Automatyczne rozpoznawanie języka tekstu

Mirosław Gajer

Przedstawiony system do automatycznego rozpoznawania języka tekstów może stanowić pierwsze ogniwo większego systemu, służącego do automatycznego tłumaczenia wielojęzycznych tekstów, gdzie zachodzi konieczność ustalenia języka tłumaczonego dokumentu. W artykule zaproponowano prostą, ale odznaczającą się dużą skutecznością, metodę rozpoznawania języka tekstu pisanego. System taki w połączeniu z programem konwertera sygnału mowy do postaci tekstowej (*speech-to-text converter*) może posłużyć do identyfikacji języka użytkownika systemu dialogowego.

W ostatnich latach występuje wzmożone zainteresowanie technikami informatycznymi, związanymi z przetwarzaniem, analizą, rozpoznawaniem i rozumieniem języka naturalnego [1]. W związku z tym tworzone są systemy informatyczne operujące na tzw. kontrolowanych językach naturalnych, stanowiących odpowiednie podzbiory języków naturalnych, które mogą być efektywnie i precyzyjnie przetwarzane w systemie komputerowym [2]. Opracowywane są także systemy, których zadaniem jest rozpoznawanie i interpretowanie sygnału mowy w czasie rzeczywistym. Badania prowadzone w tym zakresie koncentrują się zwłaszcza na zagadnieniach rozpoznawania fonemów, które są podstawowymi jednostkami, jak gdyby atomami, z których budowane są głoski, wyrazy, zdania i całe wypowiedzi. Prawidłowa identyfikacja poszczególnych fonemów stanowi podstawę budowy każdego systemu rozpoznawania mowy [3].

Rozwinięta została także teoria dotycząca metod statystycznych pozwalających na efektywny opis matematyczny prawidłowości występujących w sygnale mowy. Na szczególną uwagę zasługują prace wykorzystujące do opisu sygnału mowy niejawne modele Markowa HMM (*Hidden Markov Models*) [4]. Prowadzone badania koncentrują się szczególnie na rozpoznawaniu sygnału mowy ciągłej [5] oraz rozpoznawaniu wyrazów, należących do słownika o bardzo dużej objętości [6].

Jednym z głównych celów prowadzenia badań w dziedzinie technik przetwarzania i automatycznego rozpoznawania języków naturalnych jest opracowanie interfejsu, pozwalającego na skuteczną komunikację głosową między człowiekiem i maszyną [7]. Celem, ku któremu usilnie dążą badacze jest opracowanie tzw. mówionych systemów dialogowych, umożliwiających prowadzenie swobodnej konwersacji głosowej z systemem komputerowym [8]. Jako przykład takich systemów można wymienić opracowany w Stanach Zjednoczonych, pracujący w czasie rzeczywistym, system dialogowy Jupiter, którego zadaniem jest udzielanie precyzyjnych odpowiedzi na pytania użytkownika, dotyczące prognozy pogody dla wskazanego przez niego obszaru Ameryki Północnej [9]. Badania dotyczące automatycznych interfejsów dialogowych skupiają się głównie na niełatwym, i wciąż nie do końca poznanym, zagadnieniu automatycznej adaptacji systemu przetwarzania sygnału mowy do pewnych specyficznych cech wypowiedzi generowanych przez użytkownika aktualnie korzystającego z systemu [10].

Znane są także systemy służące do automatycznego tłumaczenia języków. Przykładowo w pracy [11] został opisany system służący do tłumaczenia w czasie rzeczywistym na język angielski sygnału mowy wypowiadanej w języku niemieckim. Konstrukcja automatycznych systemów tłumaczących wymaga opracowania specjalnych matematycznych modeli języka. Najczęściej w tym celu jest wykorzystywana tzw. technika N-gramów (*N-grams*), pozwalająca na wyliczenie prawdopodobieństwa wystąpienia w wypowiedzi arbitralnie zadanej sekwencji N wyrazów [12].

