wystarczy do scharakteryzowania łączliwości poszczególnych czasowników, trzeba także zróż-

nicować łączliwość obligatoryjną i fakultatywną. Łączliwość obligatoryjna dotyczy składników, które muszą wystąpić przy danym czasowniku. Łączliwość fakultatywna dotyczy składników, które mogą, ale nie muszą być użyte z danym czasownikiem.

Przykładowo, schemat zdaniowy podany przy haśle *biczować*

$$NP_N - NP_{Acc} + (NP_I)$$

oznacza, że czasownik łączy się obowiązkowo z rzeczownikiem w bierniku, natomiast rzeczownik w narzędniku może wystąpić przy nim lub nie.

Cechy semantyczne podane są w nawiasach kwadratowych, znak + oznacza występowanie danej cechy, natomiast znak – jej brak, np. [+Anim] = żywotność, [–Anim] = nieżywotność, [+Hum] = osobowość, [–Hum] = nieosobowość itp. Obok cech powszechnie uznanych wprowadzone są też takie rozróżnienia, jak rzeczowniki oznaczające rośliny, żywioły, informację itp.

Ze względu na to, że nie udało się sformułować definicji pozwalających w sposób jednoznaczny odróżnić dopełnienie od okolicznika, zrezygnowano z terminologii części zdań na rzecz rozróżnień kategorialnych typu: fraza nominalna (= grupa rzeczownikowa), fraza bezokolicznikowa (= wyrażenie bezokolicznikowe), przysłówek.

Fraza nominalna może pełnić w zdaniu różne funkcje – od podmiotu począwszy, a na okoliczniku skończywszy. Przy podmiocie i dopełnieniach subskrypty odnoszą się do przypadku (= fraza nominalna w nominatywie itd.), a przy okolicznikach określają funkcje (np. akcesoryjną, ablatywną, adlatywną itd.).

Każde z haseł dzieli się na kilka części.

- Wyraz hasłowy w formie bezokolicznika, np. biczować, brać itp.
- Schemat zdaniowy, zawierający symbole kategorialne połączone znakiem + względnie znakiem konkatenacji ∩. Kolejność składników połączonych znakiem + nie implikuje aktualnego szyku wyrazów w zdaniu, tylko znak konkatenacji ∩ oznacza obowiązkowy szyk elementów odpowiadający rzeczywistej kolejności występowania w zdaniu.

  Miejsce czasownika w schemacie oznaczone jest poziomą kreską –. Ponieważ podmiot najczęściej poprzedza czasownik, a dopełniacz i okoliczniki następują po nim, dlatego w schemacie przyjęto kolejność $NP_N - NP_{Acc}$.
- Charakterystyka semantyczna składników nominalnych służy tylko do określenia ograniczeń łączliwościowych. Cechy te przytacza się po strzałce wychodzącej ze skrótu symbolizującego daną frazę, np.:

  $NP_N \rightarrow$ [+Anim] (= fraza rzeczownikowa w mianowniku zawierajaca rzeczownik żywotny),

  $NP_N \rightarrow$ [–Abstr] (= fraza rzeczownikowa w bierniku zawierająca rzeczownik konkretny).

  Jeśli dana fraza charakteryzuje się łącznie dwoma cechami, to występują one w kolumnie pionowej

  $NP_I \rightarrow \begin{bmatrix} +Anim \\ +Pars \end{bmatrix}$ (= fraza rzeczownikowa zawierająca rzeczownik oznaczający część ciała; część ciała oznaczona tu została cechami [+Anim] = żywotność oraz [Pars] = część).

Jeśli w danej pozycji może wystąpić kilka fraz, przy czym każda z nich charakteryzuje się innymi cechami semantycznymi, przytacza się te cechy w szyku poziomym, np.

$NP_1 \rightarrow [Instr]\begin{bmatrix}+Anim \\ Pars\end{bmatrix}$.

Powyższy zapis oznacza, że w pozycji może wystąpić albo rzeczownik oznaczający narzędzie (np. *brać **wiadrem** wodę*) albo część ciała (np. *wziąć coś **zębami***).

