KYBERNETIKA ČÍSLO 3, ROČNÍK 2/1966

Statistický odhad sémantické informace*

VLADIMÍR DROZEN, STANISLAV LANGER

Práce popisuje metodu kvantitativního hodnocení sémantického obsahu dílčích složek (vět) zpráv, formulovaných přirozenými jazyky, na principu redukce textu a referuje o získaných výsledcích.

Ve stati „Poznámky k definici pojmu sémantické informace" [1] jsme svého času naznačili, že jedním z možných a prakticky realizovatelných přístupů ke kvantitativnímu hodnocení sémantického obsahu zpráv, formulovaných přirozenými jazyky, by byla vhodně provedená konfrontace souboru zpráv se souborem osob, které jsou uživateli příslušného jazyka.

K tomuto cíli jsme vypracovali jednoduchou metodu. Zprávy, pro jejichž dílčí složky hledáme míru sémantického obsahu, předkládáme souboru osob, které mají za úkol provést subjektivní ocenění jednotlivých složek na základě instrukce, aby z celé zprávy vybraly právě polovinu původních vět tak, aby smysl celé zprávy utrpěl co nejméně. Zdůvodnění tohoto postupu je prosté: jde tu o nejjednodušší typ rozhodování — převzít či nepřevzít určitou větu do zkrácené verze, přičemž redukce na polovinu původního rozsahu zaručuje maximální entropii výběru, neboť počet kombinací $\binom{n}{m}$ je maximální právě když $m = n/2$.

Použili jsme 12 článků po osmi větách, které jsme předložili dvěma dvacetičlenným souborům. První soubor se skládal z dospělých lidí se středoškolským vzděláním (věk 25 až 30 let), kdežto druhý soubor byly děti ve věku 12 let. V polovině článků převládala dějová náplň, zbývající měly popisný charakter. Příklad článků je uveden na konci tohoto sdělení.

Způsob vyhodnocení bude možno nejlépe objasnit na příkladu. Článek č. 1 dal v souboru dětí výsledky uvedené v tabulce. Větu č. 1 převzalo do zkrácené verse 16 členů souboru a proto jí přiřadíme váhu 16. Analogicky postupujeme u ostatních

* Předneseno na *druhé konferenci o kybernetice*, která se konala v Praze ve dnech 16.—19. listopadu 1965.

vět. Tak získáme svislé marginální hodnoty $x_{i,k}$ (k označuje pořadové číslo článku) na pravé straně tabulky, které vyjadřují relativní sémantické váhy jednotlivých vět. Marginální součty $y_{j,k}$ na spodním okraji tabulky získáme tak, že za každou ponechanou větu v příslušném sloupci dosadíme její váhu a tyto hodnoty sečteme. Sou-

Tabulka

Článek č. 1

Věta č.	Pokusná osoba č. 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	x_i
1	+	+		+	+	+		+		+	+	+	+	+	+	+		+	+	+	16
2	+		+	+	+	+		+	+		+	+	+	+		+	+	+		+	15
3			+				+	+	+	+			+			+	+				8
4	+	+		+	+	+	+				+	+								+	9
5															+				+		2
6		+	+	+			+	+	+	+		+	+	+	+		+	+			13
7									+		+				+			+	+		5
8	+	+	+		+	+	+			+				+		+	+		+	+	12
$y_{j,1}$	52	50	48	53	52	52	42	52	41	49	45	53	52	56	36	51	48	49	35	52	

čty $y_{j,k}$ vyjadřují velmi obecnou charakteristiku každého jednotlivého člena souboru ve vztahu k danému článku; tuto hodnotu bychom mohli přibližně označit jako „stupeň normality": čím vyšší $y_{j,k}$, tím více se reakce individua blíží průměru.

Z marginálních hodnot $x_{i,k}$ jsme vypočetli entropie jednotlivých článků podle vzorce

$$H'_k = \sum_{i=1}^{n} \frac{x_{i,k}}{N} \operatorname{ld} \frac{x_{i,k}}{N}$$

(ld značí dvojkový logaritmus). V našem případě n = počet vět v článku = 8, $N = \frac{1}{2}n$. počet osob = 80. Maximální možná entropie článku se u našeho pokusu rovnala 3 bitům, minimální entropie 2 bitům. Je proto účelné zavést relativní entropii článku

$$H_k = H'_k - H'_k(\min) .$$

Při zpracovávání materiálu jsme shledali, že z celkového počtu 70 možných větných kombinací se ve skutečnosti vyskytuje jen velmi omezený počet, u souboru dospělých maximálně 11; proto je možné hodnotit entropii článků i na základě celých větných kombinací. Číselné hodnoty vyjdou ovšem odlišné, už proto, že takto definovaná entropie leží v rozmezí od 0 do 4,32 bitu, avšak pořadí článků podle obou kritérií

je přibližně stejné, korelace podle pořadí činí 0,90, takže lze říci, že obě měřítka vyjadřují tutéž skutečnost.

Celkovou entropii souboru vzhledem k dané množině článků stanovíme ze skupinového rozdělení četnosti součtů $\sum_k y_{j,k}$. Opět se dá vyjít buďto z bodování jednotlivých vět nebo z celých větných kombinací.