Ostatnio dużą popularnością cieszą się prace prowadzone nad systemami rozpoznawania sygnału mowy integrującymi wiele różnych języków. Multilingwalne interfejsy dialogowe oparte na takich systemach, pozwalają użytkownikom wielu różnych języków na uzyskanie za pośrednictwem tego samego systemu potrzebnych informacji na dany temat [13]. Przykładowo, system taki może pozwalać na dokonywanie rezerwacji biletów lotniczych osobom posługującym się jednym z wielu dostępnych w rozważanym systemie języków. W celu realizacji tak postawionego zadania, należy na wstępie dokonać identyfikacji języka, którym posługuje się osoba, komunikująca się z systemem dialogowym.

Wybór cech charakterystycznych rozpoznawanych języków

Każdy z języków ma pewne cechy charakterystyczne pozwalające na jego identyfikację. Doboru cech, na podstawie których język będzie identyfikowany, należy dokonać mając na uwadze zagadnienia takie, jak: odpowiednio wysoka skuteczność procesu rozpoznawania, odporność systemu na wszelkiego rodzaju zakłócenia, łatwość realizacji technicznej, prostota obsługi, możliwość rozszerzenia funkcji systemu (dołączanie do istniejącego systemu nowych języków), szybkość działania systemu i wiele innych [14].

W przypadku systemu omawianego w artykule, rozpoznawanie języka tekstu oparto na analizie częstości występowania w zadanym fragmencie tekstu poszczególnych znaków (liter). Analiza częstości występowania

Dr inż. Mirosław Gajer jest pracownikiem Katedry Automatyki Akademii Górniczo-Hutniczej.

poszczególnych znaków została przeprowadzona dla zbioru podstawowych 26 znaków alfabetu łacińskiego, czyli: {a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z}. Spośród współczesnych języków indoeuropejskich jedynie angielski i niderlandzki wykorzystują do zapisu wyrazów wyłącznie znaki należące do wymienionego zbioru podstawowego. Praktycznie wszystkie inne języki, oprócz znaków podstawowych, wykorzystują także różnorodne ich modyfikacje. Najlepszym przykładem jest język polski, w którym dodatkowo występują znaki: ą, ę, ś, ć, ż, ź, ó, ł, ń.

W związku z powyższym rozważany system rozpoznawania języka należało wyposażyć w preprocesor, dzięki któremu wszelkie narodowe modyfikacje znaków były zastępowane znakami ze zbioru znaków podstawowych. Np. w tekście napisanym w języku polskim program preprocesora dokonywał konwersji: ą→a, ę→e, ś→s, ć→c, ż→z, ź→z, ó→o, ł→l oraz ń→n. W efekcie, niezależnie od języka tekstu wejściowego, dalszej analizie były poddawane wyłącznie próbki tekstu zapisane przy użyciu standardowego zbioru 26 znaków.

Jako cechy charakterystyczne, pozwalające na identyfikację danego języka, przyjęto wyrażoną w procentach częstość występowania poszczególnych znaków ze zbioru standardowego. Nawet pobieżna analiza tekstów napisanych w różnych językach pozwala na wysuniecie przypuszczeń odnośnie do możliwości konstrukcji efektywnego systemu rozpoznawania języka opartego na analizie częstości występowania poszczególnych znaków w zadanej próbce tekstu. Przykładowo w języku polskim praktycznie nie są spotykane znaki *x*, *q* i *v*, podczas gdy w językach angielskim i francuskim znaki te występują dość powszechnie. Podobnie w językach romańskich (francuski, włoski, hiszpański, portugalski) nie ma znaku *k* (poza nielicznymi wyrazami, będącymi zapożyczeniami z innych języków), podczas gdy w języku polskim znak ten występuje nader często.

Pomiar częstości występowania znaków w rozpoznawanych językach

Przystępując do konstrukcji systemu identyfikacji języka tekstu, należy dokonać badań statystycznych, pozwalających na oszacowanie częstości występowania poszczególnych znaków w każdym z rozpoznawanych języków. Oczywiście teksty zapisane w tych językach muszą zostać uprzednio przetworzone za pomocą programu preprocesora.

Podstawowe pytanie, na które należy odpowiedzieć, dotyczy długości tekstu, pozwalającej na dostatecznie dokładne oszacowanie częstości występowania poszczególnych znaków w badanym języku. W tabl. 1 zamieszczono wyniki takich badań, które przeprowadzono dla języka angielskiego.