3.1.2. Objaśnienia

Skróty i symbole

NP – fraza nominalna (grupa rzeczownikowa),

NP_{Abl} – fraza nominalna ablatywna (tj. oznaczająca kierunek oddalania się, np. *z domu*),

NP_{Adl} – fraza nominalna adlatywna (tj. oznaczająca kierunek zbliżania się, np. *do domu*),

NP_{Perl} – fraza nominalna perlatywna (tj. oznaczająca miejsce, przez które odbywa się ruch, np. *przez las*),

NP_{Mod} – fraza nominalna sposobowa

$NP_{N, G, D, A, I, L}$ – litery u dołu fraz nominalnych wskazują ich przypadek graniczny (= nominativus, genetivus, dativus, accusativus, instrumentalis, locativus).

Cechy semantyczne

[+Abstr] – abstrakcyjność, oderwaność,

[–Abstr] – konkretność,

[+Anim] – żywotność,

[–Anim] – nieżywotność, nieożywioność,

[+Hum] – osobowość,

[–Hum] – nieosobowość,

[Fl] – roślina,

[Pars] – część,

[Mach] – maszyna,

[Inf] – informacja.

Przykłady kombinacji cech semantycznych:

$\begin{bmatrix}-Abstr \\ -Anim\end{bmatrix}$ – konkret nieożywiony (np. *kamień, szkło*),

$\begin{bmatrix}+Hum \\ +Pars\end{bmatrix}$ – część ciała ludzkiego (np. *ręka, ząb, głowa*).

Znaki

– – pozycja czasownika hasłowego w schemacie zdaniowym,

\+ – między składnikami oznacza ich łączenie bez implikacji szyku w aktualnym zdaniu,

$\frown$ – konkatenacja (znak implikujący szyk składników),

( ) – fakultatywność składników lub grupy składników (tj. możliwość ich pominięcia),

$\left\{\frac{x}{y}\right\}$ – wymienność składników w danej pozycji,

$\rightarrow$ – strzałka odsyła do charakterystyki semantycznej za pomocą cech,

$\{x\ y\}$ – obowiązkowe występowanie w danej pozycji przynajmniej jednego spośród składników znajdujących się w takich nawiasach.

4. Algorytm

Algorytm ma za zadanie: znaleźć w słowniku schemat zdaniowy dla danego wypowiedzenia, wygenerować oczekiwania w odniesieniu do brakujących składowych i określić znaczenie wyrazów wieloznacznych w kontekście danego wypowiedzenia, przy czym ograniczymy się do jednego typu wieloznaczności – wieloznaczności leksykalnej.

Pierwszy etap polega na analizie kolejnych wyrazów zdania. Szczególną uwagę zwracamy na czasowniki i rzeczowniki. Czasownik pozwoli dotrzeć do właściwych schematów, a rzeczowniki pomogą wybrać jeden z nich (jeśli zdanie jest pełne i nie zawiera wyrazów wieloznacznych). Każdy wyraz przechodzi przez analizę leksykalną. Dla rzeczownika dodatkowo zostaje określona klasa lub klasy semantyczne. Po przeczytaniu czasownika zwracamy się do słownika syntaktyczno-generatywnego. Słownik podaje dla każdego czasownika listę schematów zdaniowych, w których dany czasownik może się pojawić. Dokonujemy wstępnego wyboru schematów z listy wykorzystując przy tym już przeczytane rzeczowniki. Po przeczytaniu pozostałych wyrazów zdania podejmujemy ostateczną decyzję co do właściwości wybranych schematów.

Drugi etap polega na rozwiązywaniu niejednoznaczności. Jeśli pierwszy etap zakończył się na wybraniu z listy jednego schematu i wszystkie jego składowe występują w zdaniu, to algorytm kończy działanie (przykład 1). Brak w zdaniu pewnych grup rzeczownikowych wymaganych przez schemat nie oznacza, że wybraliśmy niewłaściwy schemat. Taka sytuacja może wystąpić, kiedy rzeczownik został już wspomniany w poprzednich zdaniach lub będzie o nim mowa w kolejnych zdaniach. W celu ustalenia, jaki to jest rzeczownik, w pierwszym przypadku trzeba przeprowadzić pewną analizę semantyczną dotychczasowych wyników, w drugim – wygenerować oczekiwanie.

