Filozofia Nauki
Rok VI, 1998, Nr 1(21)

Maciej Gos

Metody probabilistyczne w semantyce języka naturalnego

Z interpretacją (semantyką) predykatów języka naturalnego wiążą się trudności i paradoksy, których przyczyną jest nieostrość terminów języka naturalnego. Semantyka ta jest adekwatną interpretacją języków sformalizowanych, lecz w wypadku języka naturalnego zawodzi. Interpretacja modelowa predykatów języka naturalnego w taki sposób, aby nie prowadziła ona do paradoksów jest jednak częściowo możliwa dzięki użyciu narzędzia matematycznego w postaci teorii zbiorów rozmytych i uogólnionej funkcji charakterystycznej.[1] Idea interpretacji konkretnej wartości funkcji charakterystycznej jako prawdopodobieństwa pojawiła się już podczas pierwszych badań nad zbiorami rozmytymi.[2] Celem tej pracy jest próba interpretacji liczbowej wartości uogólnionej funkcji charakterystycznej poprzez utożsamienie jej z wielkością prawdopopodobieństwa obliczoną z funkcji gęstości, a przede wszystkim dystrybuanty danego rozkładu prawdopodobieństwa na zbiorze, dla którego elementów rozkład prawdopodobieństwa jest określony. Taki zbiór utożsamimy z *quasi*-modelową interpretacją jednoargumentowego predykatu języka naturalnego.

Niech dany będzie pewien jednoargumentowy predykat języka naturalnego — przykładowo: „jest łysy", „jest inteligentny", „jest wysoki". Uniwersum *quasi*-modelu stanowić będzie zbiór ludzi, natomiast interpretacją powyższych predykatów w tym *quasi*-modelu będą odpowiednie zbiory rozmyte. Problem konstrukcji tego modelu sprowadza się do znalezienia metody przyporządkowania wartości funkcji charakterystycznej każdemu elementowi tych zbiorów. Ponieważ wzrost, iloraz inteligencji i liczba

[1] L. Zadeh, „Fuzzy Sets", *Information and Control* 8, s. 338-353.

[2] L. Zadeh, „Fuzzy Logic and Approximate Reasoning", *Synthese*, 30, s. 407-428. i B. R. Gaines, „General Fuzzy Logics", [w:] *Proceedings of the Third European Meeting on Cybernetics and Systems Research*, Vienna 1976.

włosów są zmiennymi statystycznymi mierzalnymi na zbiorze ludzi i posiadającymi rozkład normalny (Gaussa) prawdopodobieństwa na tym zbiorze, zatem dla dowolnego $x \in U$ (gdzie U jest zbiorem ludzi czyli uniwersum), określona jest wartość oczekiwana $\langle m \rangle$ i odchylenie standardowe rozkładu $\langle s \rangle$. Umożliwia to obliczenie wielkości prawdopodobieństwa wylosowania elementu, którego mierzalna cecha jest określona przedziałem liczbowym — np. elementu x wyższego niż 170 cm, o masie większej mniejszej niż 80 kg *etc.* Ponieważ zmienne statystyczne, takie jak liczba włosów, są dyskretne i posiadają skończoną liczbę argumentów, zatem rozkład normalny jest opisem przybliżonym, jednak ze względu na centralne twierdzenie graniczne jest to przybliżenie bardzo dokładne, asymptotycznie dążące do rozkładu $N\langle m, s \rangle$. Wielkość prawdopodobieństwa jest interpretowana jako wartość uogólnionej funkcji charakterystycznej dla każdego elementu zbioru będącego interpretacją danego predykatu.

Interpretacja taka pozwala wyeliminować klasyczny paradoks semantyki języka naturalnego — paradoks łysego — jak również przedstawić precyzyjnie interpretacje wielu predykatów języka naturalnego. Rozważymy ją na przykładzie predykatów „jest łysy" i „jest wysoki".

Zakładając, że rozkład zmiennej losowej X (liczba włosów) jest normalny w U, co będziemy skrótowo zapisywać $X - N\langle m, s \rangle$, w U można skonstruować zbiór rozmyty będący interpretacją predykatu „jest łysy". Ponieważ liczba włosów dowolnego $x \in U$ jest mierzalna, zatem stosując procedurę standaryzacji otrzymanej wartości, możemy obliczyć dystrybuantę standaryzowanej zmiennej, a zatem prawdopodobieństwo wylosowania elementu o danej lub mniejszej liczbie włosów.[3] W tym wypadku $1 - F(u)$ (gdzie $F(u)$ oznacza dystrybuantę standaryzowanej zmiennej losowej czyli liczbę włosów danego elementu zbioru U) przyjmuje wartości z przedziału [0, 1]. Jak łatwo wykazać rachunkowo, liczba ta będzie asymptotycznie zbliżać się do 1 zawsze i tylko wtedy, gdy liczba włosów elementu x maleje (kresem dolnym jest oczywiście 0). Ponieważ liczbę tę interpretujemy jako wartość funkcji charakterystycznej zadanej na elementach zbioru rozmytego, będącego interpretacją predykatu „jest łysy", zatem zgodnie z intuicją, wartość funkcji charakterystycznej będzie tym wyższa, im mniej włosów ma element zbioru U, poniżej pewnej liczby włosów osiągając wartość 1 w zaokrągleniu do dowolnego miejsca po przecinku, co wynika z własności dystrybuanty ciągłych zmiennych losowych.