Tabl. 1 pokazuje, że począwszy od tekstu zawierającego 500 znaków występuje stabilizacja uzyskanych wyników na określonym poziomie. Dalsze zwiększanie długości tekstu nie przynosi już większej poprawy dokładności pomiarów, zaobserwować można jedynie pewne fluktuacje statystyczne w częstości występowania poszczególnych znaków.

Jednakże w celu uzyskania pewności, że otrzymane wyniki są reprezentacyjne dla danego języka, pomiaru częstości występowania poszczególnych znaków dokonano na podstawie próbki tekstu o długości ok. 3000 znaków, będącej zlepkiem tekstów pochodzących z różnych źródeł. Chodziło o to, aby badana próbka była jak najbardziej reprezentatywna dla danego języka. W przypadku

Tablica 1. Częstość (w procentach) występowania poszczególnych znaków dla języka angielskiego, w zależności od liczby znaków występujących w tym tekście (długości badanego tekstu)

znaki / długość tekstu	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
100	12,8	0,0	1,1	3,2	16,0	2,1	3,1	4,2	8,5	0,0	0,0	2,1	4,0	7,4	2,1	1,1	0,0	5,3	8,5	9,9
200	9,4	0,5	1,9	1,9	17,8	2,4	1,9	4,2	7,0	0,0	0,5	3,8	3,3	6,1	2,8	1,9	0,0	6,1	9,9	9,9
500	7,5	1,2	4,0	2,3	15,7	3,2	1,3	4,4	7,5	0,5	0,3	3,9	2,2	6,0	5,0	1,2	0,2	7,4	8,5	9,8
1000	7,2	1,6	4,0	2,7	14,6	2,4	1,7	4,0	7,9	0,2	0,3	4,3	2,5	6,2	6,1	1,9	0,2	7,5	7,3	9,1
2000	7,3	1,4	4,4	3,0	14,3	2,3	1,5	3,8	7,3	0,2	0,2	3,9	2,7	7,0	6,5	2,3	0,2	6,9	7,0	9,2
3000	7,5	1,3	4,3	3,0	14,0	2,3	2,1	3,6	8,1	0,2	0,2	4,1	2,7	7,1	6,5	2,4	0,2	6,3	7,0	9,4

Tablica 2. Częstość (w procentach) występowania poszczególnych znaków dla wybranych języków

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
PL	10,0	1,7	5,0	2,8	10,7	0,2	1,5	1,0	0,5	3,2	3,3	3,8	3,4	4,4	7,5	3,3	0,0	3,8	5,1	5,0
D	6,1	2,7	3,4	4,0	16,3	1,9	2,5	4,9	8,7	0,0	1,9	3,8	3,0	10,1	2,9	0,7	0,1	5,5	7,1	7,1
GB	7,5	1,3	4,3	3,0	14,0	2,3	2,1	3,6	8,1	0,2	0,2	4,1	2,7	7,1	6,5	2,4	0,2	6,3	7,0	9,4
NL	7,7	1,7	1,5	4,9	17,8	1,1	3,6	2,4	6,3	1,1	3,4	4,2	2,0	8,6	5,6	1,7	0,0	6,6	3,5	7,8
F	6,5	0,9	4,0	3,7	17,5	1,3	1,1	0,8	7,1	0,2	0,0	5,1	3,0	6,9	7,0	2,4	1,1	7,2	6,9	7,5
E	10,9	1,1	5,2	5,0	13,7	1,0	1,8	0,4	7,1	0,6	0,0	5,6	2,7	7,8	9,4	2,6	0,7	6,4	7,0	4,6
I	11,8	1,1	4,0	3,3	10,9	0,8	2,0	0,6	9,8	0,0	0,0	5,0	2,5	6,5	9,9	3,9	0,8	8,1	4,8	7,2

pobrania próbki tekstu z jednego dokumentu, łatwo wyobrazić sobie sytuację, w której w rozważanym tekście bardzo często występują pewne słowa związane bezpośrednio z jego tematyką, np. w tekście dotyczącym architektur mikroprocesorów bardzo często pojawiać się będzie słowo procesor, w skutek czego częstość występowania takich znaków jak *p, r, o, c, e* i *s* może być, w próbce tekstu, znacznie większa niż ma to miejsce w innych tekstach należących do badanego języka.