Problemy pojawiają się, gdy po przeczytaniu całego zdania mamy więcej niż jeden dopasowany schemat. Jak je rozwiązać, pokażę na przykładach.

Przykład 1

Ścieżka biegła w głąb lasu.

Słownik syntaktyczno-generatywny: **biec – biegnąć**

a) $NP_N - \{(NP_{Abl}) + (NP_{Adl}) + (NP_{Perl})\}$,
$NP_N \rightarrow [+Anim]$;

b) $NP_N - \{(NP_{Abl}) + (NP_{Adl}) + (NP_{Perl})\}$,
$NP_N \rightarrow \begin{bmatrix} +Abstr \\ -Anim \end{bmatrix}$;

c) $NP_N - (NP_{Mod})$,
$NP_N \rightarrow [+Abstr]$.

ETAP I

ścieżka: *ścieżka* – rzeczownik, rodzaju żeńskiego, w liczbie pojedynczej, w mianowniku; klasa semantyczna – $\begin{bmatrix} -Abstr \\ -Anim \end{bmatrix}$.

Rzeczownik *ścieżka* (NP_N) i czasownik biec implikują wybór schematu b):

$$NP_N - \{(NP_{Abl}) + (NP_{Adl}) + (NP_{Perl})\}.$$

W głąb lasu jest frazą adlatywną (NP_{Adl}), oznaczającą kierunek zbliżania się. Taka fraza występuje w schemacie, co potwierdza prawidłowość wyboru. Pierwszy etap kończy się na wybraniu jednego schematu, którego wszystkie elementy występują w zdaniu (nawiasy {} oznaczają, że przynajmniej jeden składnik ma wystąpić w zdaniu), więc algorytm kończy działanie.

Przykład 2

Adwokat bronił pisklęta.

Słownik syntaktyczno-generatywny: **bronić**

a) $NP_N - NP_G + (NP_I^1) + (przed \cap NP_I^2)$,
$NP_N \rightarrow [+Hum]$,
$NP_G \rightarrow [-Abstr]$,
$NP_I^1 \rightarrow \begin{bmatrix} +Hum \\ +Pars \end{bmatrix}$ [broń],
$NP_I^2 \rightarrow [+Anim][+Abstr][Mach]$;

b) $NP_N - NP_G + (NP_I^1) + (przed \cap NP_I^2)$,
$NP_N \rightarrow \begin{bmatrix} +Anim \\ -Hum \end{bmatrix}$,
$NP_G \rightarrow [+Anim]$,
$NP_I^1 \rightarrow \begin{bmatrix} +Anim \\ +Pars \end{bmatrix}$,
$NP_I^2 \rightarrow [+Anim][+Abstr]$;

c) $NP_N - NP_G + (NP_I^1) + (przed \cap NP_I^2)$,
$NP_N \rightarrow [+Hum]$,
$NP_G \rightarrow [+Hum][+Abstr]$,
$NP_I^1 \rightarrow [Inf]$,
$NP_I^2 \rightarrow [+Hum][+Abstr]$;

d) $NP_N - NP_D + \begin{Bmatrix} NA_G \\ IP \end{Bmatrix}$,
$NP_N \rightarrow [+Hum]$,
$NP_D \rightarrow [+Hum]$.

Powyżej przedstawione zostały tylko te schematy, które będą uczestniczyły w analizie (w celu zwiększenia przejrzystości).

ETAP I

adwokat:

1) *adwokat* – rzeczownik, rodzaju męskiego, w liczbie pojedynczej, w mianowniku; klasa semantyczna – $\begin{bmatrix}+\text{Anim}\\-\text{Hum}\end{bmatrix}$,

2) *adwokat* – rzeczownik, rodzaju męskiego, w liczbie pojedynczej, w mianowniku; klasa semantyczna – [+Hum].