Zaletą tej koncepcji jest jej intuicyjność — dla każdego elementu zbioru U, im wyższa w stosunku do liczbowej wartości $\langle m \rangle$ (wartości oczekiwanej, czyli średniej statystycznej zmiennej losowej w całym zbiorze U) jest wartość zmiennej X (liczba włosów), tym bardziej $1 - F(u)$ zbliża się do wartości 0. Powyżej pewnej liczby włosów osiąga ona asymptotycznie 0. Intuicyjność interpretacji w semantyce języka naturalne-

[3] W wypadku rozkładu normalnego zmiennej losowej dystrybuanta określona jest całką Gaussa, której wartość może asymptotycznie dążyć do 0 lub 1.

go wartości dystrybuanty rozkładu zmiennej losowej jako liczbowej wartości uogólnionej funkcji charakterystycznej elementów zbioru rozmytego będącego interpretacją predykatu języka naturalnego, wynika z tego, że obliczenia pozwalają otrzymać wynik zgodny z oczekiwaniem. W wypadku predykatu „jest łysy" otrzymamy ze wzoru $1-F(u)$ wartość funkcji charakterystycznej 0 dla osób posiadających znacznie więcej włosów niż wynosi $\langle m \rangle$ (średnia statystyczna w populacji). Innymi słowy, osoby te nie należą do zbioru będącego interpretacją tego predykatu. W odniesieniu do predykatu „jest wysoki" dla dowolnego $x \in U$, wartość $f(x)$ (funkcji charakterystycznej tego elementu) jest równa $F(u)$, nie zaś $1 - F(u)$, gdzie u jest standaryzowanym wynikiem pomiaru wzrostu elementu x.[4] Łatwo wykazać na drodze rachunkowej, że im większy wzrost danego elementu x w stosunku do średniej $\langle m \rangle$ w całej populacji, tym większa wartość funkcji charakterystycznej (zbliża się ona nieograniczenie do 1). Dualnie, im mniejszy wzrost, tym mniejsza wartość $f(x)$. Poniżej pewnej wielkości przyjmujemy, że jest ona równa 0, a argumenty tej funkcji nie należą oczywiście do interpretacji predykatu „jest wysoki". Dowód rachunkowy tego faktu jest podobny do dowodu dotyczącego analizowanego predykatu „jest łysy".

Procedura wyznaczania zbioru rozmytego i obliczania wartości funkcji charakterystycznej jego elementów przebiega w identyczny sposób dla wszystkich predykatów języka naturalnego odnoszących się do cech, które przy zastosowaniu powyżej opisanej metody są sprowadzalne do mierzalnych zmiennych losowych (niekoniecznie o rozkładzie normalnym, choć taki rozkład jest najczęstszy w zastosowaniach praktycznych). Co więcej, nawet pewne niemierzalne (nominalne, rangowe[5]) zmienne losowe będące formalnym odpowiednikiem niemierzalnych cech denotowanych przez predykaty typu: „jest wykształcony", „jest zdrowy", można sprowadzić, dzięki centralnemu twierdzeniu granicznemu,[6] do zmiennych o rozkładzie $N\ \langle m, s \rangle$, czyli o rozkładzie normalnym z określoną średnią i odchyleniem standardowym. Tym samym pro-

[4] Standaryzacja zmiennej losowej, czyli liniowe przekształcenie jej wartości x dane jest za pomocą następujących zależności funkcyjnych:

$$u = f(x) = (x - m) / s,$$

gdzie m jest wartością oczekiwaną (średnią) rozkładu zmiennej losowej, zaś s odchyleniem standardowym tego rozkładu, umożliwia odczytanie wartości dystrybuanty tego rozkładu z tablic statystycznych (bez konieczności żmudnego obliczania całki Gaussa).