Pomiarów częstości występowania poszczególnych znaków dokonano dla języka polskiego PL oraz dla języków: niemieckiego D, angielskiego GB, niderlandzkiego NL, francuskiego F, hiszpańskiego E i włoskiego I. Uzyskane wyniki podano w procentach w tabl. 2.

Uzyskane wyniki pomiarów pozwoliły na wyznaczenie odległości między poszczególnymi językami. Odległości te zmierzono w tzw. metryce ulic, w której odległość między dwoma obiektami x i y wyznacza się jako sumę wartości bezwzględnych różnicy kolejnych współrzędnych. W rozważanym przypadku odległość między dwoma językami x i y wyraża się następującym wzorem

$$\rho = |a_x - a_y| + |b_x - b_y| + |c_x - c_y| + |d_x - d_y| + \mathrm{K} + |z_x - z_y| \quad (1)$$

a_x, b_x, c_x, ..., z_x oznaczają, wyrażone w procentach, częstości występowania znaków *a, b, c, ..., z* w języku *x*; analogiczne a_y, b_y, c_y, ..., z_y — dla języka *y*. Zmierzone za pomocą metryki ulic odległości między badanymi językami zamieszczono w tabl. 3.

Odległości między poszczególnymi językami odzwierciedlają dość dobrze genetyczne pokrewieństwa między badanymi językami. Np. wzajemne odległości między językami: francuskim, włoskim i hiszpańskim są stosunkowo niewielkie, ponieważ języki te należą do grupy języków romańskich.

w	x	y	z
0,0	0,0	3,2	0,0
0,9	0,5	3,3	0,0
1,5	0,2	2,3	0,0
1,6	0,2	2,1	0,0
1,7	0,3	2,2	0,0
1,7	0,2	1,8	0,1

w	x	y	z
3,0	0,0	4,1	6,7
1,6	0,1	0,1	1,3
1,7	0,2	1,8	0,1
1,9	0,0	0,0	1,3
0,0	0,3	0,1	0,9
0,0	0,1	0,8	0,2
0,0	0,0	0,0	1,3

Podobną sytuację można zaobserwować dla języków z germańskiej grupy językowej: niemieckiego, angielskiego i niderlandzkiego. Z kolei język polski, będący językiem słowiańskim, dzielą od wszystkich innych języków zamieszczonych w tabl. 3 zdecydowanie największe odległości. Odległość między językiem francuskim i niemieckim jest stosunkowo niezbyt wielka, co wytłumaczyć można faktem, że język francuski (mimo, że wywodzący się z łaciny i w związku z tym zaliczany do romańskiej grupy językowej) zachował również pewne cechy pierwotnego języka germańskich Franków. Również odległość między językiem angielskim i francuskim nie jest zbyt wielka, co nie budzi większego zdziwienia, ponieważ ponad połowa angielskiego słownictwa ma swój łaciński źródłosłów.

Tablica 3. Odległości między poszczególnymi językami zmierzone w metryce ulic

	PL	D	GB	NL	F	E	I
PL	-	64,8	53,1	57,8	61,8	51,4	53,1
D	64,8	-	28,0	32,4	34,4	43,6	49,3
GB	53,1	28,0	-	33,6	26,4	28,6	36,5
NL	57,8	32,4	33,6	-	34,6	43,2	48,3
F	61,8	34,4	26,4	34,6	-	25,5	29,8
E	51,4	43,6	28,6	43,2	25,5	-	25,4
I	53,1	49,3	36,5	48,3	29,8	25,4	-

Rozpoznawanie tekstów należących do poszczególnych języków

Rozpoznawanie języka tekstów należących do badanych języków przeprowadzono metodą minimalnoodległościową (zwaną również metodą najbliższego sąsiedztwa). Rozpoznawanie taką metodą przebiega w następujący sposób: Dysponując próbką tekstu zapisanego w nieznanym języku, wyznacza się częstość występowania poszczególnych znaków w tym języku. Następnie wyznacza się odległość badanego tekstu od każdego z języków ujętych w systemie. Jako język tekstu jest wskazywany ten, dla którego otrzymano najmniejszą wartość zmierzonej odległości (stąd pochodzi również nazwa metody minimalnoodległościowej).

Przed przystąpieniem do eksperymentów, mających na celu sprawdzenie skuteczności działania zaproponowanego systemu rozpoznawania języka tekstu, należy określić minimalną długość tekstu, aby wyniki jego rozpoznawania można było uznać za wiarygodne. Badania, mające na celu określenia minimalnej długości przeprowadzono dla języka angielskiego.