Rzeczownik *adwokat* (NP_N) może oznaczać ptaka lub zawód człowieka. Wybieramy wszystkie schematy, gdzie ma podane powyżej klasy semantyczne. Trzeba uwzględnić, że

$$\begin{bmatrix}+\text{Anim}\\-\text{Hum}\end{bmatrix} \subset [\text{Anim}].$$

Wszystkie przedstawione w słowniku schematy zakwalifikowały się do dalszej analizy.

Czytamy kolejny wyraz w zdaniu: *piskleta*.

piskleta: *pisklę* – rzeczownik, rodzaju nijakiego, w liczbie mnogiej, w dopełniaczu; klasa semantyczna – $\begin{bmatrix}+\text{Anim}\\-\text{Hum}\end{bmatrix}$.

Wśród wybranych w poprzednim kroku schematów wybieramy te, które zawierają frazę nominalną w dopełniaczu o podanej klasie semantycznej:

- $NP_N - NP_G + (NP_1^1) + (\text{przed} \cap NP_1^2)$,
- $NP_N - NP_G + (NP_1^1) + (\text{przed} \cap NP_1^2)$.

Zdanie zostało przeczytane w całości. Pierwszy etap kończy się na wybraniu dwóch schematów, więc przechodzimy do etapu drugiego.

ETAP II

Zdanie możemy uznać za kompletne. Wszystkie wymagane przez schematy składniki występują w zdaniu. Ale są też składniki fakultatywne, które mogły zostać pominięte. W odniesieniu do nich generujemy oczekiwania:

$$NP^1 \rightarrow \begin{bmatrix}+\text{Hum}\\+\text{Pars}\end{bmatrix} [\text{broń}],$$

$$NP^2 \rightarrow [+\text{Anim}][+\text{Abstr}][\text{Mach}],$$

$$NP^3 \rightarrow \begin{bmatrix}+\text{Anim}\\+\text{Pars}\end{bmatrix},$$

$$NP^4 \rightarrow [+\text{Anim}][+\text{Abstr}].$$

1) Algorytm nie rozstrzygnął, jakie znaczenie ma w zdaniu wyraz *adwokat*, ale i człowiek miałby trudności z określeniem, czy to jest człowiek, czy ptak. Jak wskazują wygenerowane oczekiwania, wystarczy jedna informacja, czym adwokat bronił – i problem wieloznaczności zostanie rozwiązany. Oczekiwania 2 i 4 dotyczą różnych schematów, ale są w zasadzie takie same. Odnalezienie w kolejnych lub poprzednich zdaniach odpowiadających im wyrazów nie wniosłoby niczego nowego do analizy. Nie wiedzielibyśmy, do którego schematu odnieść odnaleziony wyraz. Wydaje się, że takich oczekiwań nie trzeba generować.

Ostateczna wersja oczekiwań jest taka:

$$NP^1 \rightarrow \begin{bmatrix} +\text{Hum} \\ +\text{Pars} \end{bmatrix} [\text{broń}],$$

$$NP^2 \rightarrow [\text{Mach}],$$

$$NP^3 \rightarrow \begin{bmatrix} +\text{Anim} \\ +\text{Pars} \end{bmatrix}.$$

2) W przykładzie 2 wyraz wieloznaczny ma dwa znaczenia i dwie klasy semantyczne. To nam ułatwia rozróżnienie jednego znaczenia od drugiego. Są wyrazy o kilku znaczeniach, ale o tej samej klasie semantycznej. Nasz algorytm jest najmniej skuteczny wobec wieloznaczności dotyczącej takich wyrazów.

Możemy podejść do tego problemu z innej strony. Człowiek rozróżnia znaczenia wyrazu np. *zamek* poprzez analizę najbliższego kontekstu. Wiadomo, że jeśli mówimy o *zamku królewskim*, to nie powiemy *zasunął (naoliwił, zamontował) zamek.* Ale dla *zamka błyskawicznego* (*zamka do drzwi*) to jest poprawna fraza. Wybór czasownika determinuje znaczenie wyrazu powiązanego z nim. Zbudujemy dla wyrazów wieloznacznych pewną strukturę wiążącą różne znaczenia wyrazu i dobierzmy czasowniki, z którymi mogą one występować w tym znaczeniu.