[5] Zmienne takie są formalnym odpowiednikiem niemierzalnych cech związanych z predykatami typu „jest chory", „jest smaczny", często występującymi w języku naturalnym. Przyjmujemy wówczas jako zbiór wartości takiej zmiennej skalę rangową — np. od 0 do 10. Dla elementu uniwersum, dla którego nominalna zmienna losowa przyjmuje konkretną wartość, wartość jego funkcji charakterystycznej (czyli — intuicyjnie rzecz biorąc — stopień przynależności do interpretacji danego predykatu) jest równa wartości dystrybuanty (F) lub $1 - (F)$ obliczonej dla tej wartości zmiennej nominalnej.

[6] Intuicyjne sformułowanie twierdzenia granicznego sprowadza się do konstatacji, że zmienna losowa będąca sumą nieograniczenie wielu dowolnych zmiennych losowych ma zawsze rozkład normalny o określonych parametrach m i s. Ścisłe sformułowanie tego twierdzenia i jego dowód można znaleźć np. w książce Z. Hellwiga, *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, PWN, Warszawa 1993, s.178-182.

babilistyczna metoda konstrukcji interpretacji predykatów języka naturalnego znajduje zastosowanie również przy analizie predykatów odnoszących się do cech niemierzalnych, a dystrybuanta rozkładu normalnego jest podstawowym narzędziem obliczania wartości funkcji charakterystycznej, czyli wyznaczania zbioru rozmytego będącego interpretacją predykatu.

Na podstawie tej analizy możemy zatem wprowadzić pojęcie *quasi*-modelu dla języka naturalnego. Jest to uporządkowana struktura

$$QM: = \langle U, A, A', A'', ..., f, f', f'' ... \rangle,$$

której uniwersum U jest zbiorem wszystkich obiektów denotowanych przez język naturalny, zbiory A, A'.... są zbiorami rozmytymi (podzbiorami uniwersum), takimi że dla ich elementów wartość wyznaczających je funkcji charakterystycznych istnieje i jest obliczalna przy użyciu opisanego powyżej algorytmu — metody probabilistycznej. Zbiory te są zatem interpretacjami predykatów jednoargumentowych języka naturalnego. Relacje i funkcje f, f', f''.... są analogicznie jak w modelu języka sztucznego interpretacjami odpowiednich terminów odnoszących się do par uporządkowanych i ich zbiorów. Powyższa struktura została nazwana „*quasi*-modelem", gdyż podzbiory uniwersum będące interpretacjami predykatów jednoargumentowych są zbiorami rozmytymi i dla każdego ich elementu wartość funkcji charakterystycznej posiada jednoznaczną interpretację probabilistyczną (wartość dystrybuanty lub (1 – wartość dystrybuanty) zmiennej losowej) będącej formalnym odpowiednikiem cechy denotowanej przez predykat. Relacje i funkcje są również zbiorami rozmytymi — podzbiorami rozmytymi odpowiednich iloczynów kartezjańskich.

Należy podkreślić, że probabilistyczna interpretacja wartości funkcji charakterystycznej, którą zaproponowali Zadeh i Gaines[7] związana była z subiektywną teorią prawdopodobieństwa przedstawioną przez Gilesa[8] lub z teorią prawdopodobieństwa warunkowego Gainesa,[9] podczas gdy przedstawiona powyżej interpretacja oparta jest na klasycznej teorii prawdopodobieństwa zadanej trzema aksjomatami Kołmogorowa.[10] Współczesna probabilistyka obejmująca teorię ciągłej zmiennej losowej (oczy-

[7] Zob. przypis 1 i 2.

[8] Teoria probabilistyki R. Gilesa jest koncepcją prawdopodobieństwa subiektywnego zdefiniowanego dla zdań, nie zaś dla zdarzeń. Ponieważ jest ona sprzeczna z aksjomatyczną teorią prawdopodobieństwa Kołmogorowa, nie może służyć do wypracowania teorii ciągłej zmiennej losowej (czyli również teorii zmiennej o rozkładzie normalnym). Przykładowo, w teorii Gilesa $P(p \vee \sim p) = \max \{P(p), P(\sim p)\}$, czyli niekoniecznie jest równa 1, zaś w klasycznej probabilistyce prawdopodobieństwo sumy zdarzenia i jego dopełnienia jest zawsze równe 1. Zob. R. Giles, „A Non-classical Logic for Physics", *Studia Logica* 33, 1974, s. 394-416.