Rozpoznawano język z próbki napisanej w języku angielskim, o długości ok. 60 znaków. Przy zastosowaniu metryki ulic zmierzono odległości, jakie dzielą badany tekst od każdego z języków występujących w systemie (wymienione w tabl. 3). W sześciu przypadkach na dziesięć został prawidłowo rozpoznany język angielski. Jednakże skuteczność rozpoznawania na poziomie 60 %, mimo że znacznie wyższa niż wynikałoby to z czystego przypadku, jest zbyt mała dla zastosowań praktycznych. W związku z tym do rozpoznawania należy zastosować teksty dłuższe.

Zwiększenie długości rozpoznawanego tekstu do ok. 120 znaków poprawiło znacznie skuteczność rozpoznawania — prawidłowo rozpoznano dziewięć tekstów na dziesięć.

Analogiczne badania przeprowadzono dla tekstów zapisanych w języku angielskim o długości 180 znaków. W tym przypadku wszystkie próby rozpoznawania

zakończyły się sukcesem, za każdym razem jako język tekstu został wskazany język angielski.

Do oceny jakości procesu rozpoznawania może posłużyć współczynnik η, zdefiniowany jako stosunek średniej odległości od drugiej z kolei (pod względem bliskości) klasy przynależności do średniej odległości od właściwej dla badanych tekstów klasy przynależności. Ponieważ przy próbce długości ok. 180 znaków średnia odległość badanych tekstów od klasy reprezentującej język angielski wynosiła 27,0, natomiast średnia odległość od klasy kolejnej pod względem bliskości wynosiła 35,7, zatem współczynnik η przyjmuje wartość 1,32.

Eksperyment rozpoznawania przeprowadzono również dla tekstów o długości ok. 240 znaków. Również w tym przypadku wszystkie próby rozpoznawania języka zakończyły się pełnym sukcesem, polegającym na wytypowaniu właściwej klasy przynależności. Wartość współczynnika η wyliczona dla uzyskanych wyników wynosiła 1,46.

Wzrost wartości współczynnika η powoduje, że prawdopodobieństwo popełnienia przez system rozpoznający pomyłki jest mniejsze, ponieważ odległości od kolejnych pod względem bliskości klas przynależności ulegają zwiększeniu.

Wyniki eksperymentów

Eksperymenty rozpoznawania języka tekstów przeprowadzono dla tekstów o długości 200 znaków, ponieważ jest to długość gwarantująca odpowiedni poziom skuteczności procesu rozpoznawania. Należy zwrócić uwagę, że rozpoznawaniu były poddawane próbki tekstu, nie używane wcześniej przy uczeniu systemu.

W tabl. 4 zamieszczono wyniki rozpoznawania 15 próbek tekstu zapisanych w języku polskim. Drukiem pogrubionym zaznaczono odległości, jakie dzielą badany fragment tekstu od dwóch najbliższych (w sensie odległości wyrażonej w metryce ulic) klas przynależności.

Analizując dane zamieszczone w tabl. 4 można wyliczyć, że średnia odległość badanych próbek tekstu od klasy PL wynosi 32,53. Z kolei średnia odległość do następnej najbliższej klasy (GB, E, I) wynosi 49,10. Zatem współczynnik η charakteryzujący jakość procesu rozpoznawania wynosi 1,51, co pozwala na skuteczne odróżnienie próbek tekstu zapisanych w języku polskim od pozostałych języków występujących w systemie.

Przy rozpoznawaniu 15 próbek tekstu o długości 200 znaków, zapisanych w języku niemieckim, z wartości odległości zmierzonych przy zastosowaniu metryki ulic, jakie dzielą badany tekst od każdego z występujących w systemie języków, można było wyliczyć, że średnia odległość badanych próbek tekstu od klasy D wynosiła 25,45. Z kolei średnia odległość do następnej najbliższej klasy (GB, NL) – 32,27. Zatem współczynnik charakteryzujący jakość procesu rozpoznawania η wynosił 1,46, co pozwala na skuteczne odróżnienie próbek tekstu zapisanych w języku niemieckim od pozostałych języków występujących w systemie.