Przykładowo:

zamek

klasa semantyczna $\begin{bmatrix} -\text{Anim} \\ -\text{Abstr} \end{bmatrix}$;

– *zamek*: zbudować, oblegać, bronić, stawić, stać, remontować, zwiedzać, zamknąć;

– *zamek*: zamknąć, zatrzasnąć, przykręcić, otworzyć, oliwić, naprawić;

– *zamek*: wszyć, rozsunąć, zasunąć, rozpiąć, zapiąć.

Niektóre czasowniki mogą być powiązane z kilkoma znaczeniami wyrazu wieloznacznego.

***Zamek** zamknięty* dla zwiedzających.

Marysia *zamknęła **zamek*** na klucz.

Ustalenie powiązania między *zamek* a *zamknięty* nie daje odpowiedzi na pytanie, o jaki zamek chodzi w zdaniu. Dla takich przypadków możemy wprowadzić dodatkowe elementy pozwalające rozróżnić znaczenia. Mogą to być przymiotniki, rzeczowniki, klasy semantyczne:

– *zamek*: królewski, kamienny, obronny, drewniany; służba, ochrona, turyści, król;

– zwiedzający; fraza adlatywna, ablatywna (do zamku, z zamku);

– *zamek*: zapadkowy, bębenkowy, elektryczny; [klucz], [kod];

– *zamek*: błyskawiczny; [ubrania].

Przykład 3

Wieloznaczność leksykalna.

Drogi przyjaciel prowadził nas do sukcesu.

Słownik syntaktyczno-generatywny: **prowadzić**

a)

NP_N – do $\cap NP_G + (NP_I^1) + (NP_{Acc})$,

$NP_N \rightarrow$ [+Hum],

$NP_G \rightarrow$ [+Abstr],

$NP_I \rightarrow$ [+Abstr],

$NP_{Acc} \rightarrow$ [+Hum];

b)

$NP_N - NP_{Acc} + (NP_{Mod})$,

$NP_N \rightarrow$ [+Hum],

$NP_{Acc} \rightarrow$ [Mach] [Koń];

c)

$NP_N - NP_{Acc} + z \cap NP_I$,

$NP_N \rightarrow$ [+Hum],

$NP_{Acc} \rightarrow$ [+Abstr],

$NP_I \rightarrow$ [+Hum];

d)

$NP_N - NP_{Acc}$,

$NP_N \rightarrow$ [+Hum],

$NP_{Acc} \rightarrow$ [+Abstr] [Instit];

e)

$NP_N - \{(NP_{Adl}) + (NP_{Abl})\}$,

$NP_N \rightarrow \begin{bmatrix} -\text{Anim} \\ -\text{Abstr} \end{bmatrix}$;

f)

$NP_N -$,

$NP_N \rightarrow$ [+Hum];

g)

$NP_N - NP_{Acc}$,

$NP_N \rightarrow$ [+Hum],

$NP_{Acc} \rightarrow$ [+Fl];

h)

$NP_N - NP_{Acc} + (NP_{Mod})$,

$NP_N \rightarrow$ [+Hum],

$NP_{Acc} \rightarrow$ [+Hum];

i)

$NP_N - NP_{Acc} + (NP_{Adl}) + (NP_{Perl}) + (NP_{Abl})$,

$NP_N \rightarrow$ [+Anim],

$NP_{Acc} \rightarrow$ [+Anim].

Powyżej przedstawione zostały tylko te schematy, które będą uczestniczyły w analizie (w celu zwiększenia przejrzystości).