Uvedeme stručně některé získané výsledky. Relativní entropie článků, stanovené z výběru jednotlivých vět, se u souboru dospělých pohybují od 0,44 do 0,78 bitu, kdežto u dětí od 0,71 do 0,85 bitu; pozoruhodné je, že u dětí vycházejí pro všechny články bez výjimky větší entropie než u dospělých; diference u jednotlivých článků se různí a korelace podle pořadí mezi souborem dospělých a dětí, pokud jde o entropie jednotlivých článků, dává hodnotu 0,45 (významné na úrovni 95% spolehlivosti). Úhrnná charakteristická čísla jednotlivců – součty $\sum_k y_{j,k}$, stanovené na základě výběru jednotlivých vět, leží u dospělých v intervalu od 649 do 742 bodů, u dětí od 520 do 650 bodů, kdežto analogické hodnoty, odvozené z větných kombinací, leží u dospělých mezi 26 a 89 body, přičemž průměrná hodnota je 63,3 a $\sigma = 17{,}8$, variační koeficient = 0,28; u dětí se rozpětí pohybuje od 13 do 38 bodů, průměr je 29,2, $\sigma = 6{,}4$ a variační koeficient = 0,22. Entropie souboru, získaná ze skupinového rozdělení četnosti, činí u dospělých 2,14, u dětí 2,38 bitu.

Váhy jednotlivých vět v rámci daných článků kolísají u dospělých v plném rozsahu od 0 do 20; naproti tomu u dětí se nevyskytla ani jediná věta, kterou by alespoň jeden člen souboru nepojal do redukované verze článku.

Zkoumali jsme také otázku statistické významnosti rozdílů mezi krajními hodnotami charakteristických čísel jednotlivců. Shledali jsme, že tyto rozdíly leží nad hladinou statistické významnosti, a máme zato, že signalizují významné diference ve způsobu myšlení.

Očekáváme, že popsané metody budeme moci použít k psychodiagnostickým účelům, dále k zjišťování homogenity souborů při otázkách souvisejících s diferenciací vyučování apod. Zdá se, že soubor 10 až 12 článků rozsahu námi použitého představuje minimum, které může poskytnout prakticky použitelné výsledky; v zájmu spolehlivosti by bylo lépe použít raději rozsáhlejšího materiálu. Máme proto v úmyslu opakovat pokus v rozsáhlejším měřítku, za použití souboru 20 článků, abychom si mohli ověřit validitu metody jednak vzhledem k charakteristickým číslům jednotlivců, jednak i vzhledem k číselným parametrům, přisouzeným podle popsané metody použitým textům. Rovněž bude nutno zjistit, jak dalece budou tyto parametry záviset na případných změnách ve formulaci úlohy, např. při požadavku, aby z původního článku bylo vybráno do zkrácené verse jiné dané procento vět než právě polovina.

Pomocí vhodně utvářených textů bude nepochybně možno získávat také psychodiagnostické informace zcela specifické povahy. Na druhé straně budeme moci touto metodou hodnotit správnost koncepce a účinnost různých sdělení, jako např. učebních textů.

Počítáme s tím, že se nám podaří vypracovat analogickou metodu i pro oblast vizuálně přijímané informace.

Dodatkem uvádíme znění dvou článků použitých v pokusu: jeden článek s dějovým obsahem a jeden článek popisného typu.

Želva a zajíc

1. Zajíc a želva závodili v běhu.
2. Určili si trať a závod začal.
3. Zajíc prudce vyrazil od startu a předběhl želvu.
4. Když zajíc za chvíli uviděl, jak je želva pozadu, natáhl se do trávy a usnul.
5. Je pravdivé přísloví: Kdo druhého podceňuje, může na to doplatit.
6. Zajíc si myslel, že ho želva stejně nedohoní.
7. Želva běžela pomalu, ale vytrvale.
8. Zatímco zajíc spal, doběhla želva k cíli a zvítězila.

Sovy

1. Sovy jsou noční ptáci.
2. Sovy létají v noci.
3. Většina sov žije v lesích.
4. Několik sovích druhů žije mimo les.
5. Mnohé druhy žijí vysoko v horách.
6. Sovy mají zvláštní hlavu a nápadné oči.
7. Pro tuto nápadnost hlavy byly sovy pokládány za moudré tvory.
8. Sovy se živí drobnými škodlivými živočichy a proto jsou užitečné.

Autoři děkují ředitelství ZDŠ v Chlumci nad Cidlinou za ochotu, s níž jim umožnilo provedení pokusu na tamní škole, a Dr. A. Perezovi, DrSc. za podnětné připomínky k ověření validity metody.

(Došlo dne 5. ledna 1966.)

LITERATURA

[1] V. Drozen, P. Nádvorník, V. Pelikán: Poznámky k definici pojmu sémantické informace. Přednáška na první konferenci o kybernetice konané v Praze v listopadu 1962. Resumé otištěno ve sborníku „Kybernetika a její využití", NČSAV, Praha 1965, str. 242.

A Statistical Evaluation of Semantic Information

VLADIMÍR DROZEN, STANISLAV LANGER

One possible way how to attribute quantitative semantic weights to meassages presented in natural languages is based on the confrontation of an ensemble of messages with an ensemble of users of the language concerned.

The ensemble of human subjects is given the instruction to choose from each item (message) just half the number of the original sentences in such a manner that the sense of the message may suffer as little as possible. Semantic weight of each sentence is then defined to be the number of subjects that have taken the respective sentence up into the abbreviated version.

The values thus gained enable us to evaluate entropies of the items (messages) and to attribute "characteristic numbers" to individual subjects of the human ensemble, which may be of considerable psychodiagnostic value. Entropies of the human ensembles may also be calculated.

The authors describe the results of an experiment thus performed with two human ensembles (children and adults) of 20 subjects, the ensemble of messages consisting of 12 items with 8 sentences each.

Dr. Vladimír Drozen, katedra matematiky, Dr. Stanislav Langer, katedra pedagogiky a psychologie, Pedagogická fakulta Hradec Králové.