Przy rozpoznawaniu języka fragmentów tekstu napisanych w języku angielskim o długości 200 znaków, ze zmierzonych odległości można było wyliczyć, że średnia odległość badanych próbek tekstu od klasy GB wynosiła 27,13. Z kolei średnia odległość do następnej najbliższej klasy (D, NL, F, E, I) – 36,93. Zatem współczynnik charakteryzujący jakość procesu rozpoznawania η wynosił 1,36, co pozwala na skuteczne odróżnienie próbek tekstu zapisanych w języku angielskim od pozostałych języków występujących w systemie.

Niestety ze względu na znaczne pokrewieństwo genetyczne, występujące między językami francuskim, hiszpańskim i włoskim, rozpoznawanie próbek tekstów o długości 200 znaków nie dało zbyt dobrych rezultatów. Poprawę przyniosło zastosowanie tekstów o długości 400 znaków.

Podsumowanie

Wyniki przeprowadzonych eksperymentów przekonują o słuszności wyboru zaprezentowanej minimalnoodległościowej metody rozpoznawania języka tekstów. Dla każdego z rozpoznawanych w systemie języków przeprowadzono po 15 prób rozpoznania różnych próbek tekstów o ustalonej liczbie znaków. Wszystkie próby rozpoznania zakończyły się pełnym sukcesem, polegającym na wytypowaniu dla danego języka właściwej dla niego klasy przynależności. Miernikiem jakości procesu rozpoznawania jest wartość współczynnika η, określającego stosunek odległości do dwóch najbliższych klas przynależności wskazanych podczas zastosowania metody minimalnoodległościowej. Czym większa jest wartość współczynnika η, tym mniejsze jest prawdopodobieństwo popełnienia pomyłki podczas rozpoznawania języka tekstu. Największą wartość współczynnika η uzyska-

Tablica 4. Wyniki rozpoznawania języka tekstu uzyskane dla fragmentów tekstu napisanych w języku polskim o długości 200 znaków

Lp.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PL	**30,3**	**33,5**	**26,7**	**33,3**	**32,7**	**28,9**	**30,5**	**47,5**	**39,5**	**30,9**	**32,7**	**27,5**	**33,3**	**23,1**	**37,5**
D	64,9	58,5	52,3	52,3	57,7	71,7	65,5	70,1	55,9	56,1	57,3	56,5	53,3	64,5	64,3
GB	57,0	54,6	43,0	**45,0**	47,6	62,0	59,2	62,6	49,8	50,2	**49,4**	**44,8**	**47,2**	54,8	60,4
NL	61,5	61,5	41,9	53,0	60,5	68,3	61,1	73,5	58,9	56,3	61,7	55,5	60,7	61,5	61,9
F	65,7	63,9	50,1	53,7	56,9	68,9	64,9	69,3	58,3	57,5	59,0	54,1	54,1	62,5	69,9
E	55,3	53,9	**38,9**	51,3	**43,7**	59,3	56,3	**59,7**	49,5	47,7	57,1	49,9	49,1	50,3	58,5
I	**55,2**	**53,4**	41,4	55,0	44,0	**52,0**	**48,0**	65,8	**47,8**	**47,6**	50,6	53,2	48,0	**49,8**	**54,0**

no dla języka polskiego ($\eta = 1{,}51$), natomiast najmniejszą dla języka francuskiego ($\eta = 1{,}28$). Zatem opracowany system rozpoznawania języka tekstu działa bardzo dobrze w przypadku języków genetycznie dość daleko od siebie odległych (należących do różnych rodzin językowych). W przypadku rozpoznawania języków blisko ze sobą spokrewnionych prawdopodobieństwo popełnienia pomyłki jest oczywiście większe. Z tego powodu w przypadku rozpoznawania języków romańskich: francuskiego, hiszpańskiego i włoskiego, należało zwiększyć długość analizowanych tekstów do 400 znaków. Takie powiększenie materiału badawczego pozwoliło na skuteczne odróżnianie przez system wyżej wymienionych języków.

Zaproponowany system rozpoznawania języka tekstów może zostać wykorzystany na przykład, jako rodzaj preprocesora w większym systemie służącym do automatycznego tłumaczenia wielojęzycznych tekstów, gdzie w pierwszym rzędzie zachodzi konieczność ustalenia z jakim językiem ma się do czynienia.