ETAP I

drogi:

1) *drogi* – przymiotnik, rodzaju męskiego, w liczbie pojedynczej, w mianowniku;
2) *droga* – rzeczownik, rodzaju żeńskiego,

w liczbie – mnogiej: NP_{Acc} (biernik) lub NP_N (mianownik),
– pojedynczej: NP_G (dopełniacz);

klasa semantyczna – $\begin{bmatrix} -\text{Anim} \\ -\text{Abstr} \end{bmatrix}$.

przyjaciel: przyjaciel – rzeczownik rodzaju męskiego, w liczbie pojedynczej, w mianowniku;
klasa semantyczna – [+Hum].

Wstępnie wybieramy ze słownika dla czasownika *prowadzić* schematy zawierające składowe o powyżej podanych klasach semantycznych i przypadkach. Przymiotnik *drogi*, jako część mowy, nie uczestniczy w wyborze schematów.
Na tym etapie wybrane zostały wszystkie schematy.

nas: *my* – zaimek, w liczbie mnogiej, w bierniku;
klasa semantyczna – [+Hum].

Wśród wybranych schematów szukamy tych, które zawierają składnik NP_{Acc} [+Hum]:

a) $\mathbf{NP_N}$ – do ∩ NP_G + (NP_I) + ($\mathbf{NP_{Acc}}$),
h) $\mathbf{NP_N}$ – $\mathbf{NP_{Acc}}$ + (NP_{Mod}),
i) $\mathbf{NP_N}$ – $\mathbf{NP_{Acc}}$ + (NP_{Adl}) + (NP_{Perl}) + (NP_{Abl}).

do ∩ sukcesu: *sukces* – rzeczownik, w liczbie pojedynczej, rodzaju męskiego, w dopełniaczu;
klasa semantyczna – [+Abstr].

Tylko schemat a) zawiera składnik do ∩ NP_G [+Abstr]:

a) $\mathbf{NP_N}$ – **do** ∩ $\mathbf{NP_G}$ + (NP_I) + ($\mathbf{NP_{Acc}}$).

Zdanie zostało przeczytane w całości. Pierwszy etap kończy się na wybraniu jednego schematu.

ETAP II

Generujemy oczekiwanie na rzeczownik klasy [+Abstr].

WNIOSKI

Zadaniem algorytmu było ustalenie, czy wyraz *drogi* występuje w zdaniu w roli przymiotnika, czy rzeczownika. Schemat, w którym ten wyraz mógłby występować jako rzeczownik został wyeliminowany. Zatem jako rzeczownik wyraz *drogi* nie sprawdza się. Jedyne co pozostaje, to uznać go za przymiotnik.

Przykład 4

Określenie sprawcy akcji.

Prowadziła nas do samej granicy.

Słownik syntaktyczno-generatywny: **prowadzić** (patrz przykład 3).

ETAP I

nas: *my* – zaimek, w liczbie mnogiej, w bierniku;
klasa semantyczna – [+Hum].

Wśród wybranych schematów szukamy tych, które zawierają składnik NP_{Acc} [+Hum]:

a) $NP_N - do \cap NP_G + (NP_I) + (\mathbf{NP_{Acc}})$,
h) $NP_N - \mathbf{NP_{Acc}} + (NP_{Mod})$,
i) $NP_N - \mathbf{NP_{Acc}} + (NP_{Adl}) + (NP_{Perl}) + (NP_{Abl})$.

do samej granicy: jest frazą adlatywną (NP_{Adl}), oznaczającą kierunek zbliżania się.
Tylko schemat i) zawiera taki składnik. Etap pierwszy kończymy z wybranym jednym schematem i z nieznanymi niektórymi jego składnikami.

ETAP II

Generujemy oczekiwanie na rzeczownik należący do klasy semantycznej [+Anim] (to może być człowiek albo zwierzę). W oparciu o orzeczenie (*prowadziła*) możemy dodatkowo wskazać, że to ma być rzeczownik rodzaju żeńskiego, w liczbie pojedynczej. Dodatkowe oczekiwania dotyczą fraz ablatywnej i perlatywnej.

Przykład 5

Określenie obiektu akcji

Komary jedzą.