[9] Analogicznie teoria prawdopodobieństwa warunkowego B. R. Gainesa nie nadaje się do konstrukcji teorii ciągłej zmiennej losowej ze względu na przyjętą definicję prawdopodobieństwa sumy i iloczynu zdarzeń. Zob. B. R. Gaines, *op. cit.*

[10] Aksjomatyka Kołmogorowa teorii prawdopodobieństwa, którą zakładamy jako podstawę probabilistyki obejmuje trzy aksjomaty:

A. 1 $P(A) \in [0, 1]$;

A. 2 $P(U) = 1$, gdzie U jest zdarzeniem pewnym;

A. 3 $P(A \cup B) = P(A) + P(B)$ zawsze i tylko wtedy, gdy zdarzenia A, B są rozłączne.

wiście również zmiennej o rozkładzie $N\langle m, s\rangle$, która posiada decydujące znaczenie w tej koncepcji semantyki języka naturalnego) oparta jest na teorii miary i całki oraz aksjomatyce Kołmogorowa, co automatycznie powoduje odrzucenie teorii prawdopodobieństwa Gilesa i związanej z nią matrycowej logiki zdań o nieprzeliczalnej liczbie wartości logicznych[11] nazywanej „logiką rozmytą".[12]

Wyżej przedstawiona semantyka predykatów języka naturalnego nie jest wolna od pewnych nieintuicyjnych rozwiązań — stopień tej nieintuicyjności jest jednak nieporównanie mniejszy niż w wypadku klasycznego paradoksu łysego. Problem ten wynika z podstawowej właściwości rozkładu normalnego — jeżeli dla elementu pewnego zbioru wartość zmiennej losowej X o rozkładzie $N\langle m, s\rangle$, określonej na tym zbiorze, jest równa wartości m, czyli średniej statystycznej (wartości oczekiwanej) w całym zbiorze, to wartość dystrybuanty $F(u)$ oraz $1 - F(u)$ jest równa 0,5.[13] Oczywiście wielkość ta jest wartością funkcji charakterystycznej elementu zbioru rozmytego, będącego interpretacją danego predykatu. Znaczy to, że zarówno w wypadku predykatu „jest wysoki", jak i „jest niski", wartość $f(x)$ (funkcji charakterystycznej) dla każdego elementu uniwersum, którego wzrost jest równy średniej m w całej populacji, jest równa 0,5, w pierwszym wypadku $f(x) = F(u)$, gdzie u czyli standaryzowana wartość wzrostu jest równa 0, w drugim wypadku $f(x) = 1 - F(u)$. Otrzymana wartość funkcji charakterystycznej 0,5 dla elementu interpretacji, dla którego mierzalna cecha przyjmuje wartość równą średniej w całym uniwersum, nie jest niezgodna z intuicją. Niezgodność z intuicją pojawia się w wypadku takich predykatów, jak „jest łysy". Analogicznie przyjmując, że liczba włosów ma w rozpatrywanym uniwersum (oczywiście dla tych elementów uniwersum *quasi*-modelu, dla których ma ona określoną wartość) rozkład $N\langle m, s\rangle$, otrzymamy w wyniku obliczenia $1 - F(u)$ wniosek, że wartość funkcji charakterystycznej dla osoby posiadającej liczbę włosów równą średniej m w całej populacji (dziedzinie zmiennej losowej, będącej podzbiorem uniwersum) jest równa 0,5. Twierdzenie, że osoba posiadająca przeciętną liczbę włosów należy do interpretacji predykatu „jest łysy" z wartością $f(x) = 0{,}5$ jest nieintuicyjne, nie otrzymujemy jednak paradoksu na wzór klasycznego paradoksu łysego.

Podsumowując należy podkreślić, że zastosowanie teorii zbiorów rozmytych w połączeniu z interpretacją probabilistyczną funkcji charakterystycznej (opartą przede wszystkim na rozkładzie normalnym ciągłej zmiennej losowej) umożliwia w znacznej

---

[11] R. Giles przypisywane zdaniom prawdopodobieństwo interpretował jako wartość logiczną zdań i skonstruował system logiki zdań o continuum wartości logicznych i prawdziwościowych spójnikach. Logikę te charakteryzuje matryca identyczna z matrycą przeliczalnie wielowartościowej logiki J. Łukasiewicza. Zob. R. Giles, *op. cit.*

[12] Związek teorii zbiorów rozmytych z logiką zdań (o continuum wartości logicznych) i z teoriami prawdopodobieństwa, które są sprzeczne z teorią Kołmogorowa, omawia G. Malinowski w: *Logiki wielowartościowe*, PWN, Warszawa 1990, s. 112-119.

[13] Dowód otrzymujemy natychmiast: jeżeli bowiem $x = m$, to $[u = (x - m) / s] = 0$, zaś w standaryzowanym rozkładzie normalnym $F(u) = 0{,}5$ zawsze i tylko wtedy, gdy $u = 0$.

mierze eliminację paradoksów semantyki predykatów języka naturalnego, pozwala zdefiniować interpretację predykatu języka naturalnego jako zbioru rozmytego, takiego że funkcja charakterystyczna dla jego elementów jest określona dystrybuantą pewnego rozkładu zmiennej losowej, dzięki czemu jest ona obliczalna algorytmicznie. Pojęcia te pozwalają także zdefinować *quasi*-modele dla języka naturalnego.