Uzyskane wyniki badań stanowią również ważny przyczynek o charakterze interdyscyplinarnym. Stosując bowiem metodę minimalnoodległościową w celu wyznaczenia odległości, jaką dzieli badany tekst od każdej z klas przynależności, uzyskano swoisty rodzaj liczbowej miary wzajemnego pokrewieństwa języków. Otrzymane w ten sposób rezultaty mogą zostać wykorzystane np. w dziedzinie lingwistyki, celem weryfikacji hipotez o pochodzeniu i wzajemnym pokrewieństwie różnych języków.

Bibliografia

1 Juang B. H., Furui S.: Automatic recognition and understanding of spoken language – A first step toward natural human-machine communication. Proceedings of the IEEE, vol. 88, no. 8, August 2000, p. 1142-1165.

2 Szczepaniak L., Królikowski Z.: Kontrolowane języki naturalne – przegląd rozwiązań i zastosowań. Pro Dialog, nr 11/2000, Wydawnictwo NAKOM, Poznań, 2000, s. 47-67.

3 Bu L., Chiueh T. D.: Perceptual speech processing and phonetic feature mapping for robust vowel recognition. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, March 2000, p. 105-114.

4 Saul L. K., Rahim M. G.: Maximum likehood and minimum classification error factor analysis for automatic speech recognition. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, March 2000, p. 115-125.

5 Lleida E., Rose R. C.: Utterance verification in continuous speech recognition: Decoding and training procedures. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, March 2000, p. 126-139.

6 Bellegarda J. R.: Large vocabulary speech recognition with multispan statistical language models. IEEE Transactions on speech and audio processing, vol. 8, no. 1, January 2000, p. 76-84.

7 Levin E., Pieraccini R., Eckert W.: A stochastic model of human-machine interaction for learning dialog strategies. IEEE transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, p. 11-22.

8 Souvingnier B., Keller A., Rueber B., Schramm H., Seide F.: The thoughtful elephant: Strategies for spoken dialog systems. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, p. 51-62.

9 Zue V., Seneff S., Glass J. R., Poliforoni J., Pao C., Hazen T. J., Hetherington L.: Jupiter: A telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, p. 85-95.

10 Ricćardi G., L. Gorin A.: Stochastic language adaptation over time and state in natural spoken dialog systems. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, p. 3-10.

11 Ney H., Niessen S., Och F. J., Sawaf H., Tillmann C., Vogel S.: Algorithms for statistical translation of spoken language. IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, January 2000, p. 24-35.

12 Siu M., Ostendorf M.: Variable N-grams and extentions for conversational speech language modeling. IEEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, january 2000, p. 63-74.

13 Waibel A., Geatner P., Tomokiyo L. M., Schultz T., Woszczyna M.: Multilinguality in speech and spoken language systems. Proceedings of the IEEE, vol. 88, no. 8, August 2000, p. 1297-1313.

14 Dulas J.: Parametryzacja fonemów mowy polskiej. Materiały Konferencyjne, XXXII Międzyuczelniana Konferencja Metrologów (Rzeszów-Jawor 11-15.09.2000), Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów, 2000, s. 257-262.

Streszczenia artykułów naukowych

System automatycznego rozpoznawania języka tekstu, Mirosław Gajer — s. 29

Opisano system służący do automatycznego rozpoznawania języka tekstów. System taki może stanowić pierwsze ogniwo większego systemu, służącego do automatycznego tłumaczenia wielojęzycznych tekstów, gdzie zachodzi konieczność ustalenia języka tłumaczonego dokumentu. Opisano metodę rozpoznawania, bazującą na pomiarze częstości występowania w tekście poszczególnych znaków, pomiarze minimalnej odległości do klasy przynależności. Zamieszczono wyniki eksperymentów, potwierdzających poprawną pracę systemu.

Automatic language for text recognition system, Mirosław Gajer -p. 29

The automatic language for text recognition system is described. The system can be a first step towards a construction of automatic translation system. In case of multi-language approach, it is essential to determine at the beginning the language of text that is to be translated. In the paper the proposed recognition method is described in details. Especially the minimal-distance recognition method, which is used for determining language proprietary class, is paid a special attention. The recognition results are also presented.