Słownik syntaktyczno-generatywny: **jeść**

a) $NP_N - NP_{Acc}$,
$NP_N \rightarrow$ [+Anim],
$NP_{Acc} \rightarrow$ [Pokarm];

b) $NP_N - NP_{Acc}$,
$NP_N \rightarrow$ [owady],
$NP_G \rightarrow$ [+Anim].

ETAP I

Komary: *komar* – rzeczownik, w liczbie mnogiej, rodzaju męskiego, w mianowniku;
klasa semantyczna – [owady].

Wybieramy schemat b).

ETAP II

Oczekiwanie dotyczy rzeczownika klasy [+Anim] (coś żywego).

WNIOSKI

Celem przykładu było określenie obiektu akcji. Algorytm ustalił, jakiej klasy semantycznej to ma być obiekt. Zastąpienie wyrazu *komary* wyrazem *dzieci* spowodowałoby, że oczekiwanie dotyczyłoby rzeczownika klasy [pokarm].

5. Wnioski

Celem pracy było stworzenie metody rozstrzygającej niektóre typy wieloznaczności w trakcie analizy pojedynczych zdań. Metoda jest oparta o słownik syntaktyczno-generatywny czasowników.

Jak pokazują przykłady, algorytm oparty o tę metodę dobrze radzi sobie z określeniem brakujących elementów zdania. Możemy dla brakującego elementu określić nie tylko, do jakiej klasy semantycznej należy, ale i liczbę oraz rodzaj (dla brakującego podmiotu).

Dla wyrazów wieloznacznych, których znaczenia należą do różnych klas semantycznych, metoda dobrze wyznacza schematy i generuje oczekiwania. Oczekiwania mają pomóc w ustaleniu konkretnego znaczenia.

Wieloznaczność typu *zamek*, gdzie wszystkie znaczenia należą do tej samej klasy semantycznej, jest niemożliwa do rozwiązania przy pomocy tej metody, ale stworzenie dodatkowych struktur danych może pomóc rozwiązać ten problem.

Literatura

[1] Bolc L., Cichy M., Różańska L.: *Przetwarzanie języka naturalnego*. Warszawa, WNT 1982

[2] Kupść A., Marciniak M., Mykowiecka A.: *Komputerowe przetwarzanie języka naturalnego – wybrane zagadnienia*. Informatyka, nr 7, 1996

[3] Kamiński W.: *Automatyczne systemy pamiętania i wyszukiwania informacji*. Warszawa, WNT 1979

[4] Lubaszewski W.: *Gramatyka leksykalna w maszynowym słowniku języka polskiego*. Kraków, IP PAN 1997

[5] Lubaszewski W.: *Czy nowe językoznawstwo*. JP LXIV, 1984

[6] Lubaszewski W.: *Robot – Bibliotekarz, I – Rozumienie tekstu*. I Krajowa Konferencja Robotyki, Wrocław 1985

[7] Lubaszewski W.: *Archetype Driven Parser for Polish*. (maszynopis)

[8] Lubaszewski W.: *Rozumienie tekstu przez komputer*. Kraków, PAN 1990

[9] Polański K.: *Słownik syntaktyczno-generatywny czasowników polskich*. IP PAN 1992

[10] Raport Komisji Unii Europejskiej 1998 *LANGUAGE ENGINEERING. Progress and Prospects '98*

[11] Salton G.: *Automatic Information Organization and Rerieval*. New Jersey, Prentice Hall 1974

[12] Salton G.: *Dynamic Information and Library Processing*. New Jersey, Prentice Hall 1975

[13] Salton G.: *SMART – automatyczny system wyszukiwania informacji*. Warszawa, WNT 1975

[14] Schank R.C.: *Conceptual Dependency. A Theory of Natural Language Understanding*. Cognitive Psychology, 3, 1972

[15] Schank R.C. (red.): *Conceptual Information Processing*. Amsterdam, North Holland 1975

[16] Schank R.C.: *Conceptual Dependency Theory*. [w:] Schank R.C. (Ed.), Conceptual Information Processing, Amsterdam, North Holland, 1975, 22–82